Learning High-Order Multi-View Representation by New Tensor Canonical Correlation Analysis

Jianqin Sun, Xianchao Xiu, *Member, IEEE*, Ziyan Luo, and Wanquan Liu, *Senior Member, IEEE*

***Abstract*—Canonical correlation analysis (CCA) has attracted great interest in multi-view representation. However, most of the CCA methods heavily rely on the matrix structure, which may neglect the prior geometric information in high-order data. To deal with the above issue, we first propose a novel tensor CCA formulation with orthogonality, called TCCA-O, based on the Tucker decomposition to preserve the orthogonality. Then, we incorporate a structured sparse regularization term into the TCCA-O, called TCCA-OS, to improve feature representation. In addition, we develop an efficient alternating direction method of multipliers (ADMM)-based algorithm to solve TCCA-OS and conduct numerical comparisons on four public datasets. The results validate the advantages of the proposed methods in terms of classification accuracy, parameter sensitivity, noise robustness, and model stability. In particular, TCCA-O and TCCA-OS improve the classification accuracy by at least 10.03% and 10.36%, respectively, over the state-of-the-art CCA methods on the Caltech101-7 dataset.**

***Index Terms*—Canonical correlation analysis (CCA), multi-view learning, sparse optimization, tensor representation, Tucker decomposition.**

I. Introduction

With the rapid development of data acquisition technology, it becomes much easier for us to obtain multi-view data. For example, UCI-Ad has three views, i.e., features based on the terms in the anchor URL, in the current site of the URL, and in the image, caption, and alt text URL. More details can be found in [1]. Complement and consistent geometric information from multi-view data will facilitate a more efficient learning process [2], compared to single-view learning mechanisms. Nowadays, multi-view learning has been extensively used in many fields such as biomedicine [3]–[5], engineering [6]–[8], and computing [9]–[11].

Generally speaking, multi-view learning can be classified into co-training [12], multi-kernel learning [13], [14], and subspace learning [15]–[17]. It is suggested by [18], [19] that subspace learning targets to seek a shared subspace that can contain the most feature information from different views, thus showing advantages in multi-view representation. During the last few years, canonical correlation analysis (CCA) has become one of the most popular subspace learning methods [20]. In order to improve the multi-view representation of CCA, one can embed different regularization terms or constraints, including multi-view CCA [15], sparse CCA (SCCA) [21]–[23], kernel CCA (KCCA) [24], and deep CCA (DCCA) [25]. As demonstrated in [17], [22], SCCA can integrate sparse discriminative information into CCA. Thus, SCCA can efficiently and directly deal with data where the feature dimension is larger than the number of observations. Although SCCA has obtained excellent performance in signal and image processing [26], [27], most of the existing works are limited to matrix formulations, which can only obtain pairwise correlation information and thus may ignore high-order data features [28], [29].

Recently, tensors have proved to be promising in high-order representation [30]. On one hand, tensors can preserve the high-order geometric structure [31]. On the other hand, tensors can reduce the impact of redundant data through tensor low-rank decomposition [32]. However, most of the existing tensor-based CCA variants only focus on tensor-type input data, and then unfold the data into vectors and matrices. Obviously, the high-order multi-view structure information is lost [33]–[35]. Recently, Luo *et al.* [36] utilized the covariance tensor to construct a new formulation of tensor CCA (TCCA), and solved the TCCA by CANDECOMP/PARAFAC (CP) decomposition. It is demonstrated that compared with matrix-type CCA, TCCA has better image classification results by fully taking high-order data information into consideration. In addition, TCCA can also be combined with convolutional neural network (CNN) [37] and multi-layer perception (MLP) [38]. Unfortunately, all the above TCCA methods cannot guarantee the orthogonality of canonical variables as discussed in [39]. Therefore, a natural idea comes: *is it possible to propose a TCCA variant that can simultaneously consider the orthogonality and sparsity for better high-order multi-view representation?*

Motivated by the above analysis, we first propose a new TCCA formulation by introducing the Tucker decomposition, called TCCA with orthogonality (TCCA-O), to guarantee the orthogonality of canonical variables. Unlike the CP decomposition in [36], the Tucker decomposition is more flexible and easier to add regularization terms, which has been widely used in many fields; see, e.g., [40]–[42]. Furthermore, to improve the variable selection, we enforce a sparse regularization term into TCCA-O, called TCCA with orthogonality and sparsity (TCCA-OS). Moreover, we provide the framework of the proposed TCCA-O and TCCA-OS in Fig. 1.

This work was supported in part by the National Natural Science Foundation of China under Grant 12271022 and Grant 12001019. *(Corresponding author: Ziyan Luo.)*

J. Sun and Z. Luo are with the School of Mathematics and Statistics, Beijing Jiaotong University, Beijing 100044, China (e-mail: 21121631@bjtu.edu.cn; zyluo@bjtu.edu.cn).

X. Xiu is with the School of Mechatronic Engineering and Automation, Shanghai University, Shanghai 200444, China (e-mail: xcxiu@shu.edu.cn).

W. Liu is with the School of Intelligent Systems Engineering, Sun Yat-Sen University, Guangzhou 510275, China (e-mail: liuwq63@mail.sysu.edu.cn).

This article has been accepted for publication in IEEE Transactions on Circuits and Systems for Video Technology. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2023.3263853

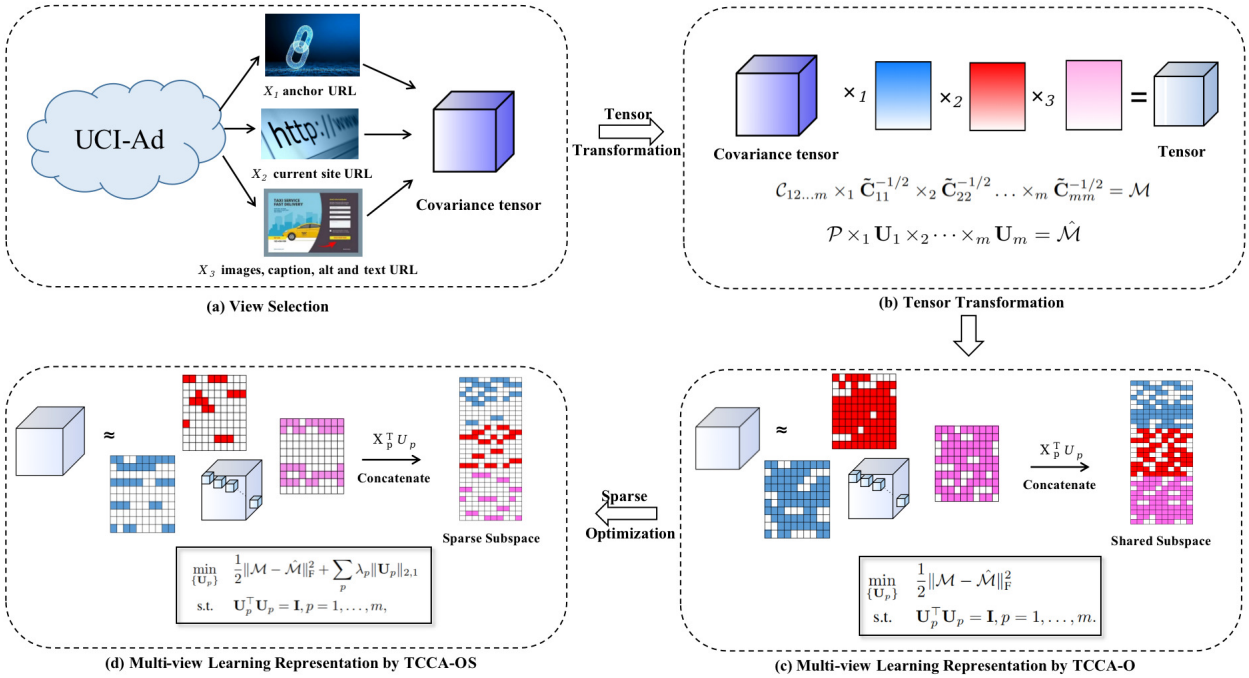


Fig. 1. The framework of TCCA-O and TCCA-OS for high-order multi-view representation. (a) The UCI-Ad dataset has three views, including the anchor URL, the current site URL, and the image, alt, and text URL, denoted by $\mathbf{X}_1, \mathbf{X}_2$, and $\mathbf{X}_3$, respectively. The covariance tensor is calculated and described as $\mathcal{C}_{12\cdots m}$. (b) Here, $\mathbf{C}_{pp}$ denotes the variance matrix, and $\mathbf{U}_p$ denotes the canonical matrix. Through some simple tensor transformations, one can get the tensor $\mathcal{M}$ and its approximation $\hat{\mathcal{M}}$. (c) TCCA-O is proposed to obtain the orthogonal canonical variables and shared subspaces. (d) TCCA-OS is constructed to alleviate the feature redundancy and achieve sparse subspaces.

Compared with the existing work, the main contributions of this paper can be summarized in the following three aspects.

1) It constructs a Tucker decomposition-based TCCA formulation (termed as TCCA-O) where the orthogonality on factor matrices is employed to ensure the irrelevance among canonical vectors, compared to the existing CP decomposition-based formulations. To the best of our knowledge, this is the first work that models CCA into such a Tucker decomposition framework.
2) It integrates a structured sparse regularization term on canonical variables of TCCA-O, and the feature redundancy in the data representation can be greatly alleviated by the resulting TCCA-OS.
3) It provides efficient optimization algorithms based on the alternating direction method of multipliers (ADMM), and validates the effectiveness of the proposed methods via numerical experiments.

The structure of this paper is listed as follows. Section II introduces notations and CCA basics. Section III formulates TCCA-O and TCCA-OS models. Section IV provides optimization algorithms. Section V validates the superiority of the proposed methods. Section VI concludes this paper.

II. Preliminaries

A. Notations

Throughout this paper, tensors are represented by calligraphic letters, say $\mathcal{X}$, matrices are represented by bold capital letters, say $\mathbf{X}$, vectors are represented by bold lowercase letters, say $\mathbf{x}$, and scalars are represented by lowercase letters, say x. For a matrix $\mathbf{X} \in \mathbb{R}^{n\times m}$, the $\ell_{2,1}$-norm is defined as $\|\mathbf{X}\|_{2,1} = \sum_{i=1}^{n}(\sum_{j=1}^{m} x_{i,j}^2)^{1/2}$, where $x_{i,j}$ denotes the ijth element. Below, some basic definitions of tensors are introduced.

Definition 2.1 (Inner product and Frobenius norm): For tensors $\mathcal{A}, \mathcal{B} \in \mathbb{R}^{I_1\times\cdots\times I_N}$, their inner product is

$$\langle \mathcal{A}, \mathcal{B} \rangle = \sum_{i_1,\ldots,i_N} a_{i_1,\ldots,i_N} b_{i_1,\ldots,i_N}. \tag{1}$$

Based on the definition of the inner product, the Frobenius norm of tensor $\mathcal{X}$ is defined as

$$\|\mathcal{A}\|_{\mathrm{F}}^2 = \langle \mathcal{A}, \mathcal{A} \rangle = \sum_{i_1=1}^{I_1}\sum_{i_2=1}^{I_2}\cdots\sum_{i_N=1}^{I_N}(a_{i_1,\ldots,i_N})^2. \tag{2}$$

Definition 2.2 (Outer product): For tensors $\mathcal{A} \in \mathbb{R}^{I_1\times\cdots\times I_N}$, $\mathcal{B} \in \mathbb{R}^{J_1\times\cdots\times J_M}$, their outer product is $\mathcal{A}\circ\mathcal{B} \in \mathbb{R}^{I_1\times\cdots\times I_N\times J_1\times\cdots\times J_M}$, whose entries are composed by

$$(\mathcal{A}\circ\mathcal{B})_{i_1,\ldots,i_N,j_1,\ldots,j_M} = a_{i_1,\ldots,i_N} b_{j_1,\ldots,j_M}. \tag{3}$$

Definition 2.3 (Mode-n unfolding matrix): For a tensor $\mathcal{A} \in \mathbb{R}^{I_1\times\cdots\times I_N}$, the mode-$n$ unfolding matrix is denoted by $\boldsymbol{A}_{(n)} \in \mathbb{R}^{I_n\times\prod_{i\neq n} I_i}$.

Definition 2.4 (n-mode product): For a tensor $\mathcal{A} \in \mathbb{R}^{I_1\times I_2\times\cdots\times I_N}$ and a matrix $\mathbf{V} \in \mathbb{R}^{r_n\times I_n}$, their n-mode

product is denoted as $\mathcal{A}\times_n \mathbf{V} \in \mathbb{R}^{I_1\times\cdots\times I_{n-1}\times r_n\times I_{n+1}\times\cdots\times I_N}$ with the element being

$$(\mathcal{A}\times_n \mathbf{V})_{i_1,\ldots,i_{n-1},r_n,i_{n+1},\ldots,i_N} = \sum_{i_n=1}^{I_n} a_{i_1,\ldots i_n,\ldots,i_N} v_{r_n,i_n}. \tag{4}$$

Moreover, the n-mode product of the tensor $\mathcal{A}$ and a vector $\mathbf{v}\in\mathbb{R}^{I_n}$ is $\mathcal{A}\,\overline{\times}_n\mathbf{v} \in \mathbb{R}^{I_1\times\cdots\times I_{n-1}\times I_{n+1}\times\cdots\times I_N}$ with the element defined as

$$(\mathcal{A}\,\overline{\times}_n\mathbf{v})_{i_1,\ldots,i_{n-1},i_{n+1},\ldots,i_N} = \sum_{i_n=1}^{I_n} a_{i_1,\ldots i_n,\ldots,i_N} v_{i_n}. \tag{5}$$

Definition 2.5 (Contracted tensor product): For a tensor $\mathcal{A}\in\mathbb{R}^{I_1\times\cdots\times I_N}$ and the matrix set $\{\mathbf{V}_p\in\mathbb{R}^{r_p\times I_p}\}_{p=1}^N$, their contracted tensor product is denoted by

$$\mathcal{B} = \mathcal{A}\times_1\mathbf{V}_1\times_2\cdots\times_N\mathbf{V}_N \in \mathbb{R}^{r_1\times\cdots\times r_N}. \tag{6}$$

Accordingly, the mode-p matricization of the tensor $\mathcal{B}$ can be given by

$$\mathcal{B}_{(p)} = \mathbf{V}_p\mathcal{A}_{(p)}(\mathbf{V}_{N-1}\otimes\cdots\otimes\mathbf{V}_{p+1}\otimes\mathbf{V}_{p-1}\otimes\cdots\otimes\mathbf{V}_1)^\top, \tag{7}$$

where $\otimes$ is the Kronecker product.

Definition 2.6 (Tucker decomposition): For tensors $\mathcal{X}\in\mathbb{R}^{I_1\times\cdots\times I_m}$, $\mathcal{G}\in\mathbb{R}^{r_1\times\cdots\times r_m}$ and matrices $\mathbf{U}_p\in\mathbb{R}^{I_p\times r_p}$, $p=1,\ldots,m$, if the following relation holds

$$\mathcal{X} = \mathcal{A}\times_1\mathbf{U}_1\times_2\mathbf{U}_2\times\cdots\times_m\mathbf{U}_m, \tag{8}$$

then (8) is called the Tucker decomposition of $\mathcal{X}$, where $\mathcal{A}$ is the core tensor, and $\mathbf{U}_1,\ldots,\mathbf{U}_m$ are the factor matrices.

Definition 2.7 (Proximal operator): For a matrix $\mathbf{X}\in\mathbb{R}^{n\times m}$ and a parameter $\beta>0$, the proximal operator $\mathrm{Prox}_{2,1}(\mathbf{X},\beta)$ is defined as

$$\mathrm{Prox}_{2,1}(\mathbf{X},\beta) = \underset{\mathbf{Y}\in\mathbb{R}^{n\times m}}{\operatorname{argmin}}\big\{\|\mathbf{Y}\|_{2,1} + \frac{1}{2\beta}\|\mathbf{Y}-\mathbf{X}\|_{\mathrm{F}}^2\big\}, \tag{9}$$

whose ith row admits the closed-form expression

$$\mathbf{y}_i = \frac{\mathbf{x}_i}{\|\mathbf{x}_i\|_2}\max\{0,\|\mathbf{x}_i\|_2-\beta\} \tag{10}$$

with $\mathbf{x}_i$ and $\mathbf{y}_i$ being its ith row of $\mathbf{X}$ and $\mathbf{Y}$, respectively. More details can be found in [43].

B. CCA Basics

Given two matrices $\mathbf{X}_1\in\mathbb{R}^{d_1\times N}$ and $\mathbf{X}_2\in\mathbb{R}^{d_2\times N}$, the objective of CCA is to find two basis vectors, $\mathbf{h}_1\in\mathbb{R}^{d_1}$ and $\mathbf{h}_2\in\mathbb{R}^{d_2}$, with the aim of maximizing the correlation between canonical variables [44]. Mathematically, it can be described as the form of

$$\max_{\mathbf{h}_1,\mathbf{h}_2}\ \frac{\mathbf{h}_1^\top\mathbf{X}_1\mathbf{X}_2^\top\mathbf{h}_2}{\sqrt{\mathbf{h}_1^\top\mathbf{X}_1\mathbf{X}_1^\top\mathbf{h}_1}\sqrt{\mathbf{h}_2^\top\mathbf{X}_2\mathbf{X}_2^\top\mathbf{h}_2}}. \tag{11}$$

For multiple canonical variables, the above formulation can be generalized to

$$\begin{aligned}\max_{\{\mathbf{h}_i\}}\quad & \sum_{i=1}^m\sum_{j=1}^m \mathbf{h}_i^\top\mathbf{X}_i\mathbf{X}_j^\top\mathbf{h}_j\\ \text{s.t.}\quad & \mathbf{h}_i^\top\mathbf{X}_i\mathbf{X}_i^\top\mathbf{h}_i = 1, i,j=1,\ldots,m,\end{aligned} \tag{12}$$

which is often called the sum of correlation CCA (SUMCOR CCA). It aims to find projection $\mathbf{h}_i$ for the ith view so that their pairwise correlations are maximized after projecting into $\mathbf{X}_i^\top\mathbf{h}_i$ [45].

However, the above CCA models can only obtain the pairwise correlation of all views, which may destroy the structure information. Suppose the multi-view data has N instances with m views, then it can be represented as $\mathbf{X}=[\mathbf{X}_1,\cdots,\mathbf{X}_m]$, where $\mathbf{X}_p\in\mathbb{R}^{d_p\times N}$, $p=1,\cdots,m$. The data of each view is assumed to be centered and arranged, for example, $\overline{\mathbf{X}}_p=[\overline{\mathbf{x}}_{p1},\ldots,\overline{\mathbf{x}}_{pN}]\in\mathbb{R}^{d_p\times N}$. Motivated by the recent developments of tensor decomposition, the following TCCA [36] is considered

$$\begin{aligned}\max_{\{\mathbf{h}_i\}}\quad & \mathcal{C}_{12\ldots m}\overline{\times}_1\mathbf{h}_1^\top\overline{\times}_2\ldots\overline{\times}_m\mathbf{h}_m^\top\\ \text{s.t.}\quad & \mathbf{h}_i^\top\mathbf{X}_i\mathbf{X}_i^\top\mathbf{h}_i = 1, i=1,\ldots,m,\end{aligned} \tag{13}$$

where $\mathcal{C}_{12\ldots m}=\frac{1}{N}\sum_{n=1}^N\overline{\mathbf{x}}_{1n}\circ\cdots\circ\overline{\mathbf{x}}_{mn}$ with $\circ$ being the outer product. In fact, $\mathcal{C}_{12\ldots m}$ is the high-order covariance tensor, which enables TCCA to maximize the whole correlation of all views.

III. New Formulations

A. TCCA-O

Although TCCA can obtain more prior geometric information from the multi-view data than SUMCOR CCA, it cannot ensure the irrelevance among canonical vectors $\mathbf{h}_{p1},\ldots,\mathbf{h}_{pr}$, where r denotes the number of extracted features and $\mathbf{h}_{pi}$ represents the ith feature extraction of the pth view. Denote the canonical matrix by $\mathbf{H}_p=[\mathbf{h}_{p1},\ldots,\mathbf{h}_{pr}]\in\mathbb{R}^{d_p\times r}$. Then the correlation tensor can be characterized as

$$\begin{aligned}\mathcal{P} &= \mathrm{corr}\left(\overline{\mathbf{X}}_1^\top\mathbf{H}_1,\ldots,\overline{\mathbf{X}}_m^\top\mathbf{H}_m\right)\\ &= \mathcal{E}\times_1\overline{\mathbf{X}}_1^\top\mathbf{H}_1\times_2\cdots\times_m\overline{\mathbf{X}}_m^\top\mathbf{H}_m\\ &= \mathcal{C}_{12\ldots m}\times_1\mathbf{H}_1^\top\times_2\cdots\times_m\mathbf{H}_m^\top\in\mathbb{R}^{r\times\cdots\times r},\end{aligned} \tag{14}$$

where $p=1,\cdots,m$, $\mathcal{E}$ is the all-one tensor with m modes, and the dimension of each mode is N. In order to exploit the shared subspace that incorporates most of the information from the multi-view data, we propose a new TCCA formulation with orthogonality (TCCA-O) given by

$$\begin{aligned}\max_{\{\mathbf{H}_p\}}\quad & \frac{1}{2}\|\mathcal{C}_{12\ldots m}\times_1\mathbf{H}_1^\top\times_2\cdots\times_m\mathbf{H}_m^\top\|_{\mathrm{F}}^2\\ \text{s.t.}\quad & \mathbf{H}_p^\top\mathbf{X}_p\mathbf{X}_p^\top\mathbf{H}_p=\mathbf{I}, p=1,\ldots,m,\end{aligned} \tag{15}$$

where $\mathbf{I}$ is the identity matrix of appropriate dimension. Obviously, different from TCCA in (13), TCCA-O can maximize the correlation between these canonical variables $\overline{\mathbf{X}}_p^\top\mathbf{H}_p$.

For the convenience of expression, let $\mathbf{U}_p=\overline{\mathbf{X}}_p^\top\mathbf{H}_p$, and then the constraint in (15) can be transformed into $\mathbf{U}_p^\top\mathbf{U}_p=\mathbf{I}$ for $p=1,\ldots,m$. Next, an equivalent but simple variant is provided in the following theorem.

Theorem 3.1: Assume that $\mathbf{C}_{ii}=\overline{\mathbf{X}}_i\overline{\mathbf{X}}_i^\top$, $\mathcal{M}=\mathcal{C}_{12\ldots m}\times_1\mathbf{C}_{11}^{-1/2}\times_2\cdots\times_m\mathbf{C}_{mm}^{-1/2}$, and $\hat{\mathcal{M}}=\mathcal{P}\times_1\mathbf{U}_1\times_2\cdots\times_m\mathbf{U}_m$. Problem (15) can be equivalently reformulated as

$$\begin{aligned}\min_{\mathcal{P},\{\mathbf{U}_p\}}\quad & \frac{1}{2}\|\mathcal{M}-\hat{\mathcal{M}}\|_{\mathrm{F}}^2\\ \text{s.t.}\quad & \mathbf{U}_p^\top\mathbf{U}_p=\mathbf{I}, p=1,\ldots,m.\end{aligned} \tag{16}$$

Proof: We first analyze the relation between tensors $\mathcal{P}$ and $\mathcal{M}$. It is observed that

$$\begin{aligned}
&(\mathcal{C}_{12\cdots m} \times_1 \mathbf{H}_1^\top \times_2 \cdots \times_m \mathbf{H}_m^\top)_{(m)} \\
&= \mathbf{H}_m^\top \mathbf{C}_{(m)} \left(\mathbf{H}_{m-1} \otimes \cdots \otimes \mathbf{H}_1\right) \\
&= \mathbf{U}_m^\top \mathbf{C}_{mm}^{-1/2} \mathbf{C}_{(m)} \left(\mathbf{C}_{m-1,m-1}^{-1/2} \mathbf{U}_{m-1}\right) \otimes \cdots \otimes \left(\mathbf{C}_{11}^{-1/2} \mathbf{U}_1\right) \\
&= \mathbf{U}_m^\top \left(\mathbf{C}_{mm}^{-1/2} \mathbf{C}_{(m)} \tilde{\mathbf{C}}_{(-m)}^{-1/2}\right) \left(\mathbf{U}_{m-1} \otimes \cdots \otimes \mathbf{U}_1\right) \\
&= \mathbf{U}_m^\top \mathcal{M} \left(\mathbf{U}_{m-1} \otimes \cdots \otimes \mathbf{U}_1\right) \\
&= (\mathcal{M} \times_1 \mathbf{U}_1^\top \times_2 \cdots \times_m \mathbf{U}_m^\top)_{(m)},
\end{aligned} \tag{17}$$

where $\mathbf{C}_{(m)} = (\mathcal{C}_{12\cdots m})_{(m)}$ and $\tilde{\mathbf{C}}_{(-m)}^{-1/2} = \mathbf{C}_{m-1,m-1}^{-1/2} \otimes \cdots \otimes \mathbf{C}_{11}^{-1/2}$. Thus, $\mathcal{P} = \mathcal{M} \times_1 \mathbf{U}_1^\top \times_2 \cdots \times_m \mathbf{U}_m^\top$. Next, the following relationships hold

$$\begin{aligned}
&\frac{1}{2}\|\mathcal{M} - \mathcal{P} \times_1 \mathbf{U}_1 \times_2 \cdots \times_m \mathbf{U}_m\|_{\mathrm{F}}^2 \\
=&\frac{1}{2}\|\mathcal{M}\|_{\mathrm{F}}^2 - \langle \mathcal{M}, \mathcal{P} \times_1 \mathbf{U}_1 \times_2 \cdots \times_m \mathbf{U}_m \rangle \\
&+ \frac{1}{2}\|\mathcal{P} \times_1 \mathbf{U}_1 \times_2 \cdots \times_m \mathbf{U}_m\|_{\mathrm{F}}^2 \\
=&\frac{1}{2}\|\mathcal{M}\|_{\mathrm{F}}^2 - \langle \mathcal{M} \times_1 \mathbf{U}_1^\top \times_2 \cdots \times_m \mathbf{U}_m^\top, \mathcal{P} \rangle + \frac{1}{2}\|\mathcal{P}\|_{\mathrm{F}}^2 \\
=&\frac{1}{2}\|\mathcal{M}\|_{\mathrm{F}}^2 - \langle \mathcal{P}, \mathcal{P} \rangle + \frac{1}{2}\|\mathcal{P}\|_{\mathrm{F}}^2 \\
=&\frac{1}{2}\|\mathcal{M}\|_{\mathrm{F}}^2 - \frac{1}{2}\|\mathcal{P}\|_{\mathrm{F}}^2.
\end{aligned} \tag{18}$$

Therefore, the desired conclusion is obtained. ∎

B. TCCA-OS

It is noted that the ith row of $\mathbf{U}_p$ represents the information extracted by all canonical vectors in the pth view for the ith feature. For multi-view learning, feature redundancy often appears. This means that not all rows in $\mathbf{U}_p$ are valuable and interpretable. Therefore, to alleviate the feature redundancy and improve the performance of data representation, we integrate the row-wise sparsity, i.e., $\ell_{2,1}$-norm, into the objective of TCCA-O, which gives the following TCCA with orthogonality and sparsity (TCCA-OS) formulation

$$\begin{aligned}
\min_{\mathcal{P}, \{\mathbf{U}_p\}} \quad & \frac{1}{2}\|\mathcal{M} - \hat{\mathcal{M}}\|_{\mathrm{F}}^2 + \sum_p \lambda_p \|\mathbf{U}_p\|_{2,1} \\
\text{s.t.} \quad & \mathbf{U}_p^\top \mathbf{U}_p = \mathbf{I}, p = 1, \ldots, m,
\end{aligned} \tag{19}$$

where λ_p $(p = 1, \ldots, m)$ are the penalty parameters for balancing the sparse regularization term with the approximation loss. By choosing different λ_p, it is possible to drop out some rows in $\mathbf{U}_p$ and thus reduce the feature redundancy.

IV. Optimization Algorithm

For the proposed TCCA-O in (16), it can be directly solved by the Tucker decomposition. First, compute the mode-p matricization of the tensor $\mathcal{M}$. After that, compute the singular value decomposition (SVD) of $\mathcal{M}_{(p)}$. Finally, the matrix $\mathbf{U}_p$ can be obtained by the leading left singular matrix, and $\mathcal{P}$ is $\mathcal{M} \times_1 \mathbf{U}_1^\top \times_2 \cdots \times_m \mathbf{U}_m^\top$. Overall, the optimization procedure for solving TCCA-O in (16) is summarized in Algorithm 1. In the following, we will discuss how to solve the proposed TCCA-OS in (19) by adopting the alternating direction method of multipliers (ADMM) [46].

Algorithm 1 Optimization Algorithm for TCCA-O

Input: Multi-view data $\mathbf{X}$, parameters d_p, r. Calculate covariance matrix $\mathbf{C}_{pp}$, covariance tensor $\mathcal{C}_{12\cdots m}$, and tensor $\mathcal{M}$.

For $p = 1, \cdots, m$

1: Compute the matrix $\mathcal{M}_{(p)} = \mathbf{C}_{pp}\mathcal{C}_{12\cdots m(p)}(\mathbf{C}_{mm-1} \otimes \cdots \otimes \mathbf{C}_{pp+1} \otimes \mathbf{C}_{pp-1} \otimes \cdots \otimes \mathbf{C}_{11})^\top \in \mathbb{R}^{d_p \times (\prod_{i \neq p} d_i)}$;

2: Compute the singular value decomposition (SVD) of $\mathcal{M}_{(p)}$ and obtain the r leading left singular matrix $\mathbf{U}_p$;

3: Compute $\mathcal{P} = \mathcal{M} \times_1 \mathbf{U}_1^\top \times_2 \cdots \times_m \mathbf{U}_m^\top$.

End for

First, by introducing variables $\mathbf{V}_p$ $(p = 1, \ldots, m)$, one can reformulate problem (19) as

$$\begin{aligned}
\min_{\mathcal{P}, \{\mathbf{V}_p\}, \{\mathbf{U}_p\}} \quad & \frac{1}{2}\|\mathcal{M} - \mathcal{P} \times_1 \mathbf{U}_1 \times_2 \cdots \times_m \mathbf{U}_m\|_{\mathrm{F}}^2 \\
& + \sum_p \lambda_p \|\mathbf{V}_p\|_{2,1} \\
\text{s.t.} \quad & \mathbf{U}_p^\top \mathbf{U}_p = \mathbf{I}, \ p = 1, \ldots, m, \\
& \mathbf{U}_p = \mathbf{V}_p, \ p = 1, \ldots, m.
\end{aligned} \tag{20}$$

The corresponding augmented Lagrangian function is

$$\begin{aligned}
&\mathcal{L}_\rho(\mathcal{P}, \{\mathbf{V}_p\}, \{\mathbf{U}_p\}, \{\mathbf{B}_p\}) \\
&= \frac{1}{2}\|\mathcal{M} - \mathcal{P} \times_1 \mathbf{U}_1 \times_2 \cdots \times_m \mathbf{U}_m\|_{\mathrm{F}}^2 \\
&\quad + \sum_p \lambda_p \|\mathbf{V}_p\|_{2,1} + \sum_{p=1}^m \langle \mathbf{B}_p, \mathbf{U}_p - \mathbf{V}_p \rangle \\
&\quad + \frac{\rho}{2} \sum_{p=1}^m \|\mathbf{U}_p - \mathbf{V}_p\|_{\mathrm{F}}^2,
\end{aligned} \tag{21}$$

where $\mathbf{B}_p$ $(p = 1, \ldots, m)$ are the Lagrangian multipliers, and $\rho \geq 0$ is the penalty parameter.

According to the framework of ADMM, we can alternately optimize the augmented Lagrangian function in a Gauss-Seidel manner as follows.

$$\begin{cases}
\mathcal{P}^{k+1} = \underset{\mathcal{P}}{\operatorname{argmin}} \ \mathcal{L}_\rho(\mathcal{P}, \{\mathbf{V}_p^k\}, \{\mathbf{U}_p^k\}, \{\mathbf{B}_p^k\}), \\
\mathbf{V}_p^{k+1} = \underset{\mathbf{V}}{\operatorname{argmin}} \ \mathcal{L}_\rho(\mathcal{P}^{k+1}, \mathbf{V}_p, \mathbf{U}_p^k, \mathbf{B}_p^k), \\
\mathbf{U}_p^{k+1} = \underset{\mathbf{U}_p^\top \mathbf{U}_p = \mathbf{I}}{\operatorname{argmin}} \ \mathcal{L}_\rho(\mathcal{P}^{k+1}, \mathbf{V}_p^{k+1}, \mathbf{U}_p, \mathbf{B}_p^k), \\
\mathbf{B}_p^{k+1} = \mathbf{B}_p^k + \rho\left(\mathbf{U}_p^{k+1} - \mathbf{V}_p^{k+1}\right),
\end{cases} \tag{22}$$

where $p = 1, \ldots, m$. Next, we will show that the resulting problems admit closed-form solutions.

A. Updating $\mathcal{P}$

When $\{\mathbf{U}_p\}, \{\mathbf{V}_p\}$, and $\{\mathbf{B}_p\}$ are determined, $\mathcal{P}$ can be updated by solving

$$\min_{\mathcal{P}} \quad \frac{1}{2}\|\mathcal{P}\|_{\mathrm{F}}^2 - \langle \mathcal{M}, \mathcal{P} \times_1 \mathbf{U}_1 \times_2 \cdots \times_m \mathbf{U}_m \rangle. \tag{23}$$

Algorithm 2 Optimization Algorithm for TCCA-OS

Input: Multi-view data $\mathbf{X}$, parameters $d_p, r, \lambda_p, t, \epsilon_1, \epsilon_2$. Calculate covariance matrix $\mathbf{C}_{pp}$, covariance tensor $\mathcal{C}_{12\cdots m}$, and tensor $\mathcal{M}$.

Initialize: Factor matrices $\{\mathbf{U}_p^0\}$ from the Tucker decomposition of $\mathcal{M}$ and $\{\mathbf{B}_p^0\}$.

While not converged **do**

1: Update $\mathcal{P}$ according to (24);
2: Update $\{\mathbf{V}_p\}$ according to (26);
3: Update $\{\mathbf{U}_p\}$ according to (29);
4: Update multipliers $\{\mathbf{B}_p\}$;
5: Check the convergence conditions.

End while

Through algebraic operations, it admits the following closed-form solution

$$\mathcal{P}^{k+1} = \mathcal{M} \times_1 \mathbf{U}_1^{k\top} \times_2 \cdots \times_m \mathbf{U}_m^{k\top}. \tag{24}$$

B. Updating $\mathbf{V}_p$

After $\mathcal{P}$ has been updated, the minimization for all $\mathbf{V}_p$'s can be handled separately by solving

$$\min_{\mathbf{V}_p} \quad \frac{\rho}{2}\|\mathbf{U}_p^k - \mathbf{V}_p + \mathbf{B}_p^k/\rho\|_{\mathrm{F}}^2 + \lambda_p\|\mathbf{V}_p\|_{2,1}, \tag{25}$$

which has a closed-form solution given by

$$\mathbf{V}_p^{k+1} = \mathrm{Prox}_{2,1}(\mathbf{U}_p^k + \mathbf{B}_p^k/\rho, \lambda_p/\rho), \ p = 1, \ldots, m. \tag{26}$$

C. Updating $\mathbf{U}_p$

Once $\mathcal{P}$ and $\{\mathbf{V}_p\}$ have been updated, $\{\mathbf{U}_p\}$ can be updated alternatively by the following problems

$$\begin{aligned}\min_{\mathbf{U}_p^\top\mathbf{U}_p=\mathbf{I}} \quad & -\langle \mathcal{M}_{(p)}\mathbf{U}_{(-p)}(\mathcal{P}_{(p)}^{k+1})^\top, \mathbf{U}_p\rangle \\ & + \frac{\rho}{2}\|\mathbf{U}_p - \mathbf{V}_p^{k+1} + \mathbf{B}_p^k/\rho\|_{\mathrm{F}}^2,\end{aligned} \tag{27}$$

where $\mathbf{U}_{(-p)} = \mathbf{U}_m^k \otimes \cdots \mathbf{U}_{p+1}^k \otimes \mathbf{U}_{p-1}^{k+1} \otimes \cdots \otimes \mathbf{U}_1^{k+1}$, $p = 1, \ldots, m$, and $\mathcal{P}_{(p)}$ is the mode-p matricization of the tensor $\mathcal{P}$. Note that (27) is equivalent to

$$\max_{\mathbf{U}_p^\top\mathbf{U}_p=\mathbf{I}} \ \langle \mathcal{M}_{(p)}\mathbf{U}_{(-p)}(\mathcal{P}_{(p)}^{k+1})^\top + \rho\mathbf{V}_p^{k+1} - \mathbf{B}_p^k, \mathbf{U}_p\rangle, \tag{28}$$

which is called the orthogonal Procrustes problem, and the optimal solution is given by

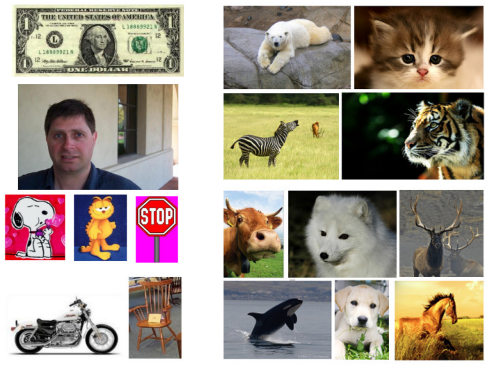
$$\mathbf{U}_p^{k+1} = \mathbf{S}_p\mathbf{D}_p^\top, \tag{29}$$

where $\mathbf{S}_p$ and $\mathbf{D}_p$ are the left and right singular matrices of $\mathcal{M}_{(p)}\mathbf{U}_{(-p)}(\mathcal{P}_{(p)}^{k+1})^\top + \rho\mathbf{V}_p^{k+1} - \mathbf{B}_p^k$.

Therefore, the whole scheme for solving TCCA-OS in (19) is summarized in Algorithm 2. Introduce the primal residual $r_1^{k+1} = \sum_{p=1}^m \|\mathbf{U}_p^{k+1} - \mathbf{V}_p^{k+1}\|_{\mathrm{F}}$, and the dual residual $r_2^{k+1} = \rho\sum_{p=1}^m \|\mathbf{V}_p^{k+1} - \mathbf{V}_p^k\|_{\mathrm{F}}$. The algorithm stops if $r_1^{k+1} \le \epsilon_1$ and $r_2^{k+1} \le \epsilon_2$ hold, where ϵ_1, ϵ_2 are two prescribed tolerance parameters.

TABLE I
THE STATISTICS OF ALL SELECTED DATASETS.

Datasets	Views	Dim	Instance	Class
Caltech101-7	Gabor Wavelet moments CENTRIST HOG	48 40 254 1984	1474	7
NUS-WIDE	color auto-correlogram wavelet texture bag of visual words	144 128 500	11647	10
UCI-Ad	image, caption, alt text current site anchor URL	588 495 472	3279	3
BBC	View 1 View 2 View 3	4569 4633 4665	685	5



(a) (b)

Fig. 2. The image examples on (a) the Caltech101-7 dataset, (b) the NUS-WIDE dataset.

D. Complexity Analysis

In this subsection, the computational complexity will be discussed. The complexity of updating $\mathcal{P}$ is $O(\prod_{p=1}^m d_p \times r^m)$, the complexity of updating $\mathbf{V}_p$ is $O(d_p \times r^2)$, and the complexity of updating $\mathbf{U}_p$ is $O(d_p \times \prod_{i\neq p}^m d_i \times r^{i-1}) + O(d_p \times r^m)$. Thus, the computational complexity of each iteration is $\max\{O(\prod_{p=1}^m d_p \times r^m), O(d_p \times \prod_{i\neq p}^m d_i \times r^{i-1}) + O(d_p \times r^m)\}$.

V. Numerical Experiments

In this section, the experiments are conducted and compared with several state-of-the-art methods including KNN[1], CCA[2], SCCA[3], and TCCA[4], to test the effectiveness of the proposed TCCA-O and TCCA-OS.

Section V-A lists the dataset description, Section V-B gives the implementation settings, Section V-C discusses the experimental results, Section V-D provides the parameter sensibility evaluation, Section V-E studies the noise robustness, Section V-F analyzes the model stability, and Section V-G presents the convergence verification.

[1]https://github.com/dingzeyuli/knn-matting
[2]https://github.com/tmarino2/scca
[3]https://github.com/htpusa/scanoncorr
[4]https://github.com/yluopku/TCCA

A. Dataset Description

In our experiments, four popular multi-view datasets are chosen, including Caltech101-7 [47], NUS-WIDE [48], UCI-Ad [1], and BBC [49]. Their statistics are given in Table I. Moreover, image examples from the Caltech101-7 and NUS-WIDE datasets are displayed in Fig. 2.

1) Caltech101-7: It is a subset of the benchmark Caltech101 dataset and consists of 1,474 images from 7 categories, i.e., Faces, Motorbikes, Dolla-Bill, Garfield, Snoopy, Stop-Sign, and Windsor-Chair. In this experiment, 48-dimensional Gabor features, 40-dimensional wavelet moments (WM), 254-dimensional CENTRIST features, and 1984-dimensional HOG features are selected as four views.

2) NUS-WIDE: It has 269,648 images from Flickr along with feature data corresponding to those images. In this experiment, 11,647 similar mammals from 10 categories are selected, including bear, cat, cow, dog, elk, fox, horse, tiger, whale, and zebra. After that 144-dimensional color auto-correlogram, 128-dimensional wavelet texture, and 500-dimensional bag of visual words are extracted as three views for comparison.

3) UCI-Ad: It contains 3,279 samples and 2 classes, including ad and nonad. The data attributes are mainly represented by binary (0-1) for features, with 0 indicating that the corresponding belongs to non-existent and 1 indicating the presence of the corresponding feature. In this experiment, the image URLs, titles, and alt texts are considered as view 1, the current site URLs are considered as view 2, and the anchor features in URLs are considered as view 3.

4) BBC: It contains 685 text documents from the BBC news and can be grouped into 5 classes, i.e., business, entertainment, politics, sport, and tech. Following a similar idea as stated in [49], different fragments from this dataset are selected as three views.

B. Implementation Settings

After obtaining the matrices $\mathbf{U}_p$, $p = 1, 2, \cdots, m$, the projected data for the pth view $\mathbf{Z}_p$ can be computed by

$$\mathbf{Z}_p = \overline{\mathbf{X}}_p^\top \tilde{C}_{pp}^{-1/2} \mathbf{U}_p. \tag{30}$$

Then all $\mathbf{Z}_p$'s are concatenated as the final representation $\mathbf{Z} = [\mathbf{Z}_1, \cdots, \mathbf{Z}_m] \in \mathbb{R}^{N \times (mr)}$ for classification. In our experiments, the classification accuracy is determined by using the k-nearest neighbor (kNN) classifier, where k is chosen from the set $\{1, \cdots, 10\}$. The test ratio is set as 0.3, and each penalty parameter is obtained through cross-validation techniques. In addition, all experiments are randomly repeated 30 times, and the mean accuracy values and the associated standard deviations are recorded.

C. Experimental Results

1) The Caltech101-7 Dataset: Table II lists the classification accuracy and standard errors (in brackets) for all compared methods. Moreover, the best results are bolded and the second-best results are underlined. It is easily concluded that most CCA-based methods perform better than KNN, which indicates that CCA plays a significant role in feature extraction and integration from different views. Unfortunately, the performance of TCCA is poor, which may be due to the lack of orthogonality constraints in the canonical variables. Moreover, the proposed TCCA-O and TCCA-OS can fully make use of the high-order structure to exploit the data information from different views, thus improving classification accuracy.

TABLE II
THE CLASSIFICATION ACCURACY (%) OF ALL COMPARED METHODS UNDER BEST DIMENSIONS.

Methods	Caltech101-7	NUS-WIDE	UCI-Ad	BBC
KNN	80.93 (± 1.42)	29.20 (± 0.32)	92.42 (± 0.35)	80.88 (± 1.14)
CCA	83.33 (± 1.55)	29.72 (± 0.50)	88.17 (± 0.9)	83.12 (± 1.13)
SCCA	83.17 (± 1.85)	30.09 (± 0.36)	92.91 (± 0.42)	83.32 (± 2.63)
TCCA	87.83 (± 2.72)	30.17 (± 1.03)	94.78 (± 1.59)	74.63 (± 2.29)
TCCA-O	<u>93.37</u> (± 0.95)	<u>33.67</u> (± 0.84)	<u>95.35</u> (± 0.64)	<u>84.98</u> (± 1.75)
TCCA-OS	**93.69** (± **1.24**)	**33.73** (± **0.81**)	**96.07** (± **0.46**)	**87.56** (± **1.94**)

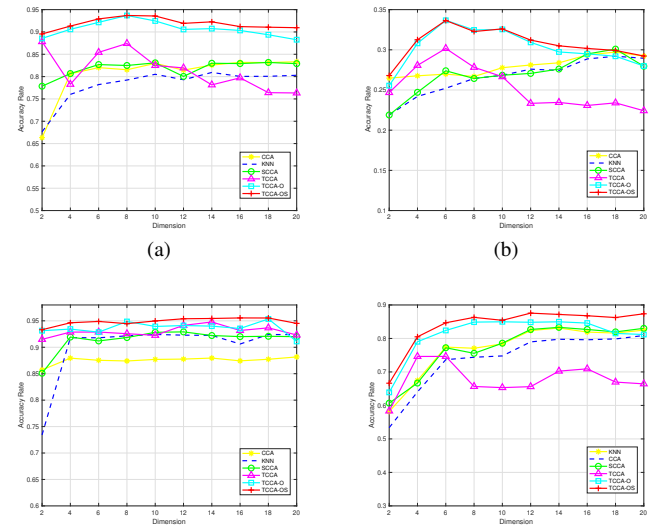


Fig. 3. The classification accuracy under different dimensions on (a) the Caltech101-7 dataset, (b) the NUS-WIDE dataset, (c) the UCI-Ad dataset, (d) the BBC dataset.

Fig. 3(a) shows the classification accuracy for all methods under different dimensions. It is found that TCCA fluctuates more obviously as the number of extracted features increases, compared to TCCA-O and TCCA-OS, which fluctuates occasionally. In most cases, TCCA-OS ranks at the top, which suggests that the proposed method is less sensitive to the choice of dimensions.

2) The NUS-WIDE Dataset: The classification accuracy under different dimensions is given in Fig. 3(b). It is seen that the classification accuracy values of TCCA-O and TCCA-OS are higher than other methods. Even compared with TCCA, they still obtain 3.5% and 3.56% increases, respectively. However, the proposed TCCA-OS does not improve significantly

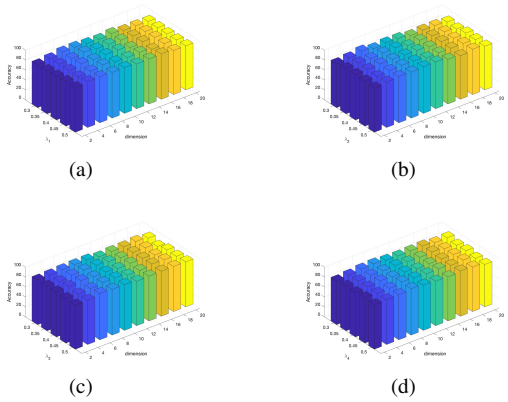


Fig. 4. The performance of TCCA-OS under different parameters on the Caltech101-7 dataset: (a) λ_1, (b) λ_2, (c) λ_3, (d) λ_4.

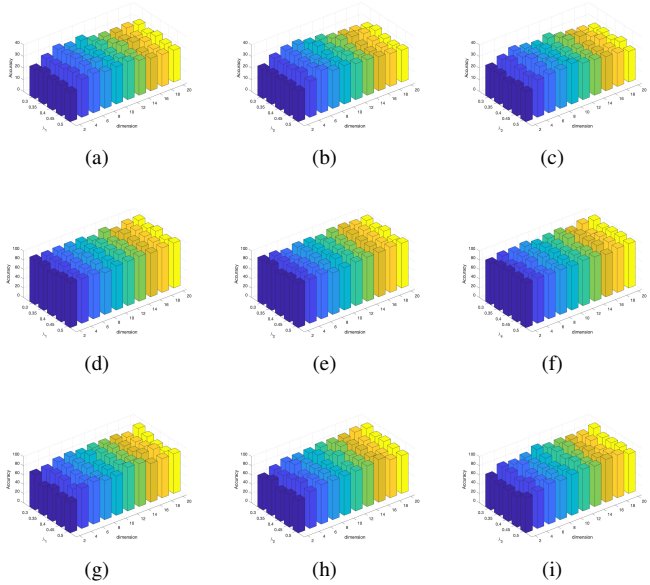


Fig. 5. The performance of TCCA-OS under different parameters, where the NUS-WIDE dataset: (a) λ_1, (b) λ_2, (c) λ_3, the UCI-Ad dataset: (d) λ_1, (e) λ_2, (f) λ_3, the BBC dataset: (g) λ_1, (h) λ_2, (i) λ_3.

compared to TCCA-O. The reason lies in that each view doesn't have sufficient data.

3) The UCI-Ad Dataset: Again from Table II, it is derived that tensor-based methods, i.e., TCCA, TCCA-O, and TCCA-OS, outperform these matrix-based methods, i.e., CCA and SCCA. In addition, TCCA-O and TCCA-OS improve the classification accuracy by 0.58% and 1.27% compared with TCCA. Fig. 3(c) shows the classification accuracy under different dimensions. The tensor-based CCA methods do not show a significant change with the increase of dimensions, however, the matrix-based CCA methods appear obvious fluctuations. In addition, TCCA-OS outperforms TCCA-O in terms of classification accuracy. For example, on the UCI-Ad dataset, TCCA-OS achieves 0.72% higher classification accuracy than TCCA-O, which shows the effectiveness of sparse regularization terms for classification.

4) The BBC Dataset: The classification results are displayed in Fig. 3(d). Clearly, the proposed TCCA-O and TCCA-OS perform much better than TCCA, which validates the importance of orthogonality and sparsity for multi-view representation.

TABLE III
THE CLASSIFICATION ACCURACY (%) UNDER BEST DIMENSIONS ON THE NUS-WIDE DATASET WITH DIFFERENT NOISE LEVELS.

Methods	30 %	60 %	90 %
KNN	25.51 (± 0.21)	23.28 (± 0.53)	21.63 (± 1.91)
CCA	24.95 (± 0.57)	21.52 (± 0.85)	21.95 (± 0.97)
SCCA	25.27 (± 0.13)	24.31 (± 0.83)	22.08 (± 0.61)
TCCA	27.57 (± 0.73)	24.06 (± 0.15)	22.36 (± 1.28)
TCCA-O	<u>30.11</u> (± 3.09)	<u>25.62</u> (± 0.70)	<u>23.41</u> (± 1.62)
TCCA-OS	**30.77** (± **0.50**)	**26.13** (± **0.70**)	**23.86** (± **1.74**)

D. Parameter Sensibility Evaluation

For the proposed TCCA-OS, the effects of the parameters λ_p $(p = 1, 2, \ldots, m)$ and the feature dimension d will be tested. Fig. 4 shows the classification accuracy under different parameters on the Caltech101-7 dataset. Fig. 5 shows the classification accuracy under different parameters on the NUS-WIDE dataset in (a)-(c), on the UCI-Ad dataset in (d)-(f), and on the BBC dataset in (g)-(i), respectively.

- All datasets are sensitive to parameter selection, but have different effects. Obviously, the Caltech101-7, UCI-Ad, and BBC datasets are less sensitive, which may be due to the fact that these datasets are sparse.
- For different feature dimension d, the classification accuracy will vary. When d is small, the performance is easily affected; when d is large, the influence will be weakened, and the performance is relatively stable.
- For different datasets, the importance of different views is different. Among these datasets, the NUS-WIDE dataset is more sensitive to λ_1.

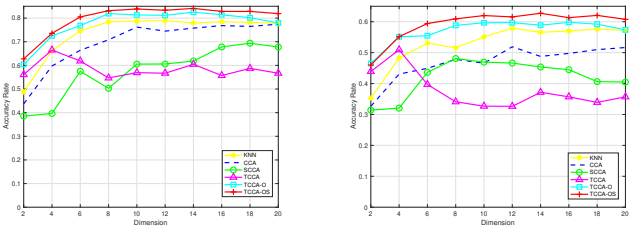
E. Noise Robustness Discussion

As mentioned above, the proposed TCCA-O and TCCA-OS are less sensitive to parameters. This subsection discusses the robustness by adding Gaussian noise (with mean zero but variance varying in 30%, 60%, 90%) to the NUS-WIDE dataset and adding normal Gaussian and non-Gaussian noise (uniform distribution noise with 0 to 1) to the BBC dataset.

Table III lists the classification accuracy and standard errors (in brackets) of all compared methods with Gaussian noise under 30%, 60%, and 90%. Moreover, the best results are bolded and the second-best results are underlined. Obviously, TCCA-OS performs the best under all noise levels, followed by TCCA-O, which shows their robustness to Gaussian noise. In particular, TCCA-OS improves the accuracy by 3.2% over TCCA and 5.82% over CCA under the 30% noise level.

Fig. 6 shows the classification performance after adding Gaussian noise and non-Gaussian noise to the BBC dataset.

This article has been accepted for publication in IEEE Transactions on Circuits and Systems for Video Technology. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2023.3263853



(a) (b)

Fig. 6. The classification accuracy under different dimensions on the BBC dataset with (a) Gaussian noise, (b) non-Gaussian noise.

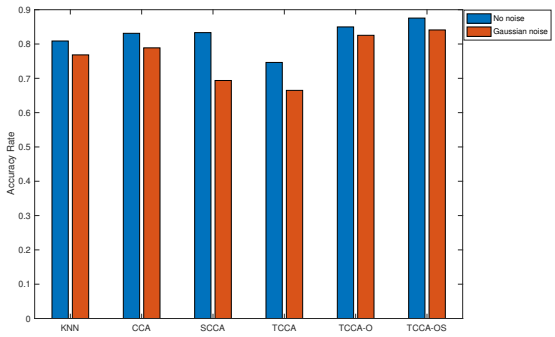


Fig. 7. The performance of all compared methods without noise and with Gaussian noise.

Again, the proposed TCCA-O and TCCA-OS outperform other compared methods, and the results are significantly better when the dimension d is small. In addition, to verify the degradation after adding noise, Fig. 7 visualizes the classification performance of all compared methods without noise and with Gaussian noise on the BBC dataset. It is concluded that the accuracy of the proposed TCCA-O and TCCA-OS does not degrade significantly, which indicates that they are less sensitive to noise.

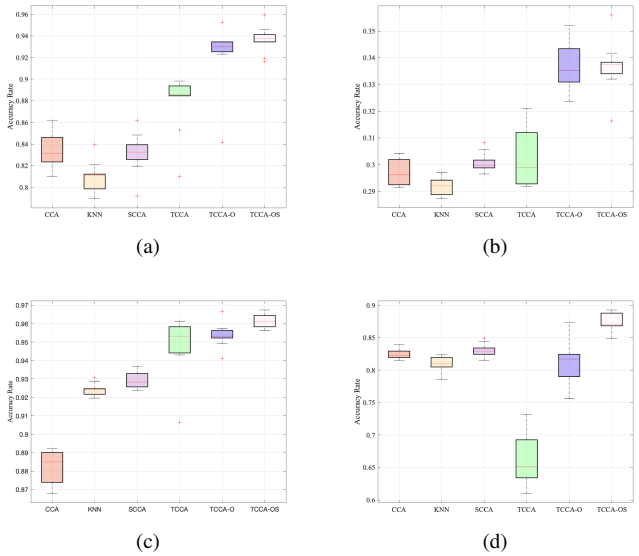
F. Model Stability Analysis

The boxplot is a classical statistical approach that describes the dispersion of a set of data, including maximum, minimum, median, upper and lower quartiles, outliers, and so on. Fig. 8 provides boxplots on the four datasets. Here, red " + " indicates outliers, and the size of boxes indicates how dispersed the data is. Some observations are

- In comparison with other methods, TCCA-OS achieves higher classification accuracy results and smaller boxes for all datasets, with more concentrated numerical classification accuracy results and better stability.
- TCCA has larger boxes and outliers on the Caltech101-7 and NUS-WIDE datasets, which reflects that the classification performance is unstable.
- Although KNN and SCCA obtain smaller box shapes than TCCA-OS on the UCI-Ad dataset, their classification accuracy is much lower than that of TCCA-OS.

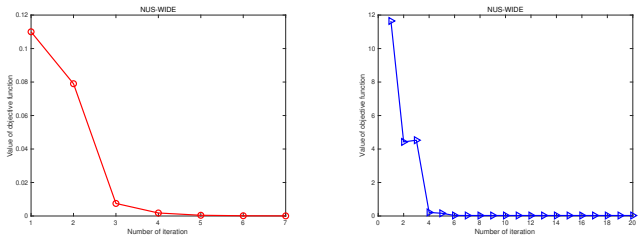
G. Convergence Verification

Since the proposed TCCA-O and TCCA-OS are both non-convex optimization problems, it is rather difficult to guarantee



(a) (b) (c) (d)

Fig. 8. The model stability analysis of all compared methods on (a) the Caltech101-7 dataset, (b) the NUS-WIDE dataset, (c) the UCI-Ad dataset, (d) the BBC dataset.



(a) (b)

Fig. 9. The values of objective functions versus the number of iterations on the NUS-WIDE dataset of (a) TCCA-O, (b) TCCA-OS.

global convergence theoretically. Below, the convergence behavior of Algorithms 1 and 2 will be verified experimentally. Fig. 9(a) and Fig. 9(b) plot the values of objective functions for TCCA-O and TCCA-OS versus the number of iterations on the NUS-WIDE dataset, respectively. It is found that the values of the objective function for TCCA and TCCA-OS decrease rapidly as the number of iterations increases. Specifically, after about 10 iterations (or even less than 10 iterations), both of them reach a steady state, which indicates their convergence.

VI. Conclusion

This paper has developed two tensor-based CCA variants, called TCCA-O and TCCA-OS, to improve the performance of multi-view representation. Unlike the existing TCCA, the proposed TCCA-O and TCCA-OS rely on the Tucker decomposition, which can preserve the orthogonality between canonical variables to select the discriminative information in multi-view data. In algorithms, an iterative optimization scheme has been designed using the ADMM and each resulting subproblem admits closed-form solutions. Numerical experiments on four benchmark datasets have validated their superiority in tasks of image and text classification.

In the future, the following issues need to be investigated. First, high-order tensor decomposition calculations are time-consuming, thus it is important to develop faster algorithms. Secondly, for practical scenarios, how to determine sparse regularization functions is worth discussing. Finally, combining the proposed methods with deep neural networks is necessary for learning nonlinear correlations of multi-view data.

REFERENCES

[1] N. Kushmerick, "Learning to remove internet advertisements," in *Proceedings of the 3rd Annual Conference on Autonomous Agents*, 1999, pp. 175–181.

[2] Y. Tian, D. Xu, and C. Zhang, "A review of multi-instance learning research," *Operations Research Transactions*, vol. 22, no. 2, pp. 1–17, 2018.

[3] Y. Li, F.-X. Wu, and A. Ngom, "A review on machine learning principles for multi-view biological data integration," *Briefings in Bioinformatics*, vol. 19, no. 2, pp. 325–340, 2018.

[4] B. Yang, M. Lange, A. Millett-Sikking, X. Zhao, J. Bragantini, S. VijayKumar, M. Kamb, R. Gómez-Sjöberg, A. C. Solak, W. Wang *et al.*, "Daxi-high-resolution, large imaging volume and multi-view single-objective light-sheet microscopy," *Nature Methods*, vol. 19, no. 4, pp. 461–469, 2022.

[5] S. Priya, M. B. Burns, T. Ward, R. A. Mars, B. Adamowicz, E. F. Lock, P. C. Kashyap, D. Knights, and R. Blekhman, "Identification of shared and disease-specific host gene-microbiome associations across human diseases using multi-omic integration," *Nature Microbiology*, pp. 1–16, 2022.

[6] C. Xu, D. Tao, and C. Xu, "Multi-view intact space learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 12, pp. 2531–2544, 2015.

[7] X. Xiu, Z. Miao, Y. Yang, and W. Liu, "Deep canonical correlation analysis using sparsity-constrained optimization for nonlinear process monitoring," *IEEE Transactions on Industrial Informatics*, vol. 18, no. 10, pp. 6690–6699, 2022.

[8] Y. Deng, H. Chen, and Y. Li, "MVF-Net: A multi-view fusion network for event-based object classification," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 12, pp. 8275–8284, 2022.

[9] S. Sun, "A survey of multi-view machine learning," *Neural Computing and Applications*, vol. 23, no. 7, pp. 2031–2038, 2013.

[10] M. Kan, S. Shan, H. Zhang, S. Lao, and X. Chen, "Multi-view discriminant analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 1, pp. 188–194, 2016.

[11] Y. Jia, H. Liu, J. Hou, S. Kwong, and Q. Zhang, "Self-supervised symmetric nonnegative matrix factorization," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 7, pp. 4526–4537, 2022.

[12] A. Kumar and H. Daumé, "A co-training approach for multi-view spectral clustering," in *Proceedings of the 28th International Conference on Machine Learning*. Citeseer, 2011, pp. 393–400.

[13] D. Guo, J. Zhang, X. Liu, Y. Cui, and C. Zhao, "Multiple kernel learning based multi-view spectral clustering," in *2014 22nd International Conference on Pattern Recognition*. IEEE, 2014, pp. 3774–3779.

[14] H. Wang, Y. Wang, Z. Zhang, X. Fu, L. Zhuo, M. Xu, and M. Wang, "Kernelized multiview subspace analysis by self-weighted learning," *IEEE Transactions on Multimedia*, vol. 23, pp. 3828–3840, 2021.

[15] D. M. Witten, R. Tibshirani, and T. Hastie, "A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis," *Biostatistics*, vol. 10, no. 3, pp. 515–534, 2009.

[16] N. Chen, J. Zhu, F. Sun, and E. P. Xing, "Large-margin predictive latent subspace learning for multiview data analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 12, pp. 2365–2378, 2012.

[17] M. Xu, Z. Zhu, X. Zhang, Y. Zhao, and X. Li, "Canonical correlation analysis with $L_{2,1}$-norm for multiview data representation," *IEEE Transactions on Cybernetics*, vol. 50, no. 11, pp. 4772–4782, 2019.

[18] H. Wang, G. Jiang, J. Peng, R. Deng, and X. Fu, "Towards adaptive consensus graph: Multi-view clustering via graph collaboration," *IEEE Transactions on Multimedia*, DOI: 10.1109/TMM.20 22.3212270.

[19] H. Wang, M. Yao, G. Jiang, Z. Mi, and X. Fu, "Graph-collaborated auto-encoder hashing for multi-view binary clustering," *IEEE Transactions on Neural Networks and Learning Systems*, DOI: 10.1109/TNNLS.2023.3239033.

[20] X. Yang, W. Liu, W. Liu, and D. Tao, "A survey on canonical correlation analysis," *IEEE Transactions on Knowledge and Data Engineering*, vol. 33, no. 6, pp. 2349–2368, 2021.

[21] D. R. Hardoon and J. Shawe-Taylor, "Sparse canonical correlation analysis," *Machine Learning*, vol. 83, no. 3, pp. 331–353, 2011.

[22] D. Chu, L.-Z. Liao, M. K. Ng, and X. Zhang, "Sparse canonical correlation analysis: New formulation and algorithm," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 12, pp. 3050–3065, 2013.

[23] X. Xiu, Y. Yang, L. Kong, and W. Liu, "Data-driven process monitoring using structured joint sparse canonical correlation analysis," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 68, no. 1, pp. 361–365, 2021.

[24] Q. Chen and Y. Wang, "Key-performance-indicator-related state monitoring based on kernel canonical correlation analysis," *Control Engineering Practice*, vol. 107, p. 104692, 2021.

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

[26] Y. Song, P. J. Schreier, D. Ramírez, and T. Hasija, "Canonical correlation analysis of high-dimensional data with very small sample support," *Signal Processing*, vol. 128, pp. 449–458, 2016.

[27] M. Kim, J. H. Won, J. Youn, and H. Park, "Joint-connectivity-based sparse canonical correlation analysis of imaging genetics for detecting biomarkers of parkinson's disease," *IEEE Transactions on Medical Imaging*, vol. 39, no. 1, pp. 23–34, 2019.

[28] L. Wang, L.-H. Zhang, C. Shen, and R.-C. Li, "Orthogonal multi-view analysis by successive approximations via eigenvectors," *Neurocomputing*, vol. 512, pp. 100–116, 2022.

[29] G. Jiang, J. Peng, H. Wang, Z. Mi, and X. Fu, "Tensorial multi-view clustering via low-rank constrained high-order graph learning," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 8, pp. 5307–5318, 2022.

[30] Y. Chen, X. Xiao, C. Peng, G. Lu, and Y. Zhou, "Low-rank tensor graph learning for multi-view subspace clustering," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 1, pp. 92–104, 2022.

[31] T. G. Kolda and B. W. Bader, "Tensor decompositions and applications," *SIAM Review*, vol. 51, no. 3, pp. 455–500, 2009.

[32] L. Qi and Z. Luo, *Tensor Analysis: Spectral Theory and Special Tensors*. Society for Industrial and Applied Mathematics, 2017.

[33] T.-K. Kim, S.-F. Wong, and R. Cipolla, "Tensor canonical correlation analysis for action classification," in *2007 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2007, pp. 1–8.

[34] D. Tao, X. Li, X. Wu, and S. J. Maybank, "General tensor discriminant analysis and gabor features for gait recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 10, pp. 1700–1715, 2007.

[35] H. Lu, "Learning canonical correlations of paired tensor sets via tensor-to-vector projection," in *23rd International Joint Conference on Artificial Intelligence*, 2013.

[36] Y. Luo, D. Tao, K. Ramamohanarao, C. Xu, and Y. Wen, "Tensor canonical correlation analysis for multi-view dimension reduction," *IEEE Transactions on Knowledge and Data Engineering*, vol. 27, no. 11, pp. 3111–3124, 2015.

[37] X. Yang, W. Liu, and W. Liu, "Tensor canonical correlation analysis networks for multi-view remote sensing scene recognition," *IEEE Transactions on Knowledge and Data Engineering*, vol. 34, no. 6, pp. 2948–2961, 2022.

[38] H. S. Wong, L. Wang, R. Chan, and T. Zeng, "Deep tensor CCA for multi-view learning," *IEEE Transactions on Big Data*, vol. 8, no. 6, pp. 1664–1677, 2022.

[39] J. Nie, L. Wang, and Z. Zheng, "Higher order correlation analysis for multi-view learning," *arXiv:2201.11949*, 2022.

[40] J. Cai, Z. Cao, and L. Zhang, "Learning a single Tucker decomposition network for lossy image compression with multiple bits-per-pixel rates," *IEEE Transactions on Image Processing*, vol. 29, pp. 3612–3625, 2020.

[41] M. Zhao, W. Li, L. Li, P. Ma, Z. Cai, and R. Tao, "Three-order tensor creation and Tucker decomposition for infrared small-target detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–16, 2021.

[42] Y. Qiu, G. Zhou, Y. Wang, Y. Zhang, and S. Xie, "A generalized graph regularized non-negative Tucker decomposition framework for tensor data representation," *IEEE Transactions on Cybernetics*, vol. 52, no. 1, pp. 594–607, 2022.

[43] F. Nie, H. Huang, X. Cai, and C. Ding, "Efficient and robust feature selection via joint $\ell_{2,1}$-norms minimization," *Advances in Neural Information Processing Systems*, vol. 23, 2010.
[44] H. Harold, "Relations between two sets of variates," *Biometrika*, vol. 28, no. 3/4, p. 321, 1936.
[45] 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.
[46] S. Boyd, N. Parikh, E. Chu, B. Peleato, J. Eckstein *et al.*, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations and Trends® in Machine learning*, vol. 3, no. 1, pp. 1–122, 2011.
[47] X. Cai, F. Nie, H. Huang, and F. Kamangar, "Heterogeneous image feature integration via multi-modal spectral clustering," in *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2011, pp. 1977–1984.
[48] T.-S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y. Zheng, "NUS-WIDE: A real-world web image database from national university of singapore," in *Proceedings of the ACM International Conference on Image and Video Retrieval*, 2009, pp. 48:1–48:9.
[49] H. Cai, B. Liu, Y. Xiao, and L. Lin, "Semi-supervised multi-view clustering based on orthonormality-constrained nonnegative matrix factorization," *Information Sciences*, vol. 536, pp. 171–184, 2020.

Jianqin Sun is currently pursuing the M.S. degree with the School of Mathematics and Statistics, Beijing Jiaotong University, Beijing, China.
Her current research interests include tensor representation and sparse optimization.

Xianchao Xiu received the Ph.D. degree in Operations Research from Beijing Jiaotong University, China, in 2019. From June 2019 to May 2021, he worked as a Postdoctoral Researcher at Peking University, China. He is a faculty member at the School of Mechatronic Engineering and Automation, Shanghai University, China.
His current research interests include large-scale sparse optimization, signal processing, deep learning, and data-driven process monitoring.

Ziyan Luo received the Ph.D. degree in Operations Research from Beijing Jiaotong University, China, in 2010. She was a visiting scholar in Stanford University, National University of Singapore, University of Southampton and Hong Kong Polytechnic University. She is currently a Full Professor at School of Mathematics and Statistics in Beijing Jiaotong University, China.
Her current research interests include large-scale sparse and low-rank optimization, tensor analysis, tensor optimization and its applications in statistical learning and high-dimensional data analysis.

Wanquan Liu received the B.S. degree in Applied Mathematics from Qufu Normal University, China, in 1985, the M.S. degree in Control Theory and Operation Research from Chinese Academy of Science in 1988, and the Ph.D. degree in Electrical Engineering from Shanghai Jiaotong University, in 1993. He once held the ARC Fellowship, U2000 Fellowship and JSPS Fellowship and attracted research funds from different resources over 2.4 million dollars. He is currently a Full Professor at the School of Intelligent Systems Engineering, Sun Yat-Sen University, Guangzhou, China.
His current research interests include large-scale pattern recognition, signal processing, machine learning, and control systems.