A Convex Discriminant Semantic Correlation Analysis for Cross-View Recognition

DOI:
10.1109/TCYB.2020.2988721

Document Version
Accepted author manuscript

Link to publication record in Manchester Research Explorer

Citation for published version (APA):
Tian, Q., Ma, C., Cao, M., Chen, S., & Yin, H. (2020). A Convex Discriminant Semantic Correlation Analysis for Cross-View Recognition. *IEEE Transactions on Cybernetics*. Advance online publication. https://doi.org/10.1109/TCYB.2020.2988721

Published in:
IEEE Transactions on Cybernetics

A Convex Discriminant Semantic Correlation Analysis for Cross-View Recognition

Qing Tian, Chuang Ma, Meng Cao, Songcan Chen*, *and* Hujun Yin, *Senior Member, IEEE*

***Abstract*—Canonical correlation analysis (CCA) is a typical statistical model used to analyze the correlation components between different view representations of the same objects. When the label information is available with the data representations, CCA can be extended to its discriminative counterparts by incorporating supervision in the analysis. Although most discriminative variants of CCA have achieved improved results, nearly all of their objective functions are nonconvex, implying that optimal solutions are difficult to obtain. More importantly, that cross-view representations from the same sample should be consistent, i.e., the cross-view semantic consistency, has however not been modelled. To overcome these drawbacks, in this paper we propose a Discriminant Semantic Correlation Analysis (DSCA) model by modelling the cross-view semantic consistency for each object in the sample space rather than in the commonly used feature space. To boost the nonlinear discriminating capability of DSCA, we extend it from the Euclidean to the geodesic space by transforming the metric and incorporating both the cross-view semantic and representation correlation information and consequently obtain our final model with convex objective, namely Convex DSCA (C-DSCA). Finally, with extensive experiments and comparisons we validate the effectiveness and superiority of the proposed method.**

***Index Terms*—Canonical correlation analysis; cross-view semantic consistency; cross-view representation correlation; discriminant semantic correlation analysis; convex discriminant semantic correlation analysis**

I. Introduction

In pattern recognition and machine learning, canonical correlation analysis (CCA) is a typical method used to analyze correlations between two or more types of feature views of a given dataset [1], [2], and is widely adopted to extract related representations from individual views and fuse them together for pattern classification tasks [3], [4], [5], [6], [7]. Specially, given two (or more) views of feature representation of interested data, traditional CCA aims to find corresponding projection directions for individual views, along which the correlations of the views are maximized. Then, with the projected and fused view representations via CCA, classification or regression decisions can be made. In above process, CCA works in an unsupervised manner without utilizing the data labels, leaving a room for performance improvement.

If class labels are available with the data representations, CCA can be extended to its discriminant counterparts by incorporating the label information. Along this line, Sun et al. [8] proposed the discriminative CCA (DCCA) method by maximizing the intra-class correlations while minimizing the inter-class relatedness. Peng et al. [9] proposed local discriminative CCA (LDCCA) by assuming data distribution following a low-dimensional manifold embedding. Su et al. [10] constructed multiple metrics instead of single metric to better measure intra-class correlations and embedded them into the CCA objective function and thus developed the multi-patch embedding CCA (MPECCA). Sun et al. [11] presented a generalized CCA (GCCA) model by unifying the formulation. Ji et al. [12] decomposed the scatter matrices into more discriminative fractional-order components to replace the original CCA objective function. Shen et al. [13] proposed to perform multi-label prediction through cross-view search.

In addition, other researchers also proposed to incorporate discriminative terms into the CCA objective to supervise its learning. Along this line, Zhou et al. [14] constructed the combined-feature-discriminability enhanced CCA (CECCA) by incorporating the linear discriminant analysis (LDA) [15] guided feature combination term into CCA. Zhao et al. [16] proposed hierarchical supervised local CCA (HSL-CCA) by penalizing the inter-class scatters within varying size of neighborhood. Haghighat et al. [17] generated the discriminant correlation analysis (DCA) by decomposing the inter-class scatter matrix. Recently, to extend correlation analysis to multi-view tasks, several multi-view models have been proposed, for instance, generalized multi-view analysis (GCA) [18], multiset canonical correlation analysis (GbMCC-DR) [19], multi-view uncorrelated discriminant analysis (MULDA) [20], semi-paired discrete hashing (SPDH) [21], and local feature based multi-view discriminant analysis (FMDA) [22]. More recently, CCA and its variants have been extended to deep architectures for higher discriminating capability, such as Deep CCA [23], Deep Soft CCA [24], Deep Complete CCA [25],

Qing Tian is with the School of Computer and Software, Nanjing University of Information Science and Technology, Nanjing 210044, China, also with the Jiangsu Key Laboratory of Big Data Analysis Technology, Nanjing University of Information Science and Technology, Nanjing 210044, China, also with the Collaborative Innovation Center of Atmospheric Environment and Equipment Technology, Nanjing University of Information Science and Technology, Nanjing 210044, China, also with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China, and also with the MIIT Key Laboratory of Pattern Analysis and Machine Intelligence, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China (e-mail: tianqing@nuist.edu.cn).

Chuang Ma and Meng Cao are with the School of Computer and Software, Nanjing University of Information Science and Technology, Nanjing 210044, China (e-mail: mcboo@nuist.edu.cn, alrash@nuist.edu.cn).

Songcan Chen is the corresponding author and is with the MIIT Key Laboratory of Pattern Analysis and Machine Intelligence, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China, and also with the College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China (e-mail: s.chen@nuaa.edu.cn, *corresponding author*).

Hujun Yin is with the Department of Electrical and Electronic Engineering, The University of Manchester, Manchester M13 9PL, U.K. (e-mail: hujun.yin@manchester.ac.uk).

Manuscript received on 22 March, 2019.

and CM-DDA [26].

Although the above CCA variants can bring accuracy improvement, nearly all of their objective functions are nonconvex [27], [10], [28], meaning that optimal solutions are difficult to obtain. Specially, their objective functions are frequently combined with constraints for the sake of preventing them from trivial solutions, which limit their solution space and render them not convex (e.g., MPECCA) that can only be solved in a time-consuming alternating manner. To pursue a closed-form solution, Jiang et al. [29] constructed a convex correlation analysis model, named CDCA, by following the scheme of geometric mean metric learning (GMML) [30], [31]. Although the CDCA model yielded much better evaluation performance than the previous models, it does not consider the semantic consistency of sample across view representations.

To overcome the drawbacks aforementioned, in this paper we first construct a discriminant semantic correlation analysis (DSCA) by modelling the cross-view representation correlation, discriminative metric, as well as the semantic consistency. To generalize the discriminating ability of DSCA, we extend it from the Euclidean space to the geodesic space and consequently obtain our final model with convex objective function, coined as Convex DSCA (C-DSCA). The objective function of C-DSCA is constructed from the perspective of geometric means, which enables our model to have a closed-form solution. Moreover, we further extend the C-DSCA model to the manifold represented by the nonlinear geodesic space, which is also convex in the geodesic metric. Finally, we conduct extensive experiments to validate the proposed methods. To sum up, the main contributions are of three folds:

1) A discriminant semantic correlation analysis (DSCA) model is constructed by incorporating both the cross-view semantic and representation correlation information in the sample representation space.
2) To enhance the discriminating ability of DSCA, it is extended with convex objective function (C-DSCA) in the geodesic metric space, enjoying closed-form solution.
3) Effectiveness and superiority of the proposed models are verified through extensive experimental evaluations.

The remainder of this paper is organized as follows. Section II briefly reviews the principle of CCA. Section III describe the proposed methods, and the experimental results are reported in Section IV. Finally, Section V concludes the paper and gives future research directions.

II. Canonical correlation analysis

In this section, we briefly review canonical correlation analysis (CCA) [1] [2]. Given two views of feature representations of the same sample set, CCA aims to seek two groups of projection directions for individual views, along which the correlations of the two views are maximized. Specially, assume $\mathbf{X} = [\mathbf{x}_1, ..., \mathbf{x}_N] \in \mathbb{R}^{p\times N}$ and $\mathbf{Y} = [\mathbf{y}_1, ..., \mathbf{y}_N] \in \mathbb{R}^{q\times N}$ are two view representations of N samples, and $\mathbf{x}_i$ and $\mathbf{y}_i$ are the two representations of the ith sample, which are already normalized. To fuse the two views together for subsequent classification, CCA seeks two sets of projection matrices, $\mathbf{W}_x \in \mathbb{R}^{p\times r}$ and $\mathbf{W}_y \in \mathbb{R}^{q\times r}$, to transform the two views into r-dimensional common space. In the projected space, the correlation of the new view representations $\mathbf{W}_x^T\mathbf{x}_i$ and $\mathbf{W}_y^T\mathbf{y}_i$ is maximized. Formally, the objective function of the CCA is formulated as

$$\max_{\{\mathbf{W}_x,\mathbf{W}_y\}} \frac{\mathbf{W}_x^T\mathbf{C}_{xy}\mathbf{W}_y}{\sqrt{\mathbf{W}_x^T\mathbf{C}_{xx}\mathbf{W}_x\mathbf{W}_y^T\mathbf{C}_{yy}\mathbf{W}_y}}, \tag{1}$$

where $\mathbf{C}_{xx} = \frac{1}{N}\sum_{i=1}^{N}(\mathbf{x}_i-\bar{\mathbf{x}})(\mathbf{x}_i-\bar{\mathbf{x}})^T$, $\mathbf{C}_{yy} = \frac{1}{N}\sum_{i=1}^{N}(\mathbf{y}_i-\bar{\mathbf{y}})(\mathbf{y}_i-\bar{\mathbf{y}})^T$, and $\mathbf{C}_{xy} = \frac{1}{N}\sum_{i=1}^{N}(\mathbf{x}_i-\bar{\mathbf{x}})(\mathbf{y}_i-\bar{\mathbf{y}})^T$, with $\bar{\mathbf{x}} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i$ and $\bar{\mathbf{y}} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{y}_i$ being sample means of the two views, respectively. The numerator terms characterize the correlation of the two views, while the denominator terms restrict the complexity of individual views to avoid degraded solutions. For convenience of solving (1), it is usually transformed to a generalized-eigenvalue problem as follows

$$\begin{pmatrix} & \mathbf{XY}^T \\ \mathbf{YX}^T & \end{pmatrix}\begin{pmatrix}\mathbf{W}_x \\ \mathbf{W}_y\end{pmatrix} = \lambda\begin{pmatrix}\mathbf{XX}^T & \\ & \mathbf{YY}^T\end{pmatrix}\begin{pmatrix}\mathbf{W}_x \\ \mathbf{W}_y\end{pmatrix}. \tag{2}$$

The concatenated projections $\begin{pmatrix}\mathbf{W}_x \\ \mathbf{W}_y\end{pmatrix}$ can be obtained by calculating the eigenvectors of $\begin{pmatrix}\mathbf{XX}^T & \\ & \mathbf{YY}^T\end{pmatrix}^{-1}\begin{pmatrix} & \mathbf{XY}^T \\ \mathbf{YX}^T & \end{pmatrix}$ corresponding to the r largest eigenvalues.

With the obtained $\mathbf{W}_x$ and $\mathbf{W}_y$, the view representations $\mathbf{x}_i$ and $\mathbf{y}_i$ of the ith sample can be fused by $\mathbf{W}_x^T\mathbf{x}_i + \mathbf{W}_y^T\mathbf{y}_i = \begin{pmatrix}\mathbf{W}_x \\ \mathbf{W}_y\end{pmatrix}^T\begin{pmatrix}\mathbf{x}_i \\ \mathbf{y}_i\end{pmatrix}$.

III. The proposed method

In the existing multi-view correlation learning (CCA and its variants), the models are typically constructed in Euclidean space. However, as mentioned previously, they mainly suffer from three aspects of drawbacks: (1) most of them were trained in unsupervised manner without using the labels information, limiting their discriminating ability; (2) even though the labels information is modeled in forms of constraints, their objective functions are induced as nonconvex and hence difficult to solve; (3) the cross-view correlations are modeled in linear or finite nonlinear (via kernel trick [32], [33]) Euclidean distance space, which is not powerful enough to distinguish the view dissimilarity or characterize the cross-view semantic consistency. To overcome these drawbacks, in this section we first build a cross-view discriminant semantic correlation analysis model (DSCA) to model the correlations across the views in the sample semantic space, rather than in the previously-adopted feature space. In this way, not only the correlations across views but also their semantic consistency can be more desirably exploited. Then, to further enhance the discriminating ability of the DSCA model, we remodel it from the Euclidean space to the Riemannian manifold space and generate the extended DSCA (C-DSCA) whose objective function is convex in the geodesic metric space. Finally, we validate the effectiveness of the proposed models with experimental evaluations.

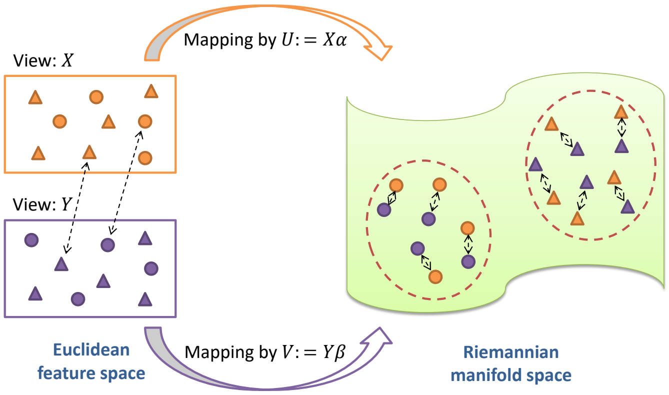


Fig. 1: Overview of the cross-view semantic correlation analysis. Samples from distinct views are denoted in different colors and different classes are denoted with diverse shapes. Through semantic discriminant correlation mapping by distinct transformations $\mathbf{U}$ and $\mathbf{V}$ that are represented in individual sample spaces $\mathbf{X}$ and $\mathbf{Y}$, the original distinct view spaces are transformed to a common Riemannian manifold space, in which the similar samples (within and across the views) are pulled much nearer while those dissimilar are pushed further away.

A. Discriminant sematic correlation learning (DSCA)

Discriminant cross-view correlation exploitation For convenience of presentation, we first define some notations. For given N training samples from C classes, they are represented in two types of view representations: $\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_C] \in \mathbb{R}^{p \times N}$ with the kth class set $\mathbf{X}_k = [\mathbf{x}_1^k, \mathbf{x}_2^k, \cdots, \mathbf{x}_{n_k}^k]$ and $\mathbf{Y} = [\mathbf{Y}_1, \mathbf{Y}_2, \cdots, \mathbf{Y}_C] \in \mathbb{R}^{q \times N}$ with the kth class set $\mathbf{Y}_k = [\mathbf{y}_1^k, \mathbf{y}_2^k, \cdots, \mathbf{y}_{n_k}^k]$. To enhance the discriminating power of estimators in cross-view analysis, we should take into account the supervision knowledge like the labels information from two aspects: the representation correlation across views and the discriminative similarities within and between data classes. To this end, we should seek two individual projections (denoted as $\mathbf{U} \in \mathbb{R}^{p \times r}$ and $\mathbf{V} \in \mathbb{R}^{q \times r}$, assuming the dimension of the projected common space is r) for the view representations of training data. Taking into account the above considerations, we consequently build the discriminant cross-view correlation learning as follows,

$$\begin{aligned}\min_{\{\mathbf{U},\mathbf{V}\}} \; & \frac{1}{N}\sum_{i=1}^{N}\left(\mathbf{U}^T\mathbf{x}_i - \mathbf{V}^T\mathbf{y}_i\right)^T\left(\mathbf{U}^T\mathbf{x}_i - \mathbf{V}^T\mathbf{y}_i\right) \\ & + \frac{1}{N_w}\sum_{k=1}^{C}\sum_{i=1}^{n_k}\sum_{j=1}^{n_k}\left(\mathbf{U}^T\mathbf{x}_i - \mathbf{V}^T\mathbf{y}_j\right)^T\left(\mathbf{U}^T\mathbf{x}_i - \mathbf{V}^T\mathbf{y}_j\right) \\ & - \frac{1}{N_b}\lambda_1\sum_{k=1}^{C}\sum_{c=1,c\neq k}^{C}\sum_{i=1}^{n_k}\sum_{j=1}^{n_c}\left(\mathbf{U}^T\mathbf{x}_i - \mathbf{V}^T\mathbf{y}_j\right)^T\left(\mathbf{U}^T\mathbf{x}_i - \mathbf{V}^T\mathbf{y}_j\right)\end{aligned} \tag{3}$$

where $N_w = \sum_{k=1}^{C} n_k^2$ and $N_b = \sum_{k=1}^{C} n_k(N - n_k)$, λ_1 is a tradeoff parameter controlling the balance between the three terms. The first term characterizes the cross-view representation correlation of each sample, the second term denotes the similarity of samples from the same class, while the third term models the dissimilarity of between-class samples. Obviously, the second and third terms are generated based on the data labels, and in this way, the supervised discriminative information is incorporated in the objective function. After equivalent reformulation, (3) can be simplified to

$$\min_{\mathbf{A}\succ 0} \; tr\left(\mathbf{AC}\right) + tr\left(\mathbf{AS}\right) - \lambda_1 tr\left(\mathbf{AD}\right), \tag{4}$$

where $\mathbf{A} = \mathbf{H}\mathbf{H}^T$ with $\mathbf{H}^T = [\mathbf{U}^T, \mathbf{V}^T]$, $\mathbf{C} = \mathbf{M}\mathbf{M}^T$ with $\mathbf{M} = [\mathbf{X}^T, -\mathbf{Y}^T]^T$, $\mathbf{S} = \sum_{k=1}^{C} 2\left(n_k\mathbf{L}_k\mathbf{L}_k^T - \mathbf{L}_k\mathbf{1}_{n_k}\mathbf{1}_{n_k}^T\mathbf{L}_k^T\right)/N_w$ with $\mathbf{L}_k = [\mathbf{X}_k^T, \mathbf{Y}_k^T]^T$ and $\mathbf{1}_{n_k}$ being an n_k-dimensional all-one vector, and $\mathbf{D} = \left(2N\mathbf{L}\mathbf{L}^T - 2\mathbf{L}\mathbf{1}_N\mathbf{1}_N^T\mathbf{L}^T - \mathbf{S}\right)/N_b$ with $\mathbf{L} = [\mathbf{L}_1, \cdots, \mathbf{L}_C]$ and $\mathbf{1}_N$ being an N-dimensional all-one vector. With $\mathbf{A}$ solved from (4), we can readily obtain $\mathbf{U}$ and $\mathbf{V}$ by decomposing it as square root. Then, we can map the distinct view representations respectively by $\mathbf{U}$ and $\mathbf{V}$ to the common representation space and then concatenate them for subsequent classifications.

Cross-view semantic consistency exploitation We can see from (3) that, although the cross-view correlation of each sample is characterized by the first term, the semantic consistency across view representations is not exploited. More accurately, the cross-view semantic *consistency* is much stricter than the semantic *correlation* characterization. We take face images as an example. As demonstrated in Figure 2, the *semantic correlation* measures the similarity degree of face images. More specifically, face A1 is more similar to A2/B2 in appearance than to A3/B3 within or between the views. In contrast, *semantic consistency* requires the cross-view representations of the same face image are semantically identical. From this viewpoint, A1 and B1 are semantically identical since they are different representations of the same subject face.

Regretfully, the aforementioned semantic representation consistency across view representations is not exploited in (3) at all. Nevertheless, it is quite challenging to model such cross-view semantic consistency since it cannot be directly characterized through the view projections $\mathbf{U} \in \mathbb{R}^{p \times r}$ and $\mathbf{V} \in \mathbb{R}^{q \times r}$, as they are with mismatched dimensions. Therefore, we need to turn to other modeling schemes. Surprisingly, the representer theorem [34], [35] provides us a direction. Motivated by this, we propose to represent the projection matrices

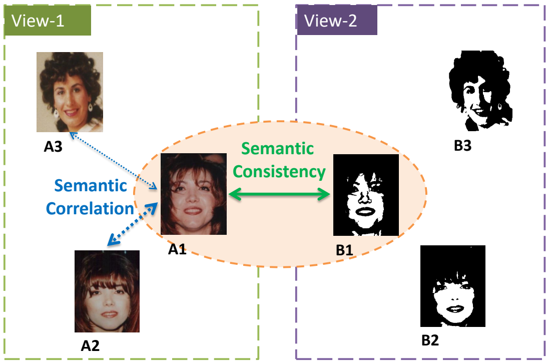


Fig. 2: Semantic correlation versus semantic consistency across views. Ai, $i = 1, 2, 3$ are three face image representations in view 1, while Bi, $i = 1, 2, 3$ are their individual representations in view 2. The *semantic correlation* measures the similarity degree of face images, for instance, face A1 is more similar to A2/B2 in appearance than to A3/B3 within or between the views. In contrast, the *semantic consistency* requires the cross-view representations of the same face image are semantically identical, for instance, A1 and B1 are semantically identical since they are different representations of the same subject face.

in respective sample space of distinct views. Specifically, we reconstruct $\mathbf{U} = \mathbf{X}\boldsymbol{\alpha}$ and $\mathbf{V} = \mathbf{Y}\boldsymbol{\beta}$, where $\boldsymbol{\alpha} \in \mathbb{R}^{N\times r}$ and $\boldsymbol{\beta} \in \mathbb{R}^{N\times r}$ denote the representation coefficient matrices. In this way, $\mathbf{H}$ involved in (4) can be reconstructed as

$$\mathbf{H} = \begin{bmatrix} \mathbf{U} \\ \mathbf{V} \end{bmatrix} = \begin{bmatrix} \mathbf{X}\boldsymbol{\alpha} \\ \mathbf{Y}\boldsymbol{\beta} \end{bmatrix} = \begin{bmatrix} \mathbf{X} & \\ & \mathbf{Y} \end{bmatrix} \begin{bmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\beta} \end{bmatrix}. \tag{5}$$

Then, we substitute (5) into $\mathbf{A} = \mathbf{H}\mathbf{H}^T$ and consequently the discriminant cross-view correlation learning (4) can be reformulated as

$$\min_{\mathbf{A}\succ 0} \; tr\left(\mathbf{AC}\right) + tr\left(\mathbf{AS}\right) - \lambda_1 tr\left(\mathbf{AD}\right), \tag{6}$$

with $\mathbf{C}$, $\mathbf{S}$ and $\mathbf{D}$ defined the same as in (4). Formally, although (6) is completely the same as (4), there is essential difference between them: (6) is modeled in the *sample representation space* while (4) in the *feature representation space*. To investigate the semantic consistency in the sample representation space and incorporate it into (6), we are surprised to find that $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are matched and essentially identical in the semantic representation. Consequently, we can characterize the cross-view semantic consistency by measuring the divergence (or say discrepancy) between $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. As a result, the cross-view semantic consistency can be measured by

$$\min \; \|\boldsymbol{\alpha} - \boldsymbol{\beta}\|_2^2. \tag{7}$$

Recall that $\mathbf{U} = \mathbf{X}\boldsymbol{\alpha}$, we can derive $\boldsymbol{\alpha} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{U}$. Similarly, we can generate $\boldsymbol{\beta} = \left(\mathbf{Y}^T\mathbf{Y}\right)^{-1}\mathbf{Y}^T\mathbf{V}$ from $\mathbf{V} = \mathbf{Y}\boldsymbol{\beta}$. Substituting $\boldsymbol{\alpha} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{U}$ and $\boldsymbol{\beta} = \left(\mathbf{Y}^T\mathbf{Y}\right)^{-1}\mathbf{Y}^T\mathbf{V}$ into (7) yields

$$\min_{\{\boldsymbol{\alpha},\boldsymbol{\beta}\}} \; \|\boldsymbol{\alpha} - \boldsymbol{\beta}\|_2^2 \Rightarrow \min_{\mathbf{A}\succ 0} \; tr\left(\mathbf{A}\boldsymbol{\Gamma}\right), \tag{8}$$

where $\boldsymbol{\Gamma} = \begin{bmatrix} 2\mathbf{P}_1^T\mathbf{P}_1 & -2\mathbf{P}_2^T\mathbf{P}_1 \\ -2\mathbf{P}_1^T\mathbf{P}_2 & 2\mathbf{P}_2^T\mathbf{P}_2 \end{bmatrix}$, with $\mathbf{P}_1 = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T$ and $\mathbf{P}_2 = \left(\mathbf{Y}^T\mathbf{Y}\right)^{-1}\mathbf{Y}^T$.

Formulation of DSCA We can see from (8) that the semantic consistency across views is formulated as a concise term elegantly. Combining it into (6) consequently generates the objective function of the proposed *discriminant semantic correlation analysis* (DSCA) model as follows

$$\begin{aligned} \mathcal{J} &= \min_{\mathbf{A}\succ 0} \; tr\left(\mathbf{AC}\right) + tr\left(\mathbf{AS}\right) - \lambda_1 tr\left(\mathbf{AD}\right) + \lambda_2 tr\left(\mathbf{A}\boldsymbol{\Gamma}\right) \\ &= \min_{\mathbf{A}\succ 0} \; tr\big(\mathbf{A}\left(\mathbf{C} + \mathbf{S} - \lambda_1\mathbf{D} + \lambda_2\boldsymbol{\Gamma}\right)\big) \\ &= \min_{\mathbf{Q}^T\mathbf{Q}=\mathbf{I}} \; tr\big(\mathbf{Q}^T\left(\mathbf{C} + \mathbf{S} - \lambda_1\mathbf{D} + \lambda_2\boldsymbol{\Gamma}\right)\mathbf{Q}\big), \end{aligned} \tag{9}$$

where λ_2 is also a tradeoff parameter to control the balance between the semantic consistency term and other terms, and $\mathbf{I}$ denotes an identity matrix of proper size. The orthogonality constraint, $\mathbf{Q}^T\mathbf{Q} = \mathbf{I}$, is guaranteed to avoid trivial solutions. In this way, not only the cross-view representation consistency of each sample, similarity and dissimilarity of the data classes (discriminative supervision information), but also the semantic consistency across individual views are modeled simultaneously in (9).

For convenience of solving (9), we introduce the Lagrangian multipliers $\boldsymbol{\Lambda}$ and rewrite it as

$$\mathcal{J}_{\{\mathbf{Q},\boldsymbol{\Lambda}\}} = tr\big(\mathbf{Q}^T\left(\mathbf{C} + \mathbf{S} - \lambda_1\mathbf{D} + \lambda_2\boldsymbol{\Gamma}\right)\mathbf{Q}\big) - tr\big(\boldsymbol{\Lambda}(\mathbf{Q}^T\mathbf{Q} - \mathbf{I})\big). \tag{10}$$

Calculating the partial derivative of $\mathcal{J}_{\{\mathbf{Q},\boldsymbol{\Lambda}\}}$ with regard to $\mathbf{Q}$ and making it to zero yields

$$(\mathbf{C} + \mathbf{S} - \lambda_1\mathbf{D} + \lambda_2\boldsymbol{\Gamma})\mathbf{Q} = \mathbf{Q}\boldsymbol{\Lambda}, \tag{11}$$

which is a generalized eigen-decomposition problem and can be solved by finding its smallest eigenvectors. Along this line, the projection matrix $\mathbf{Q}$ can be obtained by calculating a required number of smallest eigenvectors of $\mathbf{C}+\mathbf{S}-\lambda_1\mathbf{D}+\lambda_2\boldsymbol{\Gamma}$. Finally, with obtained $\mathbf{Q}$, we can recover $\mathbf{A} = \mathbf{Q}\mathbf{Q}^T$.

B. Convex DSCA (C-DSCA)

After analyzing the objective function (9) of DSCA, we can see that it is a difference combination of trace terms regarding $\mathbf{A}$. Such a objective function may be nonconvex in practice [36], [37], since its Hessian matrix with regard to $\mathbf{A}$ is a difference combination of positive definite matrices and consequently may not be positive definite. Although its optimal solutions may be obtained via the generalized eigen-decomposition algorithm, its optimality may be limited since it is solved in a conditionally-convex solution space spanned by a set of linear orthogonal eigenvectors (see (11)), whose optimality is dominated by the convexity of (9).

To overcome these shortcomings, we remodel (9) in the geodesic space and reformulate it as a convex function with preferable closed-form solution. Compared to the Euclidean space, the separability of nonlinear data patterns in the geodesic space can be significantly improved and thus benefits their subsequent recognitions. More explanations about the superiority of modelling in geodesic space to that in Euclidean space will be elaborated in Section III-C. Before presenting the new model, we first give a proposition below

Proposition 1. *Given a d-order positive definite matrix D, for d-order matrix variable A defined in the positive definite space, it holds that*

$$\min_{\boldsymbol{A}\succ 0}\ tr(\boldsymbol{A}^{-1}\boldsymbol{D}) \Leftrightarrow \max_{\boldsymbol{A}\succ 0}\ tr(\boldsymbol{AD}), \tag{12}$$

whose proof is detailed in the Appendix.

Following **Proposition** 1, we equivalently reformulate D-SCA in (9) as

$$\min_{\mathbf{A}\succ 0}\ tr\left(\mathbf{AC}\right) + tr\left(\mathbf{AS}\right) + \lambda_1 tr(\mathbf{A}^{-1}\mathbf{D}) + \lambda_2 tr\left(\mathbf{A\Gamma}\right), \tag{13}$$

in which minimizing the third term $\lambda_1 tr(\mathbf{A}^{-1}\mathbf{D})$ is equivalent to minimizing the third term $-\lambda_1 tr\left(\mathbf{AD}\right)$ of (9). For the convexity of (13), since the first, second and last terms are linear with regard to $\mathbf{A}$ and readily convex; for the third term involved with the inverse of $\mathbf{A}$, it is also convex in in the cone space [38]. As a result, (13) is entirely convex with regard to A. More interestingly, it also enjoys closed-form solution [39], [40] [41], [30].

TABLE I: Properties comparison between $tr(\mathbf{AD})$ and $tr(\mathbf{A}^{-1}\mathbf{D})$.

Term	*Convex*	*Convexity Type*	*Derivative Regarding A*	*Monotonicity With A*
$tr(\mathbf{AD})$	Yes	Linear	$\mathbf{D}$	Increasing
$tr(\mathbf{A}^{-1}\mathbf{D})$	Yes	Conical	-$\mathbf{A}^{-1}\mathbf{DA}^{-1}$	Decreasing

Formulation of C-DSCA To obviously distinguish $tr(\mathbf{A}^{-1}\mathbf{D})$ from $tr(\mathbf{AD})$, we comparatively summarize the two operators in Table I. We can see that $tr(\mathbf{A}\bullet)$ is increasing while $tr(\mathbf{A}^{-1}\bullet)$ gets decreasing with $\mathbf{A}$. In this way, (13) enjoys the following merits: (1) the discriminative similarity and dissimilarity measures can be modelled ingeniously in a unified objective function; (2) modelled with two kinds of distance characterizations $tr(\mathbf{A}\bullet)$ and $tr(\mathbf{A}^{-1}\bullet)$, the metric of (13) is extended to a convex and more discriminative geodesic space, beneficial to the subsequent recognitions; (3) since (13) is modelled from the perspective of the geometric mean metric learning [30], [31], we can readily achieve its closed-form solutions. Along this line, we reformulate (13) as

$$\min_{\mathbf{A}\succ 0}\ \gamma tr(\mathbf{A}^{-1}\mathbf{D}) + (1-\gamma)(tr\left(\mathbf{AS}\right) + tr\left(\mathbf{AC}\right) + \delta tr\left(\mathbf{A\Gamma}\right)), \tag{14}$$

where we set $\gamma \in (0,1)$ [30]. Let $J(\mathbf{A}) := \gamma tr(\mathbf{A}^{-1}\mathbf{D}) + (1-\gamma)(tr\left(\mathbf{AS}\right) + tr\left(\mathbf{AC}\right) + \delta tr\left(\mathbf{A\Gamma}\right))$. Setting the gradient of $J(\mathbf{A})$ with respect to $\mathbf{A}$ to zero yields

$$(1-\gamma)\mathbf{A}(\mathbf{S}+\mathbf{C}+\delta\mathbf{\Gamma})\mathbf{A} = \gamma\mathbf{D}, \tag{15}$$

which is a Riccati Equation [42] [43] whose solution happens to be the midpoint of the geodesic jointing $((1-\gamma)(\mathbf{S}+\mathbf{C}+\delta\mathbf{\Gamma}))^{-1}$ and $\gamma\mathbf{D}$, that is

$$\mathbf{A} = ((1-\gamma)(\mathbf{S}+\mathbf{C}+\delta\mathbf{\Gamma}))^{-1}\sharp_{1/2}(\gamma\mathbf{D}), \tag{16}$$

where $(\cdot)\sharp_{1/2}(\cdot)$ denotes the geodesic mean (midpoint) jointing two matrices. To generalize the discriminating ability of the solution, we extend the geodesic mean solution (16) from the Euclidean space to the geodesic space (i.e., Riemannian manifold space) by replacing $(\cdot)\sharp_{1/2}(\cdot)$ with $(\cdot)\sharp_t(\cdot)$, $0 \leqslant t \leqslant 1$, as demonstrated in Figure 1, which is also convex [44].

In practice, the invertibility of $(1-\gamma)(\mathbf{S}+\mathbf{C}+\delta\mathbf{\Gamma})$ does not always hold since its rank may be not full. To address this issue, we regularize (14) by the symmetrized LogDet divergence and consequently generate the proposed Convex DSCA (coined as C-DSCA), as follows

$$\begin{aligned}\min_{\mathbf{A}\succ 0}\ &\gamma tr(\mathbf{A}^{-1}\mathbf{D}) + (1-\gamma)(tr\left(\mathbf{AS}\right) + tr\left(\mathbf{AC}\right) + \delta tr\left(\mathbf{A\Gamma}\right)) \\ &+ \lambda D_{sld}(\mathbf{A},\mathbf{A}_0),\end{aligned} \tag{17}$$

where $\mathbf{A}_0$ is a prior positive definite matrix and $D_{sld}(\mathbf{A},\mathbf{A}_0) = tr(\mathbf{A}\mathbf{A}_0^{-1}) + tr(\mathbf{A}^{-1}\mathbf{A}_0) - 2(p+q)$ stands for the symmetrized version of LogDet divergence [45]. Since (17) is still convex in the geodesic space, C-DSCA enjoys closed-form solution as follows

$$\mathbf{A} = ((1-\gamma)(\mathbf{S}+\mathbf{C}+\delta\mathbf{\Gamma}) + \lambda\mathbf{A}_0^{-1})^{-1}\sharp_t(\gamma\mathbf{D}+\lambda\mathbf{A}_0). \tag{18}$$

More specifically,

$$\begin{aligned}\mathbf{A} &= ((1-\gamma)(\mathbf{S}+\mathbf{C}+\delta\mathbf{\Gamma}) + \lambda\mathbf{A}_0^{-1})^{-1}\sharp_t(\gamma\mathbf{D}+\lambda\mathbf{A}_0) \\ &= \left((1-\gamma)(\mathbf{S}+\mathbf{C}+\delta\mathbf{\Gamma}) + \lambda\mathbf{A}_0^{-1}\right)^{1/2} \\ &\Big(\left((1-\gamma)(\mathbf{S}+\mathbf{C}+\delta\mathbf{\Gamma}) + \lambda\mathbf{A}_0^{-1}\right)^{-1/2}(\gamma\mathbf{D}+\lambda\mathbf{A}_0) \\ &\left((1-\gamma)(\mathbf{S}+\mathbf{C}+\delta\mathbf{\Gamma}) + \lambda\mathbf{A}_0^{-1}\right)^{-1/2}\Big)^t \\ &\left((1-\gamma)(\mathbf{S}+\mathbf{C}+\delta\mathbf{\Gamma}) + \lambda\mathbf{A}_0^{-1}\right)^{1/2}.\end{aligned} \tag{19}$$

Without loss of generality, in this paper we set $\mathbf{A}_0$ to be $(p+q)$-order identity matrix $\mathbf{I}_{p+q}$. After obtaining $\mathbf{A}$, we can recover $\mathbf{U}$ and $\mathbf{V}$ from (5). Then, for a test instance with cross-view representations $\mathbf{x}$ and $\mathbf{y}$, we can predict its class label by estimating on its concatenated representations $\mathbf{U}^T\mathbf{x} + \mathbf{V}^T\mathbf{y} = \mathbf{H}^T[\mathbf{x}^T, \mathbf{y}^T]^T$.

Complexity analysis on C-DSCA Since the objective function of C-DSCA enjoys a closed-form solution, so its complexity mainly lies in (19). It is composed of $\left((1-\gamma)(\mathbf{S}+\mathbf{C}+\delta\mathbf{\Gamma}) + \lambda\mathbf{A}_0^{-1}\right)^{1/2}$, $(\gamma\mathbf{D} + \lambda\mathbf{A}_0)$ and $\left((1-\gamma)(\mathbf{S}+\mathbf{C}+\delta\mathbf{\Gamma}) + \lambda\mathbf{A}_0^{-1}\right)^{-1/2}$. Specially, the complexity of both $\left((1-\gamma)(\mathbf{S}+\mathbf{C}+\delta\mathbf{\Gamma}) + \lambda\mathbf{A}_0^{-1}\right)^{1/2}$ and $\left((1-\gamma)(\mathbf{S}+\mathbf{C}+\delta\mathbf{\Gamma}) + \lambda\mathbf{A}_0^{-1}\right)^{-1/2}$ is $\mathcal{O}((p+q)^3)$. Consequently, the total complexity of C-DSCA of (19) is $\mathcal{O}((p+q)^3)$.

C. Comparison between DSCA and C-DSCA

On the one hand, the convexity of C-DSCA is superior to that of DSCA. The Hessian matrix of the DSCA objective in (9), i.e. $\mathbf{C} + \mathbf{S} - \lambda_1\mathbf{D} + \lambda_2\mathbf{\Gamma}$, is a differential combination of positive definite matrices, which may incur (9) to be nonconvex (or say conditionally convex) in the Euclidean distance space. By comparison, the objective of C-DSCA in (17) is a sum of linearly convex terms and conically convex term. As a result, (17) is entirely convex, which readily brings it with optimal solutions.

On the other hand, the solution space of C-DSCA is bigger than that of DSCA. For DSCA, we can see that the Hessian matrix (i.e. $\mathbf{C}+\mathbf{S}-\lambda_1\mathbf{D}+\lambda_2\mathbf{\Gamma}$) regarding $\mathbf{A}$ of (9) is a differential combination of positive definite matrices, resulting in its conditional-convexity. Although the solution of DSCA can be constructed through generalized eigen-decomposition on its conditionally positive definite Hessian matrix, the resulting metric $\mathbf{A}$ is spanned by orthogonal eigenvectors, which may be warped by the non-positive definiteness of its Hessian matrix. As a result, the solution space of DSCA and consequently its solution optimality will be limited. In contrast, the objective (17) of C-DSCA is extended from (9) by modeling in the geodesic space instead of the Euclidean space. In this way, the metric space is consequently transformed from the orthogonal eigenvectors to much greater convex positive-definite cone space. As a result, the solution space of C-DSCA is made to be bigger than that of DSCA.

IV. Experiment

In this section, we conducted comparative experiments on several multi-view databases to evaluate the proposed methods, i.e., DSCA and C-DSCA.

A. Setup

For a fair comparison, several most related methods were implemented, i.e., CCA [1], DCCA [8], MPECCA [10], DCA [17], CECCA [14], CDCA [29]. For the tradeoff parameters involved in the compared methods, they were tuned through five-fold cross-validation following the related references. For fused cross-view representations, the k-nearest-neighbors (kNN) classifier was applied for final decision making with $k=3$. To evaluate the generalization ability of the proposed method, we conducted experiments on both non-face datasets and face datasets in the following subsections, for which accuracy (%) and mean absolute error (MAE) were adopted as performance measures, respectively.

B. Experimental results on non-face datasets

We first performed experiments on several widely used non-face multi-view datasets, i.e., handwritten digit databases MFD [46] and USPS [47], animal dataset AWA [48] and Alzheimer's Disease dataset ADNI [49]. Details about the datasets are given in Table II. On each of these datasets, we randomly took half of the samples for model training and rest for testing. And, we report the averaged results over ten runs with random data partitions in Table III.

We can reach the following observations from the results. In nearly all cases, the proposed DSCA method yielded the second highest recognition accuracies, only slightly lower than the highest results of C-DSCA, demonstrating the usefulness of our modelling schema in preserving the semantic consistency in cross-view correlation learning. More importantly, the proposed C-DSCA method achieved the highest recognition accuracies in most cases. Specially, the average improvement of C-DSCA method was significant, especially on AWA with average about 10% accuracy improvement over the best of the compared methods.

C. Experimental results on face datasets

We also conducted experiments on four challenging real-world, large-scale face datasets, i.e., AgeDB [50], Morph (Album I and II) [51], FERET [52], [53], and the Cross-Age Celebrity Dataset (CACD) [54]. Specifically, the AgeDB database contains 16,516 face images from 570 subjects and the images are annotated with accurate age, noise-free labels from 0 to over 100 years. As for the Morph dataset, the Album I contains 1,690 images taken from 631 African and European persons, aged from 15 to about 68 years, and the Album II consists of over 55,000 face images, aged from 16 to 77 years. For the FERET dataset, it contains over 11,000 images from over 900 subjects, from Asian, Hispanic, Caucasian, Melanoderm and other races. For the CACD dataset, it consists of about 163,446 face images from 2,000 celebrities aged from 0 to over 100 years. Image examples from the four databases are illustrated in Figure 3.

In the experiments, we extracted BIF [55] and HoG [56] features from these databases and respectively reduced their dimensions to 200 by PCA technique as two view representations for cross-view facial attributes estimation. Specially, we performed facial age estimation on the AgeDB and CACD datasets, facial gender classification on AgeDB and Morph II, and facial race recognition on the Morph I and FERET databases, respectively. For performance measure, we adopted the MAE on age estimation and recognition accuracy on both gender and race recognition.

1) Age estimation: To comprehensively evaluate the age estimation performance, we randomly chose 50, 100, 150 samples for training while the remainder for test from AgeDB, and 500, 1000, 1500 samples for training while the rest for test from CACD. We ran the experiments ten times with random data partitions and report the averaged results in Tables IV and V. From them, we can observe that with increased samples for training, the estimation errors (MAEs) of all the methods reduced monotonically. Moreover, the result magnitudes on CACD are lower than that on AgeDB. These validate that training with more adequate samples could improve the generalization ability of the age estimator. Moreover, the age MAEs of DSCA were generally the second lowest among all the methods. It demonstrates the rationality of preserving the semantic consistency in cross-view learning. More importantly, we also find that in all cases our proposed method C-DSCA yielded the lowest estimation MAEs (*the lower the better*) among all the compared methods, demonstrating the effectiveness and superiority of our proposed method in handling facial age estimation task. Furthermore, the performance improvements of C-DSCA over DSCA validate the superiority of modelling in the geodesic space over the Euclidean space.

2) Gender classification: For gender classification, we randomly chose 1000, 2000, 3000 samples for training while the remainder for test from both the AgeDB and Morph Album II datasets. We also ran the experiments ten times with random data partitions and report the averaged results in Tables VI and VII. We can observe that increased training samples resulted in higher gender recognition accuracies for all the compared methods. Besides, as an unsupervised model, the

TABLE II: Profiles of the MFD, AWA, ADNI and USPS datasets.

Dataset	*View Representations (# Dimension)*	*# Classes*	*# Samples*
MFD	fou(76), fac(216), kar(64), pix(240), zer(47), mor(6)	10	20000
AWA	cq(2688), lss(2000), phog(252), rgsift(2000), sift(2000), surf(2000)	2	879816
ADNI	VBM(116), FDG(116), AV(116)	2	211160
USPS	left(128), right(128)	10	11000

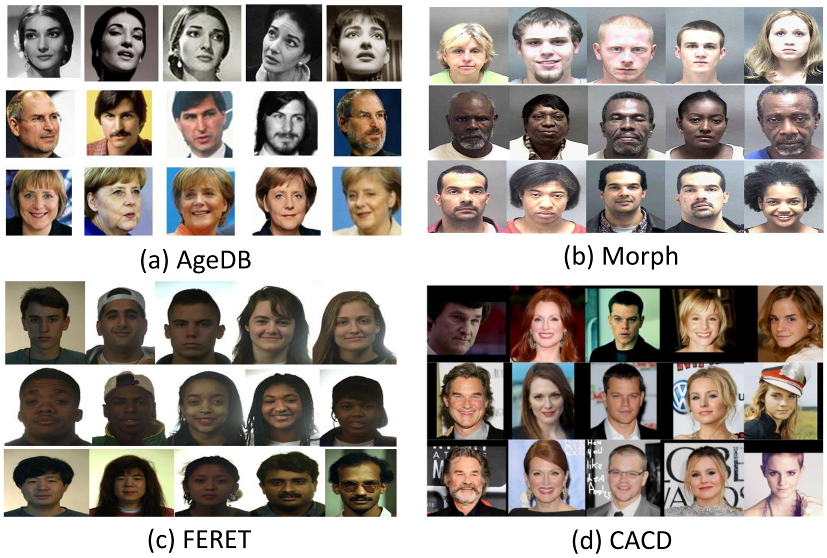
TABLE III: Recognition accuracy (%) of the methods on non-face datasets.

Dataset	View Combination		*CCA*	*DCA*	*MPECCA*	*DCCA*	*CECCA*	*CDCA*	*DSCA (ours)*	*C-DSCA (ours)*
MFD	fac	fou	80.22±0.9	80.00±0.2	90.64±1.3	95.15±0.9	96.46±2.4	97.27±0.4	97.33±0.3	**97.54±0.2**
	fac	kar	92.12±0.5	90.10±0.8	95.39±0.6	95.33±0.7	96.52±1.2	97.08±0.6	97.00±0.4	**97.31±0.4**
	fac	mor	78.22±0.8	63.22±4.3	72.32±2.4	95.22±0.9	94.23±1.0	**97.03±0.2**	95.83±0.3	96.90±0.3
	fac	pix	83.02±1.2	90.20±0.5	94.65±0.5	65.60±1.1	93.67±2.9	96.38±0.6	96.44±0.5	**96.85±0.3**
	fac	zer	84.00±0.6	71.50±2.2	93.79±0.7	96.00±0.6	97.04±0.6	97.25±0.2	97.25±0.3	**97.56±0.3**
	fou	kar	90.11±1.0	75.42±5.6	93.98±0.4	89.12±4.3	96.90±0.5	96.47±0.7	96.88±0.5	**97.80±0.6**
	fou	mor	70.22±0.4	55.82±4.6	60.62±1.6	82.30±0.9	78.25±0.6	**82.45±0.4**	80.13±0.7	79.22±0.8
	fou	pix	68.44±0.4	76.10±4.7	78.24±1.1	90.41±3.2	76.28±1.3	95.47±0.5	96.50±0.3	**97.13±0.4**
	fou	zer	74.10±0.9	62.80±4.1	79.38±1.2	79.53±4.5	83.16±1.4	86.02±0.8	84.17±0.6	**86.85±0.5**
	kar	mor	64.09±0.6	82.00±1.6	72.92±2.7	91.95±2.8	91.89±0.6	**96.35±0.2**	95.27±0.4	95.44±0.5
	kar	pix	88.37±0.9	88.85±0.8	95.07±0.6	92.59±2.0	**95.98±0.3**	93.66±0.9	95.48±0.4	95.00±0.4
	kar	zer	90.77±1.0	75.97±2.8	94.17±0.6	88.47±2.9	93.57±0.9	96.22±0.3	96.42±0.4	**96.73±0.5**
	mor	pix	68.66±1.5	82.01±2.1	67.21±2.3	93.04±0.7	90.08±1.0	95.72±0.4	95.70±0.3	**95.99±0.5**
	mor	zer	73.22±0.6	50.35±1.8	60.95±1.4	84.55±0.9	80.59±0.9	**84.70±0.4**	83.23±0.4	83.28±0.6
	pix	zer	82.46±0.6	71.16±2.8	82.81±1.2	91.67±2.1	91.81±1.2	95.20±0.4	95.64±0.4	**95.84±0.4**
AWA	cq	lss	73.11±2.1	62.08±0.3	76.19±1.0	70.51±1.3	77.53±1.7	72.31±2.3	78.44±1.3	**82.44±1.5**
	cq	phog	65.21±1.4	73.10±1.2	72.42±1.6	70.15±0.9	74.51±2.1	71.22±2.4	74.31±1.3	**85.44±1.6**
	cq	rgsift	60.22±1.3	61.40±1.7	78.04±1.3	82.87±2.4	82.83±1.4	85.15±0.5	89.67±0.5	**92.34±1.2**
	cq	sift	74.33±1.3	61.28±1.9	77.85±1.4	83.19±2.1	80.05±1.7	74.55±1.6	80.42±1.5	**85.22±1.4**
	cq	surf	75.86±1.7	69.30±2.1	79.07±0.8	73.55±2.3	81.59±1.5	89.15±1.7	89.79±1.7	**92.92±0.8**
	lss	phog	69.96±1.7	59.72±0.2	68.12±1.2	64.86±2.6	71.36±1.4	66.58±2.1	71.13±1.7	**83.77±0.9**
	lss	rgsift	78.65±0.9	63.21±1.3	73.64±1.0	78.28±2.8	77.28±1.4	74.02±2.2	88.02±1.9	**88.68±0.9**
	lss	sift	73.49±1.0	65.72±2.1	73.12±1.4	66.21±1.6	76.69±1.7	65.29±2.6	74.43±1.6	**79.93±1.3**
	lss	surf	76.30±1.4	65.33±1.8	74.84±1.6	79.06±2.8	78.52±1.3	78.33±2.8	78.40±1.6	**86.43±0.8**
	phog	rgsift	68.18±1.1	48.38±1.0	69.49±2.3	77.37±1.5	74.41±1.5	63.57±1.6	78.77±1.7	**83.16±1.0**
	phog	sift	68.26±1.1	70.24±1.1	68.97±1.3	63.16±1.3	72.14±1.5	61.10±2.3	62.40±1.8	**83.67±1.0**
	phog	surf	64.57±1.4	56.94±0.5	71.55±1.4	75.68±1.9	74.43±2.1	67.05±1.7	82.17±1.7	**87.16±0.7**
	rgsift	sift	71.35±1.3	58.56±2.3	72.85±1.1	75.28±2.5	76.69±1.7	76.54±2.2	83.43±1.3	**91.83±0.9**
	rgsift	surf	75.55±1.3	67.22±1.6	76.94±2.2	84.10±2.4	80.46±1.7	84.34±3.3	86.34±2.0	**91.01±0.9**
	sift	surf	75.33±1.3	63.36±1.6	74.27±1.2	82.14±2.7	75.51±1.1	85.72±2.2	84.85±1.9	**88.46±0.8**
ADNI	AV	FDG	65.47±1.8	73.28±2.1	75.28±2.6	76.25±2.1	76.26±2.5	74.38±6.0	75.86±4.8	**81.05±2.8**
	AV	VBM	71.02±2.4	71.02±2.8	73.24±3.1	63.47±2.1	60.67±2.7	71.22±5.3	73.42±5.1	**75.58±2.5**
	FDG	VBM	61.37±1.2	65.28±1.6	70.37±2.6	64.05±1.6	70.95±1.8	80.98±3.5	81.27±4.0	**86.55±3.8**
USPS	left	right	62.14±0.6	80.11±1.2	66.67±0.9	63.96±2.0	82.89±1.9	89.73±0.7	90.67±0.6	**91.06±0.5**

TABLE IV: Age estimation results (MAE±STD) on AgeDB (↓). Hereinafter, ↓ means *lower is better*.

#training samples from each class	CCA	DCA	MPECCA	DCCA	CECCA	CDCA	DSCA (ours)	C-DSCA (ours)
50	17.70±5.1	17.78±5.2	16.10±4.2	15.93±3.5	15.62±4.6	16.00±5.0	15.30±4.1	**14.11±4.2**
100	16.81±4.5	17.21±5.5	13.42±4.5	13.77±4.6	13.35±4.4	13.20±4.1	13.17±4.5	**12.20±4.1**
150	15.43±5.4	16.25±6.1	12.82±5.0	13.31±3.5	12.20±4.2	12.90±5.5	12.78±4.2	**11.89±3.5**

baseline CCA yielded the lowest accuracy in most cases. This demonstrates that modelling with supervision information can improve the discriminating ability of the gender classifiers. More importantly, by making use of the discriminative label


(a) AgeDB

(b) Morph

(c) FERET

(d) CACD

Fig. 3: Face examples from the (a) AgeDB, (b) Morph, (c) FERET and (d) CACD datasets.

TABLE V: Age estimation results (MAE±STD) on CACD (↓).

#training samples from each class	CCA	DCA	MPECCA	DCCA	CECCA	CDCA	DSCA (ours)	C-DSCA (ours)
500	14.50±5.0	15.00±4.3	14.01±4.2	13.20±3.59	12.60±4.3	12.81±5.8	12.40±3.4	**11.51±3.6**
1000	13.43±4.3	14.50±5.3	13.02±4.5	12.42±4.5	12.35±4.4	11.90±4.6	11.80±3.4	**11.20±3.6**
1500	13.21±5.2	13.42±5.3	11.80±5.0	11.51±3.5	11.20±4.2	11.80±4.6	11.60±3.1	**10.29±3.2**

TABLE VI: Gender classification accuracy (Acc±STD) on AgeDB (↑). Hereinafter, ↑ indicates *higher is better*.

#training samples from each class	CCA	DCA	MPECCA	DCCA	CECCA	CDCA	DSCA (ours)	C-DSCA (ours)
1000	74.51±0.6	75.75±0.6	76.80±0.7	74.76±0.6	79.25±0.6	82.50±4.9	86.59±0.9	**88.33±0.2**
2000	76.98±0.6	75.95±0.6	79.51±0.8	77.18±0.5	80.31±0.6	83.84±3.6	88.12±0.7	**89.45±0.2**
3000	78.21±0.4	76.33±0.7	79.69±0.7	78.51±0.6	80.67±0.7	84.52±3.5	88.40±0.7	**89.69±0.3**

TABLE VII: Gender classification accuracy (Acc±STD) on Morph Album II (↑).

#training samples from each class	CCA	DCA	MPECCA	DCCA	CECCA	CDCA	DSCA (ours)	C-DSCA (ours)
1000	67.86±0.8	74.46±0.7	75.10±0.6	71.59±0.7	78.90±0.8	80.53±0.2	83.64±0.3	**87.65±0.3**
2000	69.54±0.8	75.32±0.7	76.39±0.7	72.86±1.0	79.12±0.6	85.16±0.3	86.43±0.2	**87.97±0.4**
3000	69.74±0.7	77.85±0.7	78.58±0.7	74.22±0.7	80.31±0.6	86.36±0.4	87.28±0.3	**88.13±0.4**

information, the cross-view correlation information, and more importantly the cross-view semantic consistency information, as well as modelling in the Riemannian geometry space with closed-form solution, the proposed C-DSCA method yielded the remarkably highest gender accuracy in all cases among all the methods, with average about 5% accuracy improvement on the two gender estimation datasets over the best of compared method CDCA. It also demonstrates the effectiveness and superiority of our modelling scheme in dealing with gender recognition tasks.

3) Race recognition: For race recognition evaluation, we randomly chose 100, 150, 200 samples from Caucasian and Melanoderm races for training while the remainder of these two races for test from Morph Album I, and 100, 200, 300 samples from Asian, Hispanic, Caucasian, Melanoderm races for training while the rest of these four races for test from the FERET database. Similar to the above age and gender recognition settings, we also ran the evaluations ten times with random data partitions and report the averaged results in Tables VIII and IX. We can observe that for the task of race estimation, with increased training samples for each race, the recognition accuracy increased consistently across all the methods, indicating that more training samples would benefit race recognition. Moreover, similar to the above age and gender evaluations, the baseline model CCA yielded the lowest accuracies (*the higher the better*) among all the methods. It illustrates that making use of the supervision information is also desired to race recognition. On the other hand, among all the methods, our proposed C-DSCA method yielded the highest accuracies in all cases, with about 6% and 6.5% accu-

TABLE VIII: Race recognition accuracy (Acc±STD) on Morph Album I (↑).

#training samples from each class	CCA	DCA	MPECCA	DCCA	CECCA	CDCA	DSCA (ours)	C-DSCA (ours)
100	63.73±2.3	66.74±0.7	68.50±0.7	68.63±0.6	72.31±0.6	77.95±0.3	80.65±0.6	**86.32±0.4**
150	65.18±2.6	67.28±0.8	69.89±0.6	69.75±1.0	73.05±0.7	80.92±1.4	82.13±0.6	**87.41±0.5**
200	66.74±2.5	67.88±0.6	71.27±0.8	70.13±0.8	76.80±0.7	83.40±1.3	82.94±0.4	**87.85±0.4**

TABLE IX: Race recognition accuracy (Acc±STD) on FERET (↑).

#training samples from each class	CCA	DCA	MPECCA	DCCA	CECCA	CDCA	DSCA (ours)	C-DSCA (ours)
100	50.99±2.0	51.22±1.8	54.02±1.3	53.89±1.9	55.16±1.6	57.66±1.8	60.39±1.2	**66.77±0.1**
200	51.48±2.7	52.19±1.8	54.89±1.7	54.00±1.9	56.24±1.5	62.43±1.8	64.10±1.4	**68.92±0.1**
300	51.84±2.4	53.82±1.5	55.34±1.3	54.12±2.0	58.62±1.6	64.91±1.4	64.73±1.2	**69.21±0.1**

racy improvements over the best compared method on Morph Album I and FERET, respectively. It demonstrates again the effectiveness and superiority of our modeling scheme. Last but not least, we can observe that the race recognition accuracies on FERET are much lower than that on Morph. It is because that four races were chosen for evaluation on FERET while only two races on the Morph dataset.

D. Parameters analysis

In order to comprehensively explore the proposed model C-DSCA, we also performed parameter analysis on the geometric weighting parameter t, the metric balance parameter γ, the metric prior parameter λ, and the semantic consistency parameter δ involved in (18), respectively. Without loss of generality, we performed gender recognition on AgeDB by randomly choosing 2000 samples per class for training while the rest for test, race recognition on Morph Album I by randomly choosing 150 samples per class for training while the rest for test, and age estimation on CACD by randomly choosing 1000 samples per class for training while the rest for test, respectively, and repeated ten times on each dataset with random data partitions. Consequently, the evaluation results are demonstrated in Figures 4 to 7.

Geometric weighting parameter t of C-DSCA: For the geometric weighting parameter t, we can observe some interesting rules from Figure 4. First, a generally similar performance rule is shared by all the tasks. That is, with t increasing from 0 to 1, their performance accuracies increased (error decreased) first and then decreased (error increased), and achieved the best performance around $t = 0.5$. It means that not only the sample similarities within the same class but also their dissimilarities between different classes are desired to be modelled for the geometric mean solution (as formulated in (18)).

Metric balance parameter γ of C-DSCA: From the results shown in Figure 5, we can observe that with increasing metric balance parameter γ, the evaluation performances on gender, race and age estimation tasks first improved gradually and then became worse. Furthermore, the performance was not so sensitive around $0.3 < \gamma < 0.7$. This observation illustrates that the metric components regarding the cross-view correlation, the intra-class similarity and inter-class dissimilarity, as well as the cross-view semantic consistency are simultaneously crucial to the cross-view metric learning.

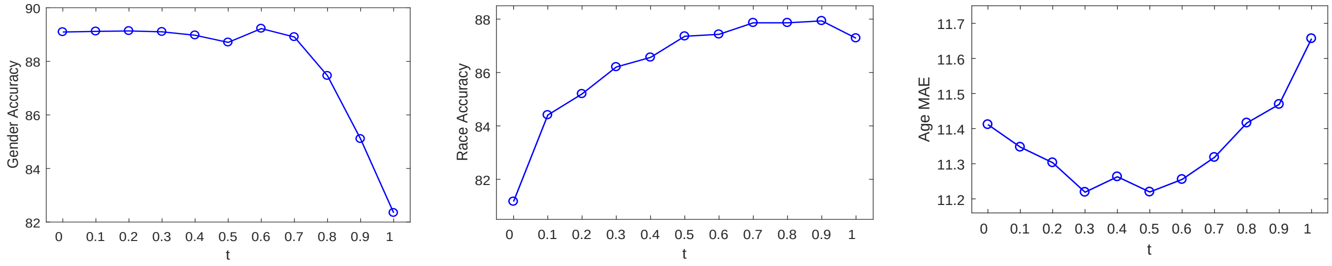
Metric prior parameter λ of C-DSCA: Regarding the metric prior parameter λ, the evaluation performances on all the tasks kept stable when $\lambda <$ 1e-4, but got worse when $\lambda >$ 1e-2. By recalling (18), we can see that λ controls the contribution degree of the metric prior $\mathbf{A}_0$ to the metric to be learned. These results in Figure 6 show that proper extent rather than too large of metric prior is desirable, because the desired metric is mostly guided by the training data.

Semantic consistency parameter δ of C-DSCA: For the parameter δ (see (18)), it plays the role of controlling a tradeoff between the cross-view semantic consistency and the correlation and similarities across views. As the results in Figure 7 showing, the evaluation performance got better with increased extent of semantic consistency at first, but then gradually became worse when δ is too large. This observation shows that not only the cross-view semantic consistency is desirable for metric modelling, but the cross-view correlation and similarities are also necessary.

Metric rank of DSCA: By recalling (9), we can observe that the solution of DSCA is composed of certain number of singular vectors of its objective Hessian matrix. That is, the rank of the metric matrix $\mathbf{A}$ is proportional to the number of the singular vectors. Therefore, to explore the performance of DSCA regarding the rank of the metric matrix, we conducted evaluations on the AgeDB, Morph Album I and CACD datasets for age, gender and race estimations with the same setup as the above parameter analysis experiments. The results are shown in Figure 8. We can observe that with increased rank of $\mathbf{A}$), the performance of DSCA improved, demonstrating that more dimensions (ranks) of the metric space would benefit the subsequent cross-view recognition.

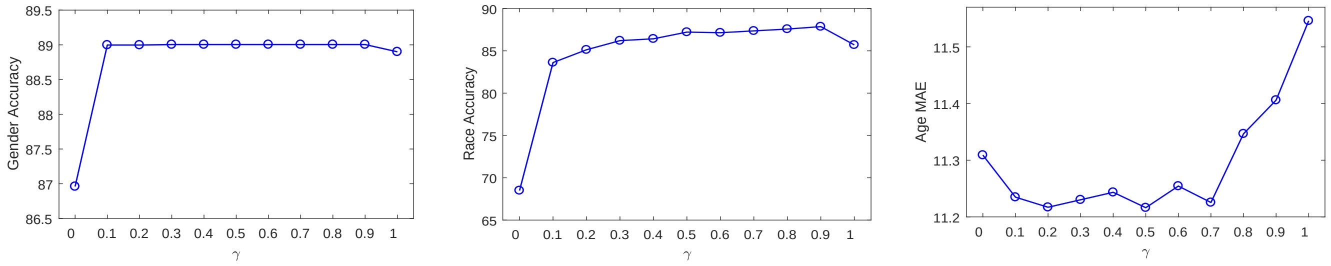
V. Conclusion

For cross-view recognition tasks, in this paper we proposed a Discriminant Semantic Correlation Analysis (DSCA) method by modelling the cross-view semantic consistency for each object in the sample space rather than in the commonly used feature space. Then, to enhance the nonlinearly discriminating ability of DSCA, we further extended it from the Euclidean to



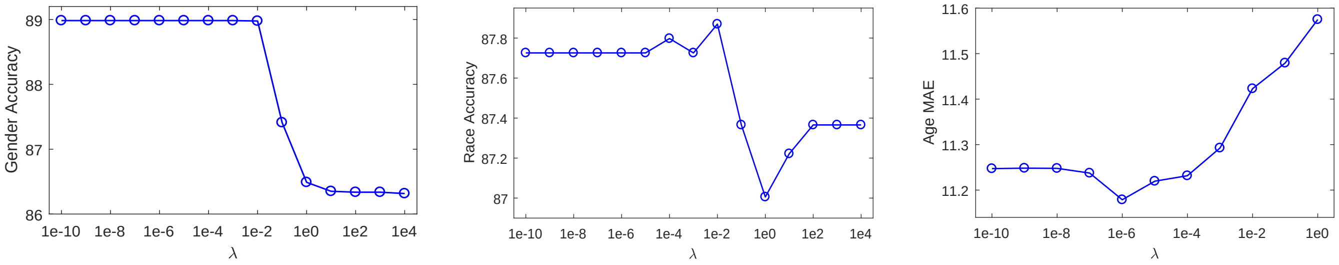
(a) Gender recognition on AgeDB (↑) (b) Race recognition on Morph Album I (↑) (c) Age estimation on CACD (↓)

Fig. 4: Performance of the proposed C-DSCA method regarding the geometric weighting parameter t.



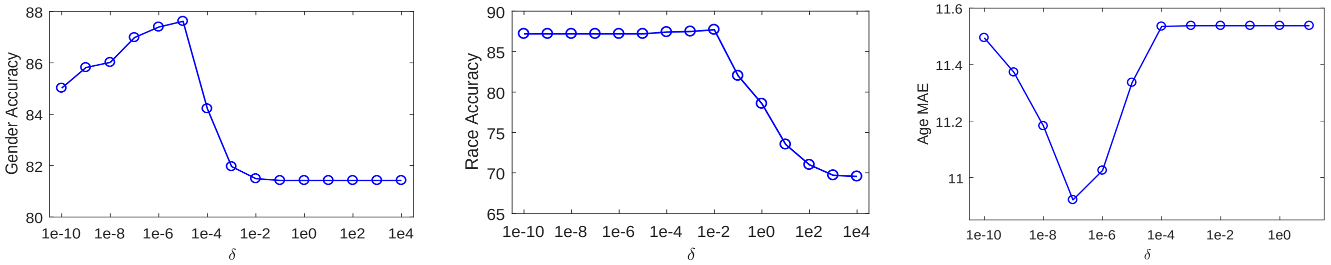
(a) Gender recognition on AgeDB (↑) (b) Race recognition on Morph Album I (↑) (c) Age estimation on CACD (↓)

Fig. 5: Performance of the proposed C-DSCA method regarding the metric balance parameter γ.



(a) Gender recognition on AgeDB (↑) (b) Race recognition on Morph Album I (↑) (c) Age estimation on CACD (↓)

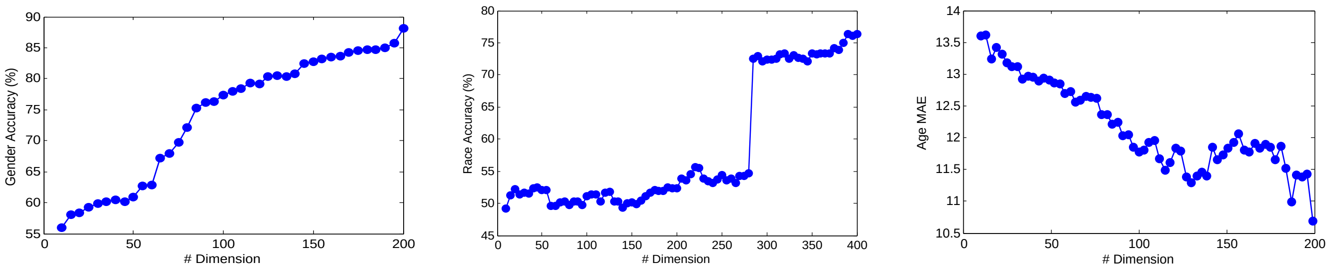
Fig. 6: Performance of the proposed C-DSCA method regarding the metric prior parameter λ.



(a) Gender recognition on AgeDB (↑) (b) Race recognition on Morph Album I (↑) (c) Age estimation on CACD (↓)

Fig. 7: Performance of the proposed C-DSCA method regarding the semantic consistency parameter δ.

the geodesic space by transforming the metric and incorporating both the cross-view semantic and representation correlation information and consequently obtained the Convex DSCA (C-DSCA), which enjoys closed-form solution in the geodesic metric space. Finally, we conducted extensive experiments to validate the effectiveness of the proposed methods and performed parameters analysis. In the future, we will consider to extend the models with deep network architectures [57].



(a) Gender recognition on AgeDB (↑)

(b) Race recognition on Morph Album I (↑)

(c) Age estimation on CACD (↓)

Fig. 8: Performance of the proposed DSCA method regarding the metric rank.

Acknowledgment

This work was partially supported by the National Natural Science Foundation of China under Grants 61702273 and 61672281, the Natural Science Foundation of Jiangsu Province under Grant BK20170956, the Natural Science Foundation of the Jiangsu Higher Education Institutions of China under Grant 17KJB520022, the Open Projects Program of National Laboratory of Pattern Recognition under Grant 202000007, the Fundamental Research Funds for the Central Universities under Grant NJ2019010, the Priority Academic Program Development of Jiangsu Higher Education Institutions (PAPD) Fund. Besides, portions of the research in this paper used the FERET database of facial images collected under the FERET program, sponsored by the DOD Counterdrug Technology Development Program Office.

Appendix: Proof of Proposition 1

Proof. Let us assume

$$\mathbf{A}^* = \arg\max_{\mathbf{A}\succ 0} tr(\mathbf{AD}). \tag{20}$$

Then, we have

$$tr(\mathbf{A}^*\mathbf{D}) > tr(\mathbf{AD}) \tag{21}$$

and

$$\mathbf{A}^* \succ \mathbf{A}. \tag{22}$$

Performing singular values decomposition (SVD) on $\mathbf{A}^*$ and $\mathbf{A}$ yields

$$\mathbf{A}^* = \mathbf{U}^*\Sigma^*(\mathbf{V}^*)^T = \sum_{i=1}^{d} \sigma_i^* \mathbf{u}_i^* (\mathbf{v}_i^*)^T,\ \sigma_i^* > 0 \tag{23}$$

and

$$\mathbf{A} = \mathbf{U}\Sigma\mathbf{V}^T = \sum_{i=1}^{d} \sigma_i \mathbf{u}_i \mathbf{v}_i^T,\ \sigma_i > 0. \tag{24}$$

Based on (22)-(24), we have

$$\sum_{i=1}^{d} \sigma_i^* > \sum_{i=1}^{d} \sigma_i, \tag{25}$$

meaning that

$$\sum_{i=1}^{d} \frac{1}{\sigma_i^*} < \sum_{i=1}^{d} \frac{1}{\sigma_i}. \tag{26}$$

As a result,

$$(\mathbf{A}^*)^{-1} = \sum_{i=1}^{d} \frac{1}{\sigma_i^*} \mathbf{u}_i^* (\mathbf{v}_i^*)^T \prec \mathbf{A}^{-1} = \sum_{i=1}^{d} \frac{1}{\sigma_i} \mathbf{u}_i \mathbf{v}_i^T, \tag{27}$$

which implies that

$$tr((\mathbf{A}^*)^{-1}\mathbf{D}) < tr(\mathbf{A}^{-1}\mathbf{D}). \tag{28}$$

That is,

$$\mathbf{A}^* = \arg\min_{\mathbf{A}\succ 0} tr(\mathbf{A}^{-1}\mathbf{D}). \tag{29}$$

Combing (20) and (29) results in the desired conclusion

$$\max_{\mathbf{A}\succ 0} tr(\mathbf{AD}) \Leftrightarrow \min_{\mathbf{A}\succ 0} tr(\mathbf{A}^{-1}\mathbf{D}). \tag{30}$$

□

References

[1] P. L. Lai and C. Fyfe, "Kernel and nonlinear canonical correlation analysis." *International Journal of Neural Systems*, vol. 10, no. 5, pp. 365–377, 2000.

[2] D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor, "Canonical correlation analysis: an overview with application to learning methods," *Neural Computation*, vol. 16, no. 12, pp. 2639–2664, 2004.

[3] V. Sindhwani and D. S. Rosenberg, "An rkhs for multi-view learning and manifold co-regularization," in *International Conference on Machine Learning*, 2008, pp. 976–983.

[4] L. Bazzani and V. Murino, "A unifying framework for vector-valued manifold regularization and multi-view learning," in *International Conference on Machine Learning*, 2013, pp. 92–100.

[5] J. Zhao, X. Xie, X. Xu, and S. Sun, "Multi-view learning overview: Recent progress and new challenges," *Information Fusion*, vol. 38, pp. 43–54, 2017.

[6] G. Zhang, H. Sun, F. Porikli, Y. Liu, and Q. Sun, "Optimal couple projections for domain adaptive sparse representation-based classification," *IEEE Transactions on Image Processing*, vol. 26, no. 12, pp. 5922–5935, 2017.

[7] X. Gao, S. Niu, and Q. Sun, "Two-directional two-dimensional kernel canonical correlation analysis," *IEEE Signal Processing Letters*, vol. 26, no. 11, pp. 1578–1582, 2019.

[8] T. Sun, S. Chen, J. Yang, and P. Shi, "A novel method of combined feature extraction for recognition," in *IEEE International Conference on Data Mining*, 2008, pp. 1043–1048.

[9] Y. Peng, D. Zhang, and J. Zhang, "A new canonical correlation analysis algorithm with local discrimination," *Neural Processing Letters*, vol. 31, no. 1, pp. 1–15, 2010.

[10] S. Su, H. Ge, and Y. H. Yuan, "Multi-patch embedding canonical correlation analysis for multi-view feature learning," *Journal of Visual Communication and Image Representation*, vol. 41, pp. 47–57, 2016.

[11] Q. S. Sun, Z. D. Liu, P. A. Heng, and D. S. Xia, "Rapid and brief communication: A theorem on the generalized canonical projective vectors," *Pattern Recognition*, vol. 38, no. 3, pp. 449–452, 2005.

[12] H. K. Ji, Q. S. Sun, Y. H. Yuan, and Z. X. Ji, "Fractional-order embedding supervised canonical correlations analysis with applications to feature extraction and recognition," *Neural Processing Letters*, vol. 45, no. 1, pp. 1–19, 2016.

[13] X. Shen, W. Liu, I. W. Tsang, Q.-S. Sun, and Y.-S. Ong, "Multilabel prediction via cross-view search," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 29, no. 9, pp. 4324–4338, 2018.

[14] X. D. Zhou, X. H. Chen, and S. C. Chen, "Combined-feature-discriminability enhanced canonical correlation analysis," *Pattern Recognition and Artificial Intelligence*, vol. 25, no. 2, pp. 285–291, 2012.

[15] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, "Eigenfaces vs. fisherfaces: Recognition using class specific linear projection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 711–720, 2002.

[16] F. Zhao, L. Qiao, F. Shi, P. T. Yap, and D. Shen, "Feature fusion via hierarchical supervised local cca for diagnosis of autism spectrum disorder." *Brain Imaging and Behavior*, vol. 11, no. 4, pp. 1–11, 2016.

[17] M. Haghighat, M. Abdel-Mottaleb, and W. Alhalabi, *Discriminant Correlation Analysis: Real-Time Feature Level Fusion for Multimodal Biometric Recognition*. IEEE Press, 2016.

[18] A. Sharma, A. Kumar, H. Daume, and D. W. Jacobs, "Generalized multiview analysis: A discriminative latent space," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 2160–2167.

[19] X. B. Shen, Q. Sun, and Y. Yuan, "A unified multiset canonical correlation analysis framework based on graph embedding for multiple feature extraction," *Neurocomputing*, vol. 148, pp. 397–408, 2015.

[20] S. Sun, X. Xie, and M. Yang, "Multiview uncorrelated discriminant analysis," *IEEE Transactions on Cybernetics*, vol. 46, no. 12, pp. 3272–3284, 2016.

[21] X. Shen, F. Shen, Q. S. Sun, Y. Yang, and H. T. Shen, "Semi-paired discrete hashing: Learning latent hash codes for semi-paired cross-view retrieval," *IEEE Transactions on Cybernetics*, vol. 47, no. 12, pp. 4275–4288, 2017.

[22] H. Peng, D. Peng, J. Guo, and L. Zhen, "Local feature based multi-view discriminant analysis," *Knowledge-Based Systems*, vol. 149, pp. 34–46, 2018.

[23] G. Andrew, R. Arora, J. Bilmes, and K. Livescu, "Deep canonical correlation analysis," in *International Conference on Machine Learning*, 2013, pp. 1247–1255.

[24] X. Chang, T. Xiang, and T. M. Hospedales, "Scalable and effective deep cca via soft decorrelation," in *International Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1488–1497.

[25] Y. Liu, Y. Li, and Y. H. Yuan, "A complete canonical correlation analysis for multiview learning," in *International Conference on Image Processing*, 2018, pp. 3254–3258.

[26] X.-M. Dai and S.-G. Li, "Cross-modal deep discriminant analysis," *Neurocomputing*, vol. 314, pp. 437–444, 2018.

[27] D. Y. Gao, *Canonical Duality Theory and Solutions to Constrained Nonconvex Quadratic Programming*. Kluwer Academic Publishers, 2004.

[28] X. Fu, K. Huang, M. Hong, N. D. Sidiropoulos, and M. C. So, "Scalable and flexible multiview max-var canonical correlation analysis," *IEEE Transactions on Signal Processing*, vol. 65, no. 16, pp. 4150–4165, 2017.

[29] F. Jiang and S. Chen, "Convex discriminant canonical correlation analysis," *Pattern Recognition and Artifical Intelligence*, vol. 30, no. 08, pp. 70–76, 2017.

[30] P. H. Zadeh, R. Hosseini, and S. Sra, "Geometric mean metric learning," in *IEEE International Conference on Machine Leaaning*, 2016, pp. 2464–2471.

[31] H. Wang, L. Feng, and Y. Liu, "Metric learning with geometric mean for similarities measurement," *Soft Computing*, vol. 20, no. 10, pp. 3969–3979, 2016.

[32] P. L. Lai and C. FyFe, "Kernel and nonlinear canonical correlation analysis," *International Journal of Neural Systems*, vol. 10, no. 05, pp. 365–377, 2000.

[33] K. Fukumizu, F. R. Ism., AC. JPBach, A. Mines. Orggretton, and D. E. TUEBINGEN. MPG., "Statistical consistency of kernel canonical correlation analysis," *Journal of Machine Learning Research*, vol. 8, no. 2007, pp. 361–383, 2007.

[34] O. Chapelle, "Training a support vector machine in the primal," *Neural Computation*, vol. 19, no. 5, pp. 1155–1178, 2007.

[35] A. Argyriou, C. Micchelli, and M. Pontil, "When is there a representer theorem? vector versus matrix regularizers," *Journal of Machine Learning Research*, vol. 10, no. 4, pp. 2507–2529, 2008.

[36] P. D. Tao, H. M. Le, H. A. L. Thi, and F. Lauer, "A difference of convex functions algorithm for switched linear regression," *IEEE Transactions on Automatic Control*, vol. 59, no. 8, pp. 2277–2282, 2014.

[37] T. Liu and T. K. Pong, "Further properties of the forwardcbackward envelope with applications to difference-of-convex programming," *Computational Optimization and Applications*, vol. 67, no. 3, pp. 489–520, 2017.

[38] S. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge University Press, 2004.

[39] T. Rapcsk, "Geodesic convexity in nonlinear optimization," *Journal of Optimization Theory and Applications*, vol. 69, no. 1, pp. 169–183, 1991.

[40] ARSIGNY, Vincent, FILLARD, Pierre, PENNEC, Xavier, AYACHE, and Nicholas, "Geometric means in a novel vector space structure on symmetric positive-definite matrices," *Siam Journal on Matrix Analysis and Applications*, vol. 29, no. 1, pp. 328–347, 2011.

[41] A. Papadopoulos, *Metric Spaces, Convexity and Nonpositive Curvature*. European Mathematical Society, 2014.

[42] W. M. Wonham, "On a matrix riccati equation of stochastic control," *Siam Journal on Control and Optimization*, vol. 6, no. 4, pp. 681–697, 1968.

[43] J. R. Cloutier, "State-dependent riccati equation techniques: an overview," in *American Control Conference, 1997. Proceedings of the*, 1997, pp. 932–936 vol.2.

[44] R. Bhatia, *Positive definite matrices*. Princeton University Press, 2007.

[45] C. Anoop, S. Suvrit, B. Arindam, and P. Nikolaos, "Jensen-bregman logdet divergence with application to efficient similarity search for covariance matrices," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 9, pp. 2161–2174, 2013.

[46] C. L. Liu, K. Nakashima, H. Sako, and H. Fujisawa, "Handwritten digit recognition: investigation of normalization and feature extraction techniques," *Pattern Recognition*, vol. 37, no. 2, pp. 265–279, 2004.

[47] T. F. Pawlicki, D. S. Lee, J. J. Hull, and S. N. Srihari, "Neural network models and their application to handwritten digit recognition," in *IEEE International Conference on Neural Networks*, 2002, pp. 63–70.

[48] C. H. Lampert, H. Nickisch, and S. Harmeling, "Learning to detect unseen object classes by between-class attribute transfer," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 951–958.

[49] C. R. Jack, M. A. Bernstein, N. C. Fox, T. Paul, H. Danielle, B. Bret, P. J. Britson, L. W. Jennifer, and W. Chadwick, "The alzheimer's disease neuroimaging initiative (adni): Mri methods." *Alzheimers and Dementia the Journal of the Alzheimers Association*, vol. 27, no. 4, pp. 685–691, 2010.

[50] S. Moschoglou, A. Papaioannou, C. Sagonas, J. Deng, and S. Zafeiriou, "Agedb: The first manually collected, in-the-wild age database," in *IEEE International Conference on Computer Vision and Pattern Recognition Workshops*, 2017, pp. 1997–2005.

[51] K. Ricanek and T. Tesafaye, "Morph: A longitudinal image database of normal adult age-progression," in *International Conference on Automatic Face and Gesture Recognition*, 2006, pp. 341–345.

[52] P. Phillips, H. Wechsler, J. Huang, and J. Patrick, "The feret database and evaluation procedure for face recognition," *Image and Vision Computing*, vol. 16, no. 5, pp. 295–306, 1998.

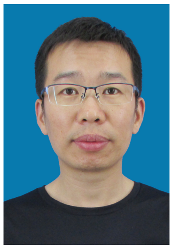
[53] P. Phillips, H. Moon, S. A. Rizvi, and P. J. Rauss, "The feret evaluation methodology for face-recognition algorithms," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 10, pp. 1090–1104, 2000.

[54] B. C. Chen, C. S. Chen, and W. H. Hsu, "Cross-age reference coding for age-invariant face recognition and retrieval," in *European Conference on Computer Vision*, 2014, pp. 768–783.

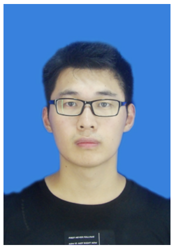
[55] G. Guo, G. Mu, F. Yun, and T. S. Huang, "Human age estimation using bio-inspired features," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 112–119.

[56] Q. Zhu, M. C. Yeh, K. T. Cheng, and S. Avidan, "Fast human detection using a cascade of histograms of oriented gradients," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2006, pp. 1491–1498.

[57] S. Chang, H. Wei, J. Tang, G. J. Qi, C. C. Aggarwal, and T. S. Huang, "Heterogeneous network embedding via deep architectures," in *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2015, pp. 119–128.



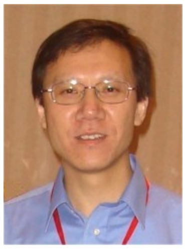
Qing Tian received his Ph.D. degree in computer science from Nanjing University of Aeronautics and Astronautics, China, in 2016. He is currently an Associate Professor in the School of Computer and Software, Nanjing University of Information Science and Technology, China. He was an Academic Visitor at the University of Manchester, UK, from 2018 to 2019. He is the recipient of the *National PhD Scholarship Award* of China in 2015, the *Best Scientific Paper Award of ICPR* in 2016, the *Excellent Doctoral Dissertation Award of Jiangsu Province* of China in 2017, etc. He has served as program committee member for several renowned international conferences, such as IJCAI, PRICAI, IDEAL, and reviewer for many prestigious international journals and conferences, such as IEEE TPAMI, IEEE TNNLS, IEEE TCYB, IEEE TIFS, ACM TIST, IJCAI, ICDM, CVPR. His research interests include machine learning, pattern recognition and computer vision.



Chuang Ma received his B.S. degree in computer science from Nanjing University of Information Science and Technology (NUIST) in 2018, China. He is currently pursuing his master degree at the NUIST. His research interests include machine learning and pattern recognition.

Meng Cao received his B.S. degree in computer science from Nanjing University of Information Science and Technology (NUIST) in 2017, China. He is currently pursuing his master degree at the NUIST. His research interests include machine learning and pattern recognition.

Songcan Chen received the B.S. degree from Hangzhou University (merged into Zhejiang University), the M.S. degree from Shanghai Jiao Tong University and the Ph.D. degree from Nanjing University of Aeronautics and Astronautics (NUAA) in 1983, 1985, and 1997, respectively. He joined in NUAA in 1986, and since 1998, he has been a full-time Professor with the Department of Computer Science and Engineering. He has authored/co-authored about 200 peer-reviewed scientific papers and ever obtained Honorable Mentions of 2006, 2007 and 2010 Best Paper Awards of Pattern Recognition Journal respectively. His current research interests include pattern recognition, machine learning, and neural computing. He is a Fellow of the IAPR and the CAAI.



Hujun Yin (SM'03) received the Ph.D. degree in neural networks from University of York, York, UK, and the B.Eng. degree in electronic engineering and the M.Sc. degree in signal processing from Southeast University, Nanjing, China. He is currently a Reader with the School of Electrical and Electronic Engineering, The University of Manchester, UK. He has published over 150 peer-reviewed articles in a wide range of topics from density modeling, image processing, face recognition, text mining and knowledge management, gene expression analysis, to novelty detection. He has served or is serving as an Associate Editor for the *IEEE Transactions on Neural Networks*, the *IEEE Transactions on Cybernetics*, the *International Journal of Neural Systems* and several other journals. He has also served as the General Co-Chair for *IDEAL* (since 2005) and Program Committee Co-Chair for the *International Symposium on Neural Networks*. His current research interests include neural networks, self-organizing learning, deep learning and pattern recognition. He is a Turing Fellow (since 2018), a Member of the EPSRC Peer Review College (since 2006) and a Senior Member of the IEEE (since 2003).