Tensor Canonical Correlation Analysis for Multi-view Dimension Reduction

Yong Luo*†‡ Dacheng Tao‡ Yonggang Wen†
Kotagiri Ramamohanarao§ Chao Xu*
yluo180@gmail.com, dacheng.tao@uts.edu.au, ygwen@ntu.edu.sg
rao@csse.unimelb.edu.au, xuchao@cis.pku.edu.cn

08 February 2015

Abstract

Canonical correlation analysis (CCA) has proven an effective tool for two-view dimension reduction due to its profound theoretical foundation and success in practical applications. In respect of multi-view learning, however, it is limited by its capability of only handling data represented by two-view features, while in many real-world applications, the number of views is frequently many more. Although the ad hoc way of simultaneously exploring all possible pairs of features can numerically deal with multi-view data, it ignores the high order statistics (correlation information) which can only be discovered by simultaneously exploring all features.

Therefore, in this work, we develop tensor CCA (TCCA) which straightforwardly yet naturally generalizes CCA to handle the data of an arbitrary number of views by analyzing the covariance tensor of the different views. TCCA aims to directly maximize the canonical correlation of multiple (more than two) views. Crucially, we prove that the multi-view canonical correlation maximization problem is equivalent to finding the best rank-1 approximation of the data covariance tensor, which can be solved efficiently using the well-known alternating least squares (ALS) algorithm. As a consequence, the high order correlation information contained in the different views is explored and thus a more reliable common subspace shared by all features can be obtained. In addition, a non-linear extension of TCCA is presented. Experiments on various challenge tasks, including large scale biometric structure prediction, internet advertisement classification and web image annotation, demonstrate the effectiveness of the proposed method.

1 Introduction

The features utilized in many real-world data mining tasks are frequently of high dimension and extracted from multiple views (or sources). For example, both the page content and hyperlink represented by bag-of-words (BOW) are usually used in web page classification Blum and Mitchell (1998); Foster et al (2008), and it is common to combine the global (such as GIST Oliva and Torralba (2001)) and local (such as SIFT Lowe (2004)) descriptors in image annotation Chua et al (2009); Guillaumin et al (2009). In these applications, the features can have dimensions of up to several hundred or thousand.

*Key Laboratory of Machine Perception (Ministry of Education), School of Electronics Engineering and Computer Science, Peking University, Beijing, China.
†Division of Networks and Distributed Systems, School of Computer Engineering, Nanyang Technological University, Singapore.
‡Centre for Quantum Computation & Intelligent Systems and the Faculty of Engineering & Information Technology, University of Technology, Sydney, Sydney, Australia.
§Department of Computer Science and Software Engineering, The University of Melbourne, Australia.

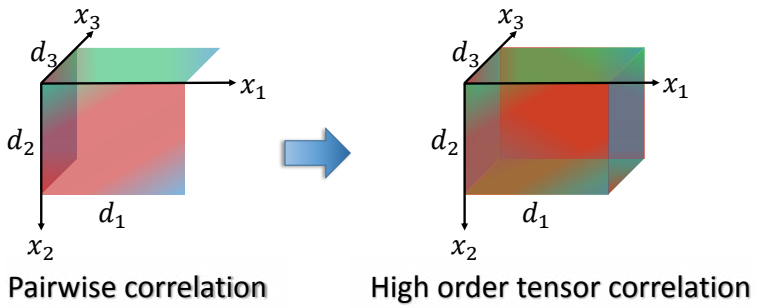

Figure 1: The tensor CCA motivation. Only the pairwise correlation is explored in the traditional extensions of CCA, while much more information (i.e., the high order correlation) that can only be obtained by simultaneously examining all views is explored in the proposed TCCA.

Multi-view dimension reduction Foster et al (2008) seeks a low-dimensional common subspace to compactly represent the heterogeneous data, in which each of the data examples is associated with multiple high-dimensional features. It often benefits the subsequent learning process significantly in that the curse-of-dimensionality is alleviated and the computation-al efficiency is improved Hou et al (2010); Han et al (2012). Canonical correlation analysis (CCA), which is designed to inspect the linear relationship between two sets of variables Hardoon et al (2004); Bach and Jordan (2005), was formally introduced as a multi-view dimension reduction method in Foster et al (2008), where the authors prove that the labeled instance complexity can be effectively reduced under certain weak assumptions. In addition, CCA has been widely used for multi-view classification Farquhar et al (2005), regression Kakade and Foster (2007), clustering Blaschko and Lampert (2008); Chaudhuri et al (2009), etc. Theoretically, Bach and Jordan Bach and Jordan (2005) interpreted CCA probabilistically as a latent variable model, and thus it is able to be involved in a larger probabilistic model.

In spite of the profound theoretical foundation and practical success of CCA in multi-view learning, it can only handle data that is represented by two-view features. The features utilized in many real-world applications, however, are usually extracted from more than two views. For example, different kinds of color, texture and shape features are popular used in visual analysis-based tasks such as image annotation and video retrieval. A typical approach for generalizing CCA to several views is to maximize the sum of pairwise correlations between different views Vía et al (2007). The main drawback of this strategy is that only the statistics (correlation information) between pairs of features is explored, while high-order statistics that can only be obtained by simultaneously examining all features is ignored.

To tackle this problem, we develop tensor CCA (TCCA) to generalize CCA to handle an arbitrary number of views in a straightforward and yet natural way. In particular, TCCA aims to directly maximize the correlation between the canonical variables of all views, and this is achieved by analyzing the high-order covariance tensor over the data from all views. We prove that maximizing the correlation is equivalent to approximating the covariance tensor with a rank-1 tensor in an optimal least square sense. This approximation has been investigated in the literature and an efficient alternating least square (ALS) algorithm can be adopted for optimization Kroonenberg and De Leeuw (1980); De Lathauwer et al (2000b); Comon et al (2009). With respect to the traditional pairwise correlation maximization, the statistics (correlation information) explored can be measured using the $m(m-1)/2$ covariance matrices of size (d^2), where m is the number of views and d represents the average feature dimensions, whereas in the proposed TCCA, the size of the covariance tensor is (d^m). Fig. 1 is an illustrative example, where $m = 3$. Much more correlation information is encoded in the common subspace shared by all features in multi-view dimension reduction, and thus hopefully better performance can be achieved. Furthermore, we extend the proposed TCCA to the non-linear case, which is useful when the feature dimensions are very high and limited instances are available. We perform extensive experiments on a variety of challenge tasks, including large scale biometric structure prediction, internet advertisement classification and web image annotation. We compare the proposed method with the traditional CCA Foster et al (2008) and its multi-view extension Vía et al (2007), as well as two representative unsupervised multi-view dimension reduction approaches Long et al (2008); Han et al (2012). The results confirm the effectiveness of the proposed TCCA.

The article is organized as follows. We summarize closely related works in Section 2. A brief introduction of CCA and its traditional multi-view extension is presented in Section 3. Section 4 includes the description, formulation, and analysis of the proposed TCCA, as well as its non-linear extension kernel TCCA (KTCCA) for multi-view dimension reduction. Extensive experiments are presented in Section 5 and the paper is concluded in Section 6.

2 Related Work

2.1 Multi-view Dimension Reduction

Dimension reduction is a key technique in machine learning. The goal of dimension reduction is to find a low dimensional representation for high dimensional data Xia et al (2010). Feature selection and feature transformation and the two main approaches for dimension reduction. The former aims to select a subset of variables from the original, while the latter transforms the data to a new space of fewer dimensions. The dimension reduction can be performed in an either unsupervised (e.g., principal component analysis (PCA) and Laplacian eigenmaps (LE) Belkin and Niyogi (2001)), semi-supervised Benabdeslem and Hindawi (2014), or supervised (e.g., linear discriminant analysis (LDA)) setting, differed in the amount of labeled information being utilized.

In another research line, multi-view learning has attracted much attention recently. The multi-view we refer to here is the multiple feature representations of an object, not the spatial viewpoints in some other vision and graphics applications Su et al (2009). We generally classify the multi-view learning algorithms into three families: weighted view combination Lanckriet et al (2004); McFee and Lanckriet (2011), multi-view dimension reduction Hardoon et al (2004); White et al (2012), and view agreement exploration Blum and Mitchell (1998); Kumar et al (2011). Multi-view dimension reduction focuses on removing irrelevant or redundant information Benabdeslem and Hindawi (2014) and reducing the feature dimension of data that consists of multiple views by leveraging the dependencies, coherence, and complementarity of those views. The different views are often assumed to be conditionally independent, thus a latent representation shared by all views can be obtained by exploiting the conditional independence structure of the multi-view data Foster et al (2008); Long et al (2008); White et al (2012); Han et al (2012); Chen et al (2012). For example, canonical correlation analysis (CCA) is employed for multi-view dimension reduction in Foster et al (2008) to exploit the underlying conditional independence and redundancy assumption in multi-view learning. A general unsupervised learning method is presented in Long et al (2008) for multi-view data, where a consensus representation is learned by first applying dimension reduction technique (such as spectral embedding Belkin and Niyogi (2001)) on each view and then combining the results via matrix factorization. In Han et al (2012), the structured sparsity Jenatton et al (2011) is enforced among the different views in the learning of low-dimension consensus representation, to allow information being shared across subsets of features adaptively. In contrast to unsupervised multi-view dimension reduction, the similarity/dissimilarity pairwise constraints are utilized in Hou et al (2010) for semi-supervised multi-view dimension reduction. In Chen et al (2012), the supervising information is also incorporated in the learned latent shared subspace by the use of a large-margin latent Markov network. In these methods, local optimal subspace can usually be obtained. Therefore, White et al. White et al (2012) proposed a convex formulation for learning a shared subspace of multiple sources. In the learned subspace, conditional independence constraints are enforced.

2.2 Canonical Correlation Analysis and Its Extensions

Canonical correlation analysis (CCA), originally proposed by Hotelling (1936), finds bases for two random variables (or sets of variables) so that the coordinates of the variable pairs projected on these bases are maximally correlated Hardoon et al (2004). Much success has been achieved by applying CCA to pattern recognition and data mining. For example, SVM-2K was proposed in Farquhar et al (2005) for two-view classification. It combines kernel CCA and support vector machine (SVM) in a single optimization problem, and the authors prove that the Rademacher complexity of SVM-2K is significantly lower than the individual SVMs. Kakade and Foster Kakade and Foster (2007) presented a multi-view regression algorithm regularized with a norm that is derived by applying CCA on unlabeled data. The authors show that the intrinsic dimension of the regression problem with the induced norm can be characterized by the correlation coefficients

obtained in CCA. Under the conditionally uncorrelated assumption, a simple and efficient subspace learning algorithm based on CCA was proposed in Chaudhuri et al (2009) for multi-view clustering. The algorithm was shown to work well under much weaker separation conditions than the previous clustering methods.

In addition to these applications, there have been dozens of developments for CCA, most of which concentrate on inspecting the relationship between two sets of tensors rather than vectors. For example, the classical CCA was extended in Lee and Choi (2007) to 2D-CCA, which directly analyzes 2D images without reshaping them into vectors. Some of its extensions are local 2D-CCA Wang (2010), sparse 2D-CCA Yan et al (2012), and multilinear CCA (MCCA) Lu (2013). Considering that the two high-order tensors to be studied may share multiple modes (e.g., the video volume data), Kim and Cipolla Kim and Cipolla (2009) presented two architectures for tensor correlation maximization by applying canonical transformation on the non-shared modes. In this way, features that have a good balance between flexibility and descriptive power may be obtained. This method is also termed "tensor CCA" (TCCA), but is quite different from the approach proposed in this paper. The main difference lies in that the latter focuses on analyzing two high-order tensor data sets, while our objective is to analyze the high-order statistics among multiple vector data sets (views).

The most closely related works to our methods, as far as we are concerned, are the maximum variance CCA (CCA-MAXVAR) Kettenring (1971) and an adaptive CCA algorithm termed CCA-LS Vía et al (2007), which is based on least square (LS) regression. The CCA-MAXVAR algorithm is performed by weighted combination of the canonical variables (projected vectors) of all views to approximate a latent common representation. This approach requires costly singular value decomposition (SVD) for optimization and cannot be trained in an adaptive fashion. To avoid these drawbacks, Via et al. Vía et al (2007) reformulated CCA-MAXVAR as a set of coupled LS regression problems, which seeks to minimize the distance between each pair of canonical variables. The reformulation is proved to be equivalent to the original CCA-MAXVAR formulation, but is much more efficient and can be learned adaptively. Nevertheless, there is still a disadvantage to both CCA-LS and CCA-MAXVAR, namely that only the pairwise correlations are exploited, while the high order correlations between all views are ignored. We developed the following tensor CCA framework to rectify this shortcoming.

3 Canonical Correlation Analysis (CCA) and Its Multi-view Generalization

This section briefly introduces standard canonical correlation analysis (CCA) and its traditional generalizations on several data sets Kettenring (1971); Vía et al (2007). Given two sets of column vectors $\mathbf{x}_{1n} \in \mathbb{R}^{d_1}$, $\mathbf{x}_{2n} \in \mathbb{R}^{d_2}$, $n = 1, \ldots, N$. The objective of CCA is to find a pair of projections (usually called canonical vectors) $\mathbf{h}_1$, $\mathbf{h}_2$, such that correlations between the two vectors of canonical variables $\mathbf{z}_1 \in \mathbb{R}^N$ and $\mathbf{z}_2 \in \mathbb{R}^N$ with each $z_{1n} = \mathbf{x}_{1n}^T \mathbf{h}_1$, $z_{2n} = \mathbf{x}_{2n}^T \mathbf{h}_2$, are maximized. The optimization problem is thus given by

$$\underset{\mathbf{z}_1, \mathbf{z}_2}{\operatorname{argmax}} \rho = \operatorname{corr}(\mathbf{z}_1, \mathbf{z}_2) = \frac{\mathbf{h}_1^T C_{12} \mathbf{h}_2}{\sqrt{\mathbf{h}_1^T C_{11} \mathbf{h}_1 \mathbf{h}_2^T C_{22} \mathbf{h}_2}}, \tag{3.1}$$

where $C_{11} = X_1 X_1^T$, $C_{22} = X_2 X_2^T$ are data variance matrices, and $C_{12} = X_1 X_2^T$ is the covariance matrix. Here, $X_1 \in \mathbb{R}^{d_1 \times N}$ and $X_2 \in \mathbb{R}^{d_2 \times N}$ are the stacked data matrices. The optimization of problem (3.1) leads to the main solution of CCA, and the remaining solutions are given by maximizing the same correlation under the constraint of being orthogonal to the previous solutions.

CCA-MAXVAR Kettenring (1971) generalizes CCA to m views. Suppose the data matrix for the p'th view is $X_p \in \mathbb{R}^{d_p \times N}$, then the optimization problem of CCA-MAXVAR for finding the canonical vectors $\{\mathbf{h}_p\}_{p=1}^m$ is

$$\begin{aligned} &\underset{\mathbf{z}, \mathbf{a}, \{\mathbf{h}_p\}_{p=1}^m}{\operatorname{argmin}} \frac{1}{m} \sum_{p=1}^{m} \|\mathbf{z} - \alpha_p \mathbf{z}_p\|_2^2, \\ &\text{s.t. } \|\mathbf{z}_p\|_2 = 1, \end{aligned} \tag{3.2}$$

where $\mathbf{z}_p = X_p^T \mathbf{h}_p$ is the vector of canonical variables, $\mathbf{z}$ is the best possible one-dimensional PCA representation, and $\mathbf{a} = [\alpha_1, \ldots, \alpha_m]^T$ is the vector of combination weights. To avoid a trivial solution, an additional

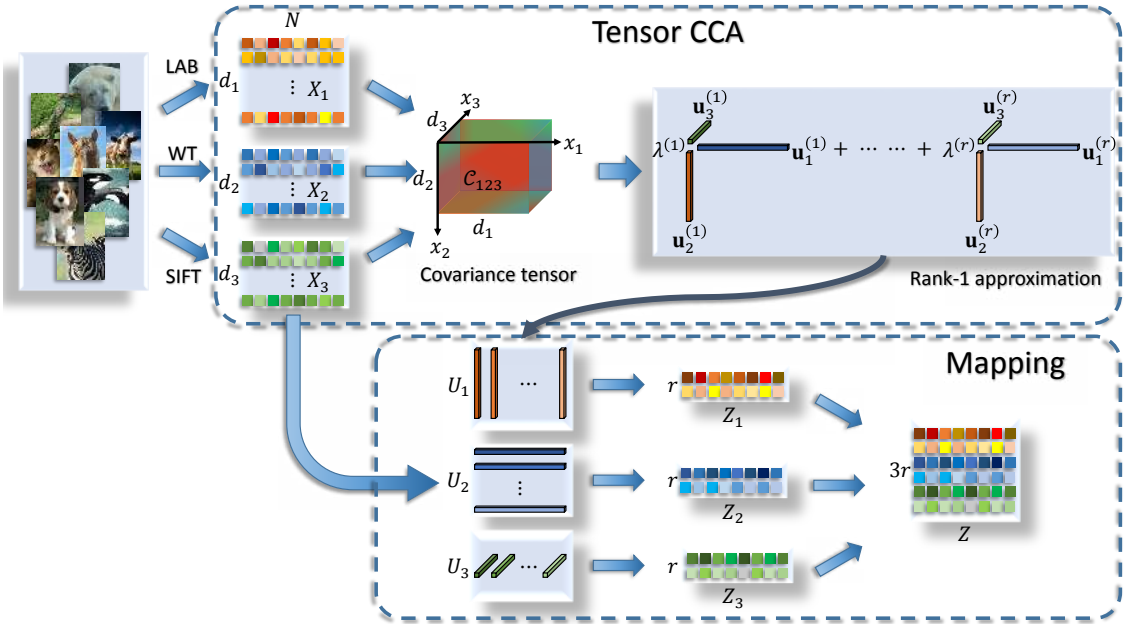

Figure 2: System diagram of the multi-view dimension reduction method by the use of the proposed TCCA. Firstly, different kinds of features are extracted to represent the available instances in different views. Then a covariance tensor is calculated on the obtained representations $X_p, p = 1, \ldots, m$ to discover the correlation information between all views. By approximating the covariance tensor with a set of rank-1 tensors, we obtain the transformation matrix U_p for the p'th view. Each U_p maps the original X_p to the low dimensional Z_p in the common subspace, and the final representation is a concatenation of $Z_p, p = 1, \ldots, m$.

constraint such as $\|\alpha_p\|_2^2 = m$ is enforced. The solutions of (3.2) can be obtained using the SVD of X_p. To develop an efficient and adaptive algorithm, Via et al. Vía et al (2007) reformulated (3.2) as

$$\begin{aligned} &\underset{\mathbf{z},\mathbf{a},\{\mathbf{h}_p\}_{p=1}^m}{\operatorname{argmin}} \frac{1}{2m(m-1)} \sum_{p,q=1}^{m} \|X_p^T \mathbf{h}_p - X_q^T \mathbf{h}_q\|_2^2, \\ &\text{s.t.} \ \frac{1}{m} \sum_{p=1}^{m} \mathbf{h}_p^T C_{pp} \mathbf{h}_p = 1. \end{aligned} \tag{3.3}$$

The orthogonal constraint $(z^{(i)})^T z^{(j)} = 0, i \neq j$ is imposed on the different solutions, which can be obtained by using an iterative algorithm based on LS regression Vía et al (2007). Here, $\mathbf{z}^{(i)} = \frac{1}{m} \sum_{p=1}^{m} \mathbf{z}_p^{(i)}$ and $\mathbf{z}_p^{(i)}$ is a vector of canonical variables projected using the i'th canonical vector in the p'th view.

4 Tensor Canonical Correlation Analysis (TCCA)

In contrast to CCA-MAXVAR Kettenring (1971) and CCA-LS Vía et al (2007), where only the pairwise correlations are considered, we propose tensor CCA (TCCA) for multi-view dimension reduction by exploiting the high-order tensor correlation between all views. The diagram of the multi-view dimension reduction method using the proposed TCCA is shown in Fig. 2. Different kinds of features, such as LAB color histogram (LAB), wavelet texture (WT), and the local SIFT features (SIFT), are first extracted to represent the instances in different views. This leads to multiple feature matrices $\{X_p \in \mathbb{R}^{d_p \times N}\}_{p=1}^m$. Here, m is set at 3 for intuitive illustration without loss of generality. The different sets of features are then used to calculate the data covariance tensor C_{123}, which is subsequently decomposed as a weighted sum of rank-1 tensors, i.e., $C_{123} \approx \sum_{k=1}^{r} \lambda^{(k)} \mathbf{u}_1^{(k)} \circ \mathbf{u}_2^{(k)} \circ \mathbf{u}_3^{(k)}$, where $r \leq \min(d_1, d_2, d_3)$ is the reduced dimension and $\circ$ is the tensor (outer) product. The vectors $\{\mathbf{u}_p^{(k)}\}_{k=1}^r$ are stacked as a transformation matrix U_p, which is used to map

the original high dimensional features into the low dimensional common subspace. The projected features $\{Z_p\}_{p=1}^m$ are concatenated as the final representation of the instances. The details of this technique are given below, but first we briefly introduce several useful notations and concepts of multilinear algebra.

4.1 Notations

Let $\mathcal{A}$ be an m-order tensor of size $I_1 \times I_2 \times \ldots \times I_m$, and U be a $J_p \times I_p$ matrix. The p-mode product of $\mathcal{A}$ and U is then denoted as $\mathcal{B} = \mathcal{A} \times_p U$, which is an $I_1 \times \ldots I_{p-1} \times J_p \times I_{p+1} \times I_m$ tensor with the element

$$\mathcal{B}(i_1, \ldots, i_{p-1}, j_p, i_{p+1}, \ldots, i_m) = \sum_{i_p=1}^{I_m} \mathcal{A}(i_1, i_2, \ldots, i_m) U(j_p, i_p). \tag{4.1}$$

The product of $\mathcal{A}$ and a sequence of matrices $\{U_p \in R^{J_p \times I_p}\}_{p=1}^m$ is a $J_1 \times J_2 \times \ldots \times J_m$ tensor denoted by

$$\mathcal{B} = \mathcal{A} \times_1 U_1 \times_2 U_2 \ldots \times_m U_m. \tag{4.2}$$

The mode-p matricization of $\mathcal{A}$ is denoted as an $I_p \times (I_1 I_{p-1} I_{p+1} I_m)$ matrix $A_{(p)}$, which is obtained by mapping the fibers associated with the p'th dimension of $\mathcal{A}$ as the rows of $A_{(p)}$, and aligning the corresponding fibers of all the other dimensions as the columns. Here, the columns can be ordered in any way. The p-mode multiplication $\mathcal{B} = \mathcal{A} \times_p U$ can be manipulated as matrix multiplication by storing the tensors in metricized form, i.e., $B_{(p)} = U A_{(p)}$. Specifically, the series of p-mode product in (4.2) can be expressed as a series of Kronecker products and is given by

$$B_{(p)} = U_p A_{(p)} (U^{(c_{p-1})} \otimes U^{(c_{p-2})} \otimes \ldots \otimes U^{(c_1)})^T, \tag{4.3}$$

where $\{c_1, c_2, \ldots, c_L\} = \{p+1, p+2, \ldots, m, 1, 2, \ldots, p-1\}$ is a forward cyclic ordering for the indices of the tensor dimensions that map to the column of the matrix. Finally, the Frobenius norm of the tensor $\mathcal{A}$ is given by

$$\|\mathcal{A}\|_F^2 = \langle \mathcal{A}, \mathcal{A} \rangle = \sum_{i_1=1}^{I_1} \sum_{i_2=1}^{I_2} \ldots \sum_{i_m=1}^{I_m} \mathcal{A}(i_1, i_2, \ldots, i_m)^2. \tag{4.4}$$

4.2 Problem Formulation

Given m views $\{X_p\}_{p=1}^m$ of N instances, and each $X_p = [\mathbf{x}_{p1}, \mathbf{x}_{p2}, \ldots, \mathbf{x}_{pN}] \in \mathbb{R}^{d_p \times N}$ is assumed to have been centered (i.e., have zero mean). The variance matrices are then

$$C_{pp} = \frac{1}{N} \sum_{n=1}^{N} \mathbf{x}_{pn} \mathbf{x}_{pn}^T, p = 1, \ldots, m,$$

and the covariance tensor among all views is calculated as

$$\mathcal{C}_{12\ldots m} = \frac{1}{N} \sum_{n=1}^{N} \mathbf{x}_{1n} \circ \mathbf{x}_{2n} \circ \ldots \circ \mathbf{x}_{mn},$$

where $\mathcal{C}$ is a tensor of dimension $d_1 \times d_2 \times \ldots \times d_m$. Following the objective of the traditional two-view CCA Hardoon et al (2004), the proposed tensor CCA seeks to maximize the correlation between the canonical variables $\mathbf{z}_p = X_p^T \mathbf{h}_p, p = 1, \ldots, m$, where $\{\mathbf{h}_p \in \mathbb{R}^{d_p \times 1}\}_{p=1}^m$ are usually called the canonical vectors. Therefore, the optimization problem is

$$\begin{aligned} \underset{\{\mathbf{h}_p\}}{\text{argmax}}\, \rho =& \text{corr}(\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_m) \\ \text{s.t. } & \mathbf{z}_p^T \mathbf{z}_p = 1, p = 1, \ldots, m. \end{aligned} \tag{4.5}$$

Here $\text{corr}(\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_m) = (\mathbf{z}_1 \odot \mathbf{z}_2 \odot \ldots \odot \mathbf{z}_m)^T \mathbf{e}$ is the canonical correlation, and $\odot$ is the element-wise product, $\mathbf{e} \in \mathbb{R}^N$ is an all ones vector. We can prove that it is equivalent to $\mathcal{C}_{12\ldots m} \times_1 \mathbf{h}_1^T \times_2 \mathbf{h}_2^T \ldots \times_m \mathbf{h}_m^T$, where $\times_p$ is the p-mode tensor-matrix product.

Theorem 1. *The high order canonical correlation is given by*

$$\rho = (\mathbf{z}_1 \odot \mathbf{z}_2 \odot \ldots \odot \mathbf{z}_m)^T \mathbf{e} = \mathcal{C}_{12\ldots m} \times_1 \mathbf{h}_1^T \times_2 \mathbf{h}_2^T \ldots \times_m \mathbf{h}_m^T. \tag{4.6}$$

The proof is presented in the Appendix. By further considering that $X_p X_p^T = C_{pp}, p = 1, \ldots, m$, the problem (4.5) becomes

$$\begin{aligned} \underset{\{\mathbf{h}_p\}}{\operatorname{argmax}} \rho =& \mathcal{C}_{12\ldots m} \times_1 \mathbf{h}_1^T \times_2 \mathbf{h}_2^T \ldots \times_m \mathbf{h}_m^T \\ & \text{s.t. } \mathbf{h}_p^T C_{pp} \mathbf{h}_p = 1, p = 1, \ldots, m. \end{aligned} \tag{4.7}$$

We further add a regularization term in the constraints to control the model complexity, and thus the constraints of problem (4.7) become

$$\mathbf{h}_p^T (C_{pp} + \epsilon I) \mathbf{h}_p = 1, p = 1, \ldots, m, \tag{4.8}$$

where I is an identity matrix and ϵ is a nonnegative trade-off parameter. Let each $\mathbf{u}_p = \tilde{C}_{pp}^{1/2} \mathbf{h}_p$ and $\mathcal{M} = \mathcal{C}_{12\ldots m} \times_1 \tilde{C}_{11}^{-1/2} \times_2 \tilde{C}_{22}^{-1/2} \ldots \times_m \tilde{C}_{mm}^{-1/2}$, we can reformulate (4.7) as

$$\begin{aligned} \underset{\{\mathbf{u}_p\}}{\operatorname{argmax}} \rho =& \mathcal{M} \times_1 \mathbf{u}_1^T \times_2 \mathbf{u}_2^T \ldots \times_m \mathbf{u}_m^T \\ & \text{s.t. } \mathbf{u}_p^T \mathbf{u}_p = 1, p = 1, \ldots, m, \end{aligned} \tag{4.9}$$

where $\tilde{C}_{pp} = C_{pp} + \epsilon I$. The equivalence of the problem (4.7) and (4.9) is ensured by the following theorem.

Theorem 2. *The problems (4.7) and (4.9) are equivalent.*

Proof. It is straightforward that the constraints of problems (4.7) and (4.9) are equivalent, and now we prove that the objective of the two problems is the same as follows,

$$\begin{aligned} & \mathcal{C}_{12\ldots m} \times_1 \mathbf{h}_1^T \times_2 \mathbf{h}_2^T \times_m \mathbf{h}_m^T \\ =& \mathbf{h}_m^T C_{(m)} (\mathbf{h}_{m-1} \otimes \ldots \otimes \mathbf{h}_2 \otimes \mathbf{h}_1) \\ =& \mathbf{u}_m^T \tilde{C}_{mm}^{-1/2} C_{(m)} ((\tilde{C}_{m-1,m-1}^{-1/2} \mathbf{u}_{m-1}) \otimes \ldots \otimes (\tilde{C}_{11}^{-1/2} \mathbf{u}_1)) \\ =& \mathbf{u}_m^T (\tilde{C}_{mm}^{-1/2} C_{(m)} (\tilde{C}_{m-1,m-1}^{-1/2} \otimes \ldots \otimes \tilde{C}_{11}^{-1/2}) (\mathbf{u}_{m-1} \otimes \ldots \otimes \mathbf{u}_1)) \\ =& \mathbf{u}_m^T \mathcal{M} (\mathbf{u}_{m-1} \otimes \ldots \otimes \mathbf{u}_2 \otimes \mathbf{u}_1) \\ =& \mathcal{M} \times_1 \mathbf{u}_1^T \times_2 \mathbf{u}_2^T \ldots \times_m \mathbf{u}_m^T, \end{aligned}$$

where the metricizing property of the tensor-matrix product presented in (4.3) and some basic properties of the Kronecker product are applied. □

4.3 Solutions

It has been presented in De Lathauwer et al (2000b) that the problem (4.9) is equivalent to finding the best rank-1 approximation of the tensor $\mathcal{M}$, i.e., if we define $\hat{\mathcal{M}} = \rho \mathbf{u}_1 \circ \mathbf{u}_2 \circ \ldots \circ \mathbf{u}_m$, then the optimization problem becomes

$$\underset{\{\mathbf{u}_p\}}{\operatorname{argmin}} \|\mathcal{M} - \hat{\mathcal{M}}\|_F^2. \tag{4.10}$$

The solution can be obtained using the alternating least square (ALS) algorithm Kroonenberg and De Leeuw (1980); Comon et al (2009). Some other algorithms, such as the high-order power method (HOPM) De Lathauwer et al (2000b) and the tensor power method Allen (2012), can also be applied here for optimization, but our empirical findings indicate that the ALS algorithm performs the best in our experiments.

As in the two-view CCA, we perform a recursive maximization of the correlation between linear combinations of $X_p, p = 1, \ldots, m$. However, we cannot expect the different linear combinations $\mathbf{h}_p^{(1)}, \ldots, \mathbf{h}_p^{(r)}$ of X_p to be uncorrelated with each other, where r is the rank of $\mathcal{M}$ (the determination of the rank value is still an

open problem for the high-order tensors De Lathauwer et al (2000a)). That is, the orthogonality constraints cannot be imposed on $\mathbf{u}_p^{(1)}, \ldots, \mathbf{u}_p^{(r)}$, since the sum of rank-1 decomposition and orthogonal decomposition of high-order tensors cannot be satisfied simultaneously De Lathauwer et al (2000a).

Based on the solutions $\mathbf{u}_p$, we obtain the canonical variables $\mathbf{z}_p = X_p^T \mathbf{h}_p = X_p^T \tilde{C}_{pp}^{-1/2} \mathbf{u}_p$. Let $U_p = [\mathbf{u}_p^{(1)}, \ldots, \mathbf{u}_p^{(r)}]$ and $\mathbf{z}_p^{(1)}, \ldots, \mathbf{z}_p^{(r)}$ be the column vectors of Z_p, we obtain the projected data for the p'th view:

$$Z_p = X_p^T \tilde{C}_{pp}^{-1/2} U_p. \tag{4.11}$$

Following Foster et al (2008), where it is suggested that the dimension be reduced to $2r$ in the standard CCA, we concatenate the different $\{Z_p\}$ as the final representation $Z \in \mathbb{R}^{(mr)\times N}$ for the subsequent learning, such as classification Farquhar et al (2005); Fisch et al (2014), clustering Yang et al (2014); Wu et al (2015), regression Kakade and Foster (2007), search ranking Xu et al (2015); Zhu et al (2015), collaborative filtering Liu et al (2014), and so on.

4.4 Non-linear Extension

The projections $\{\mathbf{h}_p\}$ are linear in TCCA and thus may be not appropriate for instances that lie in quite non-linear feature space. To this end, we develop kernel tensor CCA (KTCCA) that extends the proposed TCCA to the non-linear case. KTCCA aims to find non-linear projections by first projecting the data into higher dimensional space induced by the feature mapping ϕ:

$$\phi(X_p) = [\phi(\mathbf{x}_{p1}), \phi(\mathbf{x}_{p2}), \ldots, \phi(\mathbf{x}_{pN})] \in \mathbb{R}^{D_p \times N},$$

where the mapped dimension D_p may be infinite. Then the variance matrices

$$C_{pp} = \frac{1}{N} \sum_{n=1}^{N} \phi(\mathbf{x}_{pn}) \phi^T(\mathbf{x}_{pn}), p = 1, \ldots, m,$$

the covariance matrix

$$\mathcal{C}_{12\ldots m} = \frac{1}{N} \sum_{n=1}^{N} \phi(\mathbf{x}_{1n}) \circ \phi(\mathbf{x}_{2n}) \circ \ldots \circ \phi(\mathbf{x}_{mn}),$$

and the canonical variables $\mathbf{z}_p = \phi^T(X_p)\mathbf{h}_p$. It follows from the Representer Theorem Scholkopf and Smola (2002) that $\mathbf{h}_p$ can be rewritten as a linear combination of the given instances, i.e.,

$$\mathbf{h}_p = \phi(X_p)\mathbf{a}_p, \tag{4.12}$$

where $\mathbf{a}_p \in \mathbb{R}^{N\times 1}$ is a vector of the combination coefficients. The problem (4.7) then becomes

$$\begin{aligned} \underset{\{\mathbf{a}_p\}}{\operatorname{argmax}}\, \rho =& \mathcal{K}_{12\ldots m} \times_1 \mathbf{a}_1^T \times_2 \mathbf{a}_2^T \ldots \times_m \mathbf{a}_m^T \\ \text{s.t. } & \mathbf{a}_p^T K_{pp} \mathbf{h}_p = 1, p = 1, \ldots, m, \end{aligned} \tag{4.13}$$

where $K_{pp} = \phi^T(X_p)\phi(X_p)$ is the kernel matrix of the p'th view. The derivation is similar to Theorem 2. Here $\mathcal{K}_{12\ldots m} = \mathcal{C}_{12\ldots m} \times_1 \phi^T(X_1) \times_2 \phi^T(X_2) \ldots \times_m \phi^T(X_m)$ and can be calculated according to the following theorem.

Theorem 3. *The following equality holds:*

$$\mathcal{C}_{12\ldots m} \times_1 \phi^T(X_1) \times_2 \phi^T(X_2) \ldots \times_m \phi^T(X_m) = \frac{1}{N} \sum_{n=1}^{N} \mathbf{k}_{1n} \circ \mathbf{k}_{2n} \circ \ldots \circ \mathbf{k}_{mn},$$

in which $\mathbf{k}_{pn} = \phi^T(X_p)\phi(\mathbf{x}_{pn})$, *i.e., the n'th column of the kernel matrix* K_{pp}, $p = 1, \ldots, m$.

We give the proof in the Appendix. To avoid trivial learning, we follow Hardoon et al (2004) and introduce a partial least square (PLS) term to penalize the norms of the weight vectors $\{\mathbf{a}_p\}$. That is, the constraints of problem (4.13) become

$$\mathbf{a}_p^T(K_{pp}^2 + \epsilon K_{pp})\mathbf{a}_p = 1, p = 1, \ldots, m. \tag{4.14}$$

Because the matrix $(K_{pp}^2 + \epsilon K_{pp})$ is positive definite, it has a unique Cholesky decomposition, and we can denote its decomposition as $(K_{pp}^2 + \epsilon K_{pp}) = L_p^T L_p$. Let $\mathbf{b}_p = L_p\mathbf{a}_p$ and $\mathcal{S} = \mathcal{K}_{12\ldots m} \times_1 (L_1^{-1})^T \times_2 (L_2^{-1})^T \ldots \times_m (L_m^{-1})^T$, we can reformulate (4.13) as

$$\begin{aligned} \underset{\{\mathbf{b}_p\}}{\operatorname{argmax}}\, \rho =& \mathcal{S} \times_1 \mathbf{b}_1^T \times_2 \mathbf{b}_2^T \ldots \times_m \mathbf{b}_m^T \\ & \text{s.t. } \mathbf{b}_p^T\mathbf{b}_p = 1, p = 1, \ldots, m, \end{aligned} \tag{4.15}$$

Similar to TCCA, this problem is equivalent to finding the best rank-1 approximation of $\mathcal{S}$, and the solution can be found using the ALS algorithm. By recursively maximizing the correlation, we obtain $\mathbf{b}_p^{(1)}, \ldots, \mathbf{b}_p^{(r)}$. Let $B_p = [\mathbf{b}_p^{(1)}, \ldots, \mathbf{b}_p^{(r)}]$, and the canonical variables $\mathbf{z}_p = \phi^T(X_p)\mathbf{h}_p = \phi^T(X_p)\phi(X_p)\mathbf{a}_p = K_{pp}L_p^{-1}\mathbf{b}_p$ and the projected data for the p'th view are then

$$Z_p = K_p p L_p^{-1} B_p. \tag{4.16}$$

The concatenated $Z \in \mathbb{R}^{(mr)\times N}$ is the final representation of the instances.

4.5 Complexity Analysis

The time and space complexities of the proposed TCCA model are both closely related to the size of tensor $\mathcal{M}$. Straightforwardly, the space complexity is $O(d_1 d_2 \ldots d_m)$. Because the tensor $\mathcal{M}$ can be calculated offline, the time complexity is dominated by the rank-r decomposition using the ALS algorithm. Considering that it is common that $r \ll \min(d_1, d_2, \ldots, d_m)$, we can speculate the time complexity of ALS is $O(trd_1 d_2 \ldots d_m)$ according to Comon et al (2009), where the time cost of the ALS algorithm for the three modes tensor is presented. Here, t is the number of iterations in ALS.

According to the above analysis, we can see that the complexity of TCCA is independent of the number of instances, and thus our method can be scaled in very large sample size problems. Similarly, the complexities of KTCCA are determined by the tensor $\mathcal{S}$, the size of which is N^m. The space and time complexities are $O(N^m)$ and $O(trN^m)$ respectively. This means that KTCCA is capable of being scaled in problems that have very high feature dimensions and a small number of instances.

5 Experiments

In this section, we empirically validate the effectiveness of the proposed TCCA on a biometric structure prediction and an advertisement classification problem following Foster et al (2008), as well as on a challenging web image annotation task Chua et al (2009). In all of the following experiments, five random choices of the labeled instances are used. Twenty percent of the test data (or unlabeled data in the transductive setting) are used for validation, which means that the parameters (if not specified) corresponding to the best performance on the validation set are used for testing. The evaluation criterion is the classification accuracy.

5.1 Evaluation of the Linear Formulation

In the first two sets of experiments (biometric structure prediction and advertisement classification), we use regularized least squares (RLS) as the base learner following Foster et al (2008). Given N_l labeled instances $\{(\mathbf{x}_n, y_n)\}_{n=1}^{N_l}$, the optimization problem for RLS is given by $\operatorname{argmin}_{\mathbf{w}} \frac{1}{N_l}\sum_{n=1}^{N_l}(\mathbf{w}^T\mathbf{x}_n - y_n)^2 + \gamma\|\mathbf{w}\|_2^2$, where the positive trade-off parameter γ is is set as 10^{-2} according to Foster et al (2008). A constant feature of 1 is appended to each instance to include a bias term in $\mathbf{w}$. In web image annotation, the k-nearest-neighbor (kNN) classifier is utilized, where the candidate set for k is $\{1, 2, \ldots, 10\}$. Specifically, we compare the following methods:

- **BSF:** using the single view feature that achieves the best performance in RLS/kNN-based classification.
- **CAT:** concatenating the normalized features of all the views into a long vector, and then performing RLS/kNN-based classification.
- **CCA Foster et al (2008):** using the CCA formulation presented in Foster et al (2008) to find a common representation of two different views. In this formulation, a regularization term ϵI is added to control the model complexity, and we set the parameter ϵ as 10^{-2} in biometric structure prediction and advertisement classification according to Foster et al (2008). The parameter is tuned over the set $\{10^i | i = -5, \ldots, 4\}$ in web image annotation. The implementation details can be found in Foster et al (2008). For m different views, there are $m(m-1)/2$ subsets of two views. The subset that achieves the best performance is termed **CCA (BST)**. To combine the results of all subsets, we average their predicted scores in RLS-based classification and adopt the majority voting strategy in kNN. This combination approach is termed **CCA (AVG)**.
- **CCA-LS Vía et al (2007):** a generalization of CCA to multiple views based on least square (LS) regression.
- **DSE Long et al (2008):** a general and popular unsupervised multi-view dimension reduction method based on spectral embedding.
- **SSMVD Han et al (2012):** a recently proposed unsupervised multi-view dimension reduction method based on the structured sparsity-inducing norm Jenatton et al (2011).
- **TCCA:** the proposed tensor CCA. The regularization parameter ϵ is optimized the same as in CCA.

In the first step of DSE and SSMVD, PCA is taken as the dimension reduction method for each view, and the result dimension (of each view) is set to be 100 empirically.

5.1.1 Biometric Structure Prediction

The dataset used in this set of experiments is SecStr[1], which is a benchmark dataset for evaluating semi-supervised systems Chapelle et al (2006). The task associated with this dataset is "to predict the secondary structure of a given amino acid in a protein based on a sequence window centered around that amino acid" Chapelle et al (2006). The SecStr dataset is large-scale and contains $84K$ instances. We randomly select 100 instances as labeled samples. There are also $1200K$ unlabeled instances which we use to observe the performance of three CCA-based methods (CCA, CCA-LS and TCCA) with respect to different amounts of unlabeled data. Following Foster et al (2008), all the provided data are used (as unlabeled instances) to find the common subspace in the CCA-based methods. The performance is evaluated in a transductive setting on the unlabeled samples (except those for validation) of the $84K$ instances. Both DSE and SSMVD are naturally transductive, since they learn the low-dimensional representation of given data directly, and no projection matrix is learned for new data. Therefore, these two methods cannot handle very large datasets and the experiments are conducted only on the $84K$ instances. In particular, DSE needs to solve an eigen-decomposition problem of size $N \times N$. The time cost or memory cost is intolerable when N is $84K$, and thus a subset of $10K$ samples are utilized.

The features provided are 15 categorical attributes, each of which is generated at a position in $[-7, +7]$ from the sequence window of amino acid, and represented by a 21-dimensional sparse binary vector. We divided the $315(15 \times 21)$ features into three views:

- View-1: attributes based on the left context (positions in $[-7, -3]$);
- View-2: attributes based on the current position and middle context (positions in $[-2, 2]$);
- View-3: attributes based on the right context (positions in $[3, 7]$).

The dimension of each view is 105.

[1] http://www.kyb.tuebingen.mpg.de/ssl-book

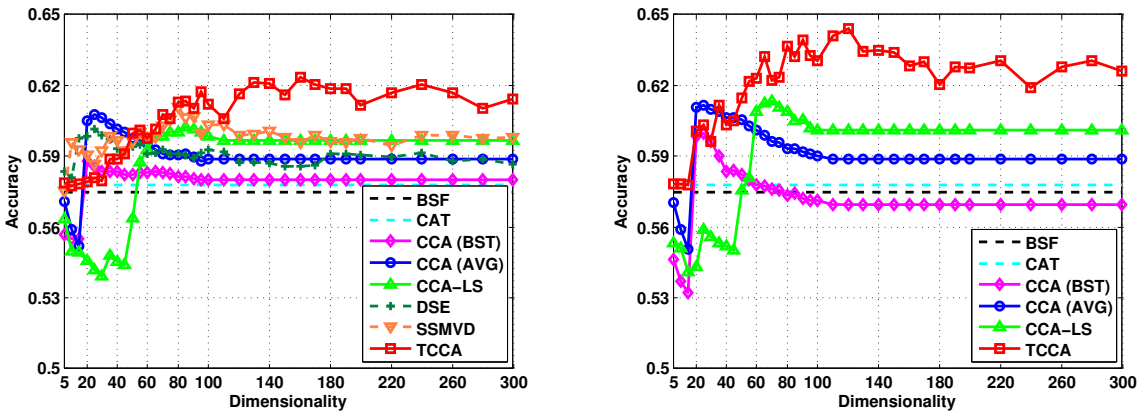

Figure 3: Prediction accuracy vs. dimension of the common subspace on the SecStr dataset. (Top: 100 labeled instances and $84K$ unlabeled instances; Bottom: 100 labeled instances and the entire unlabeled set (about $1.3M$ instances).)

Table 1: Prediction accuracies (%) of the different methods at their best dimensions on the SecStr dataset (100 labeled instances).

Methods	#unlabeled = $84K$	#unlabeled = $1.3M$
BSF	57.48±1.90	
CAT	57.77±2.03	
CCA (BST)	58.78±2.97	59.97±2.46
CCA (AVG)	60.75±1.92	61.15±1.73
CCA-LS	60.23±1.70	61.32±1.65
DSE	60.15±0.81	No Attempt
SSMVD	61.08±1.58	
TCCA	**62.36±1.27**	**64.42±1.70**

The performance of the compared methods in relation to the dimension of the common subspace is shown in Fig. 3. Accuracy is averaged over 5 runs for each dimension r in $\{5, 10, \ldots, 100, 110, \ldots, 200, 220, \ldots, 300\}$. The performance of the different methods at their best dimensions are summarized in Table 1. From the results, we observe that: 1) the concatenation strategy (CAT) is comparable to and slightly better than the strategy of only using the best single view features (BSF); 2) by learning the common subspace, all the compared multi-view dimension reduction methods are significantly better than the BSF and CAT baselines, if the dimensionalities are properly set according to the accuracy on the validation dataset. In particular, CCA (BST) is superior to CAT, although only a subset of two views is utilized in the former; 3) the accuracy of all three CCA-based methods increases with an increasing number of unlabeled data. By combining the results of different subsets, CCA (AVG) is better than CCA (BST); 4) CCA-LS is superior to CCA (BST), but their performance at their best dimension is comparable. When the number of unlabeled data is $84K$, DSE and SSMVD are comparable to CCA (BST) and CCA-LS respectively; 5) the performance of TCCA does not decease significantly as CCA-LS and CCA do when the number of dimensions is high. The main reason is that the ALS algorithm used in TCCA seeks to maximize the canonical correlations for all the r factors simultaneously, but not to greedily find orthogonal decomposition components Allen (2012). That is, the main variance tends to be explained uniformly by all factors, not only by the first several factors. This is also the reason why there are some oscillations in TCCA; 6) the proposed TCCA significantly outperforms all the other methods on most dimensionalities. This demonstrates that the high order correlation information between all features is well discovered, and that exploring this kind of information is much better than only

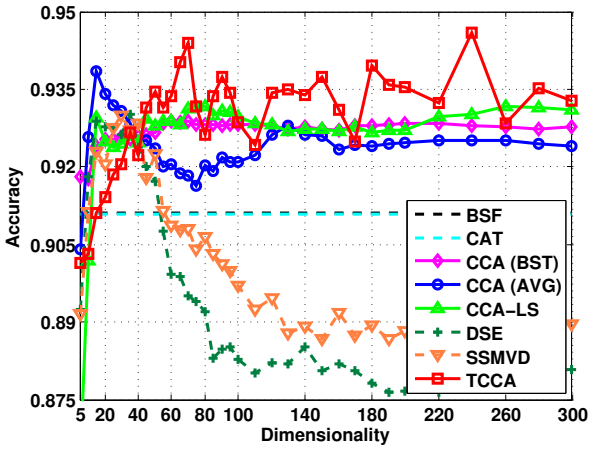

Figure 4: Classification accuracy vs. dimension of the common subspace on the Ads dataset. 100 labeled training samples are utilized.

exploring the correlation information between pairs of features, as in CCA-LS.

5.1.2 Advertisement Classification

This set of experiments is conducted on the Ads (internet advertisements)[2] dataset from the well-known UCI Machine Learning Repository. The task is to predict whether or not a given hyperlink (associated with an image) is an advertisement. There are $3,279$ instances in this dataset. We randomly choose 100 instances as labeled training samples, and all the instances except those for validation are utilized as unlabeled samples to find the common subspace. The performance is evaluated in a transductive setting on the unlabeled samples.

We use the features as described in Kushmerick (1999), and omit the attributes that have missing values, such as the height (and width) of the image. The remained attributes are represented by binary (1/0) features which indicate the presence/absence of corresponding terms. For CCA-LS and TCCA, we divide all these features into three views as follows:

- View-1: features based on the terms in the images URL, caption, and alt text. 588 dimensions;
- View-2: features based on the terms in the URL of the current site. 495 dimensions;
- View-3: features based on the terms in the anchor URL. 472 dimensions.

Fig. 4 shows the classification accuracy of the compared methods (in relation to the dimension r), and the accuracies at their best dimensions are summarized in Table 2. In contrast to the observations of the last set of experiments, we can see that: 1) the accuracy of the concatenation strategy (CAT) and the best single view (BSF) are almost the same. The performance of CAT is relatively worse since the feature dimension in this set of experiments is high ($1,555$ dimensions), and over-fitting occurs given the limited number of labeled samples; 2) the performance of DSE and SSMVD first increase and then decrease sharply with an increasing number of the dimension r, while the CCA-based methods are much steady; 3) the improvement of TCCA compared with the other CCA-based methods is not as great as in the last set of experiments. This is because we need more samples to approximate the true underlying high order correlation compared with the traditional pairwise correlation, since there are more variables to be estimated in the high order statistics. The unlabeled instances utilized in this set of experiments are much fewer, thus the high order correlation information is not well explored. CCA-LS is only comparable to CCA for the same reason.

[2] http://archive.ics.uci.edu/ml/datasets/Internet+Advertisements

Table 2: Classification accuracies (%) of the different methods at their best dimensions on the Ads dataset.

Methods	#labeled = 100
BSF	91.10±1.65
CAT	91.08±1.74
CCA (BST)	92.88±1.11
CCA (AVG)	93.84±0.85
CCA-LS	93.17±1.10
DSE	93.01±0.96
SSMVD	92.99±0.91
TCCA	**94.59±0.27**

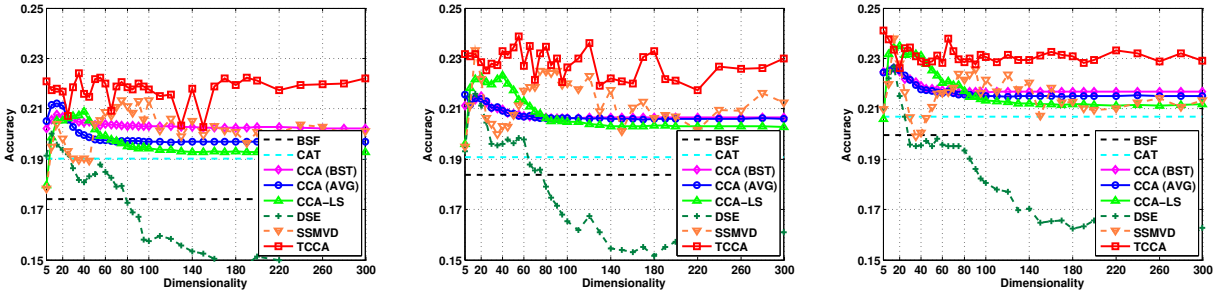

Figure 5: Anotation accuracy vs. dimension of the common subspace on the NUS-WIDE mammal subset. (Left: 4 labeled instances for each mammal concept; Middle: 6 labeled instances; Right: 8 labeled instances.)

5.1.3 Web Image Annotation

We further verify the effectiveness of the proposed algorithm on a natural image dataset NUS-WIDE Chua et al (2009). This dataset contains 269, 648 images, and our experiments are conduct on a subset that consists of 11, 189 images belonging to 10 mammal concepts: bear, cat, cow, dog, elk, fox, horse, tiger, whale, and zebra. We randomly split the images into a training set of 5, 597 images and a test set of 5, 592 images. Distinguishing between these concepts is very challenging, since many of them are similar to each other, e.g., cat and tiger. We randomly choose $\{4, 6, 8\}$ labeled instances for each concept in the training set, and all the training instances are utilized as unlabeled samples to find the common subspace.

In this dataset, we choose three types of visual feature, namely 500-D bag of visual words based on SIFT Lowe (2004) descriptors, 144-D color auto-correlogram, and 128-D wavelet texture, to represent each image Chua et al (2009).

The annotation performance of the compared methods is shown in Fig. 5 and Table 3. It can be seen from the results that: 1) in general, performance improves with an increased number of labeled instances; 2) CCA-LS is comparable to CCA (BST) and CCA (AVG), while the best performance (peak of the curve) of CCA-LS is usually higher; 3) the performance of DSE is poor when r is large, while SSMVD is much steady and can be superior to CCA (AVG) and CCA-LS sometimes; 4) the accuracies of CCA (AVG) and CCA-LS first increase and then decrease with an increasing number of the dimension r, while the results of the proposed TCCA are satisfactory even though r is large; 3) the accuracy of TCCA is significantly better than that of all the other methods under most dimensionalities.

Table 3: Annotation accuracies (%) of the different methods at their best dimensions on the NUS-WIDE mammal dataset.

Methods	#labeled = 4	#labeled = 6	#labeled = 8
BSF	17.42±1.37	18.37±1.21	19.96±1.19
CAT	19.01±1.86	19.07±2.23	20.70±1.44
CCA (BST)	20.77±1.52	21.51±2.38	22.61±1.76
CCA (AVG)	21.21±1.47	21.57±2.04	22.61±1.21
CCA-LS	20.90±1.84	22.31±2.53	23.50±2.48
DSE	20.02±1.23	21.59±1.13	22.67±0.74
SSMVD	21.34±2.08	23.32±1.08	23.79±1.30
TCCA	**22.40±1.96**	**23.86±1.41**	**24.11±0.32**

5.2 Evaluation of the Non-linear Extension

We evaluate the non-linear extension of the proposed TCCA in the web image annotation task. As discussed in Section 4.5, the non-linear extension is able to handle the small sample size problem, where the feature dimensions can be very high and possibly infinite. We thus randomly choose a small set of 500 samples from the animal subset. To perform the non-linear classification, we construct a kernel for each kind of feature. The kernel is defined by

$$k(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\lambda^{-1} d(\mathbf{x}_i, \mathbf{x}_j)),$$

where $d(\mathbf{x}_i, \mathbf{x}_j)$ denotes the distance between $\mathbf{x}_i$ and $\mathbf{x}_j$, and $\lambda = \max_{i,j} d(\mathbf{x}_i, \mathbf{x}_j)$. We choose the χ^2 distance for the visual word histogram. For other features, the $L2$ distance is utilized. Specifically, we compare the following methods:

- **BSK:** using the single view kernel that achieves the best performance in the kNN-based classification.
- **AVG:** averaging the normalized kernels of all the views, and then performing kNN-based classification.
- **KCCA Hardoon et al (2004):** using the KCCA formulation presented in Hardoon et al (2004) to find a common representation of two different views. The regularization parameter is optimized over the set $\{10^i | i = -7, \ldots, 2\}$. The setup of **KCCA (BST)** and **KCCA (AVG)** are similar as **CCA (BST)** and **CCA (AVG)** in the experiments of the linear version.
- **KTCCA:** the non-linear extension of the proposed tensor CCA. The regularization parameter ϵ is optimized in the same way as in KCCA.

The experimental results are shown in Fig. 6 and Table 4. Compared with the results in Fig. 5, we can see that: 1) although a small number of unlabeled samples is utilized, the performance is better since the separability is improved by the non-linear projection, which is implemented via the kernel trick Shawe-Taylor and Cristianini (2004); 2) the simple AVG view combination strategy outperforms the best single view kernel (BSK) significantly, and is comparable to KCCA (BST); 3) KCCA (AVG) is slightly better than KCCA (BST), and the proposed KTCCA achieves the best performance under most dimensionalities.

5.3 Empirical analysis of the computational complexity

In this subsection, we empirically analyze the computational complexity of the different methods. The experiments are conducted in Matlab R2012b on a 2×3.33 GHz Intel Xeon (6 cores) computer, where the memory is 48GB 1333MHz ECC DDR3-RAM. The results (time cost and memory cost) on the different datasets are shown in Fig. 7-10. From the results, we observe that: 1) the costs of the proposed TCCA are higher than the other CCA-based methods in general. This is because the decomposition is performed on a large $d_1 \times d_2 \times \ldots \times d_m$ covariance tensor, instead of one or multiple $d_p \times d_q$ covariance matrices, where $p, q = 1, \ldots, m$ are the view indices. The tensor decomposition method we adopt in this paper is the ALS

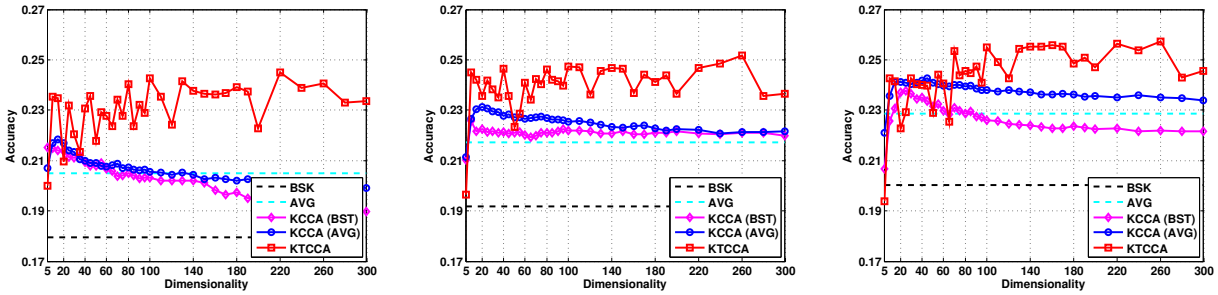

Figure 6: Annotation accuracy (of the non-linear methods) vs. dimension of the common subspace on the NUS-WIDE mammal subset, where a small set of 500 samples is utilized. (Left: 4 labeled instances for each mammal concept; Middle: 6 labeled instances; Right: 8 labeled instances.)

Table 4: Annotation accuracies (%) of the different non-linear methods at their best dimensions on the NUS-WIDE mammal dataset.

Methods	#labeled = 4	#labeled = 6	#labeled = 8
BSK	17.96±1.29	19.17±2.01	20.04±1.66
AVG	20.49±1.65	21.73±2.74	22.86±1.87
KCCA (BST)	21.51±2.44	22.58±1.91	23.78±1.57
KCCA (AVG)	21.85±1.38	23.13±1.77	24.28±1.04
KTCCA	**24.51±0.78**	**25.18±0.58**	**25.74±0.90**

algorithm Kroonenberg and De Leeuw (1980); Comon et al (2009), which could result in satisfactory accuracy but is not efficient; 2) TCCA is much more efficient than DSE or SSMVD when the feature dimensions are not very high and the number of instances is large (see Fig. 7 for example). This demonstrates the superiority of TCCA compared with the existed unsupervised multi-view dimension reduction methods on the large sample size problems.

6 Conclusion

Standard CCA cannot deal with multi-view data, and its typical multi-view extensions ignore the high order statistics (correlation information) among all feature views. To resolve this problem, we have presented tensor CCA (TCCA) to discover such statistics by analyzing the covariance tensor of all views.

From the experimental validation on a variety of application tasks, we conclude that: 1) finding a common subspace for all views using the CCA-based strategy is often better than simply concatenating all the features, especially when the feature dimension is high; 2) examining more statistics, which may require more unlabeled data to be utilized, often leads to better performance; 3) by exploring the high order statistics, the proposed TCCA outperforms the other methods, especially when the dimension of the common subspace is high.

Compared with CCA and its traditional multi-view extensions, the main disadvantage of the proposed TCCA is the high computational cost. Most of the TCCA cost lies in the tensor decomposition, which is not the point of this paper. In the future, we will devote efficient tensor decomposition methods that could speed up TCCA, or introduce the parallel computing technique by utilizing GPU to accelerate the ALS tensor decomposition.

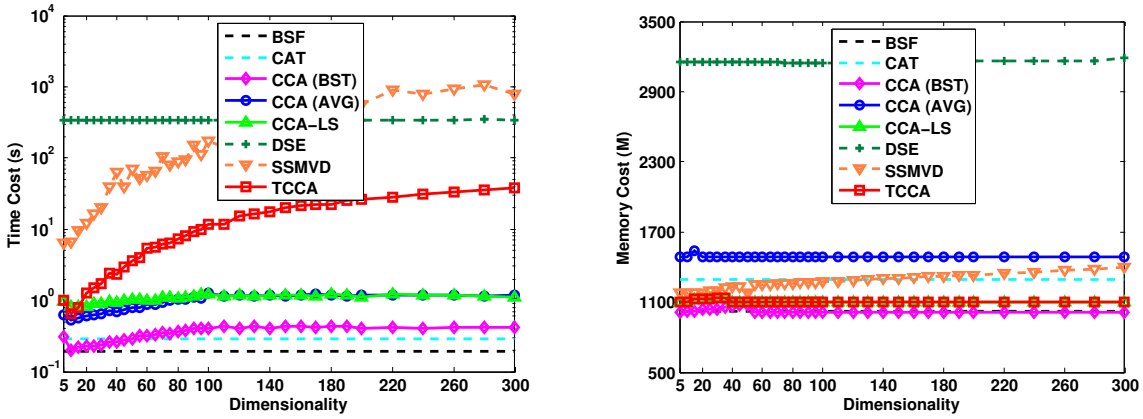

Figure 7: Computational complexity vs. dimension of the common subspace on the SecStr dataset. (Top: time cost in seconds; Bottom: memory cost in Megabits.)

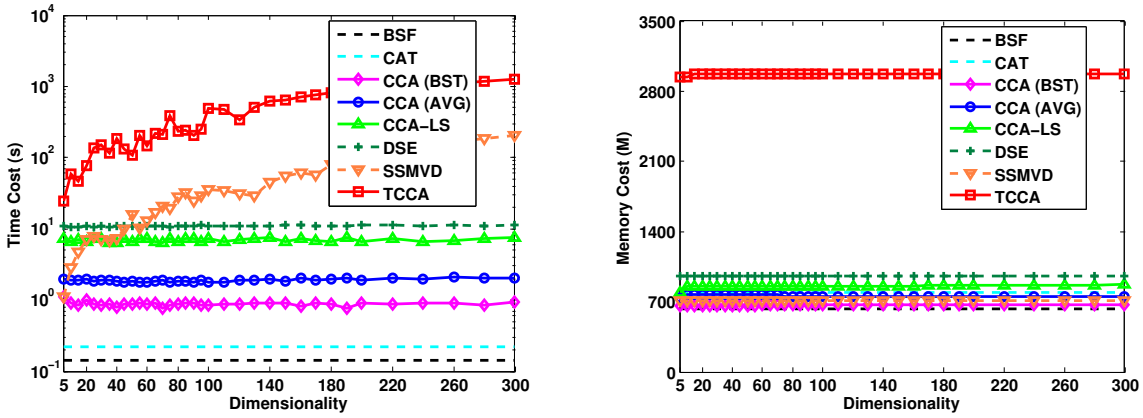

Figure 8: Computational complexity vs. dimension of the common subspace on the Ads dataset. 100 labeled training samples are utilized. (Top: time cost in seconds; Bottom: memory cost in Megabits.)

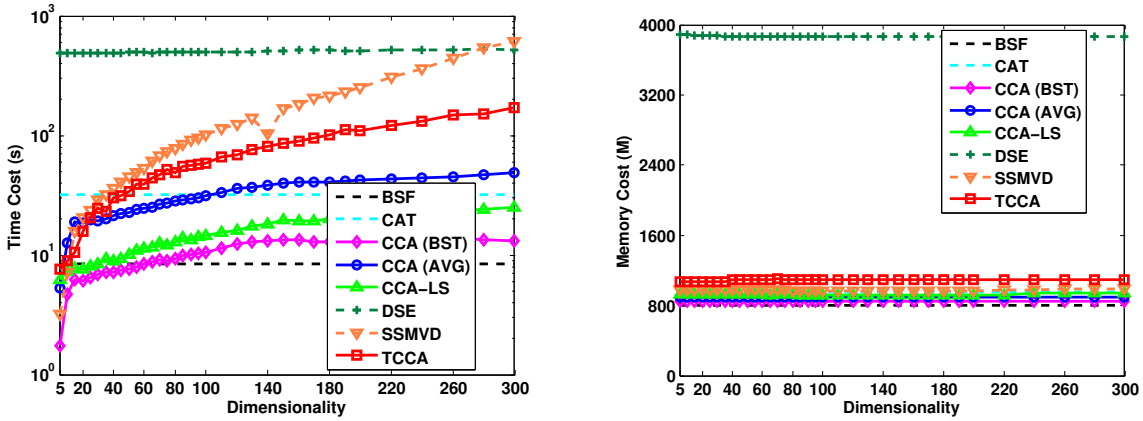

Figure 9: Computational complexity vs. dimension of the common subspace on the NUS-WIDE mammal subset. 6 labeled samples for each mammal concept are utilized. (Top: time cost in seconds; Bottom: memory cost in Megabits.)

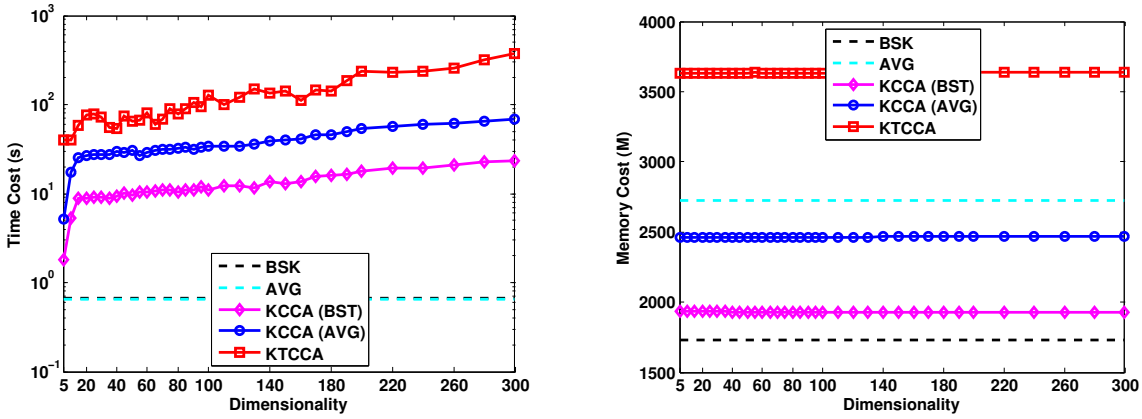

Figure 10: Computational complexity (of the non-linear methods) vs. dimension of the common subspace on the NUS-WIDE mammal subset, where a small set of 500 instances and 6 labeled samples for each mammal concept are utilized. (Top: time cost in seconds; Bottom: memory cost in Megabits.)

A Proof of Thoerem 1

Proof. According to the definition of the element-wise product, we have

$$\rho = (\mathbf{z}_1 \odot \mathbf{z}_2 \odot \ldots \odot \mathbf{z}_m)^T \mathbf{e} = \sum_{n=1}^{N} \mathbf{z}_1(n)\mathbf{z}_2(n)\ldots\mathbf{z}_m(n) = \sum_{n=1}^{N}\prod_{p=1}^{m}\mathbf{z}_p(n) = \sum_{n=1}^{N}\prod_{p=1}^{m}\left(\sum_{j_p=1}^{d_p}\mathbf{x}_{pn}(j_p)\mathbf{h}(j_p)\right), \tag{A.1}$$

where $\mathbf{z}_p(n)$ denotes the n'th entry of the vector $\mathbf{z}_p$, and the same notation is used for $\mathbf{x}_{pn}$ and $\mathbf{h}$. Additionally,

$$\mathcal{C}_{12\ldots m}(j_1, j_2, \ldots, j_m) = \sum_{n=1}^{N}\mathbf{x}_{1n}(j_1)\mathbf{x}_{2n}(j_2)\ldots\mathbf{x}_{mn}(j_m) = \sum_{n=1}^{N}\prod_{p=1}^{m}\mathbf{x}_{pn}(j_p). \tag{A.2}$$

According to the definition of the p-mode product of a tensor and matrix, we have

$$\begin{aligned}&(\mathcal{C} \times_p \mathbf{h}_p^T)(j_1, \ldots, j_{p-1}, 1, j_{p+1}, \ldots, j_m)\\ &= \sum_{j_p=1}^{d_p}\mathcal{C}(j_1, j_2, \ldots, j_m)\mathbf{h}(j_p) = \sum_{j_p=1}^{d_p}\left(\sum_{n=1}^{N}\prod_{p=1}^{m}\mathbf{x}_{pn}(j_p)\right)\mathbf{h}(j_p) = \sum_{n=1}^{N}\sum_{j_p=1}^{d_p}\left(\prod_{p=1}^{m}\mathbf{x}_{pn}(j_p)\right)\mathbf{h}(j_p).\end{aligned} \tag{A.3}$$

Therefore,

$$(\mathcal{C} \times_1 \mathbf{h}_1^T \times_2 \mathbf{h}_2^T \ldots \times_m \mathbf{h}_m^T)(1, \ldots, 1, 1, 1, \ldots, 1) = \sum_{n=1}^{N}\prod_{p=1}^{m}\left(\sum_{j_p=1}^{d_p}\mathbf{x}_{pn}(j_p)\mathbf{h}(j_p)\right). \tag{A.4}$$

This completes the proof. □

B Proof of Theorem 3

Proof. Let $\mathcal{F} = \mathcal{C}_{12\ldots m} \times_1 \phi^T(X_1) \times_2 \phi^T(X_2) \ldots \times_m \phi^T(X_m)$ and $\mathcal{G} = \frac{1}{N}\sum_{n=1}^{N}\mathbf{k}_{1n} \circ \mathbf{k}_{2n} \circ \ldots \circ \mathbf{k}_{mn}$, then according to the definition of the outer product, the $(j_1, j_2, \ldots, j_m)$'th entry of $\mathcal{G}$ is

$$\mathcal{G}(j_1, j_2, \ldots, j_m) = \sum_{n=1}^{N}\mathbf{k}_{1n}(j_1)\mathbf{k}_{2n}(j_2)\ldots\mathbf{k}_{mn}(j_m), \tag{B.1}$$

where $\mathbf{k}_{pn}(j_p)$ is the j_p'th element of the vector $\mathbf{k}_{pn}, p = 1, \ldots, m$. Additionally, the $(j_1, j_2, \ldots, j_m)$'th entry of $\mathcal{C}$ is

$$\mathcal{C}(j_1, j_2, \ldots, j_m) = \sum_{n=1}^{N} \phi_{1n}(i_1)\phi_{2n}(i_2)\ldots\phi_{mn}(i_m), \tag{B.2}$$

where $\phi_{pn}(i_p)$ is the i_p'th element of the vector $\phi(x_{pn}), p = 1, \ldots, m$. According to the definition of the tensor-matrix product, we have

$$\begin{aligned}
&(\mathcal{C} \times_p \phi^T(X_p))(i_1, \ldots, i_{p-1}, j_p, i_{p+1}, \ldots, i_m) \\
=&\sum_{i_p=1}^{D_p} C(i_1, i_2, \ldots, i_m)\phi_p^T(j_p, i_p) \\
=&\sum_{i_p=1}^{D_p}\sum_{n=1}^{N} \phi_{1n}(i_1)\phi_{2n}(i_2)\ldots\phi_{mn}(i_m)\phi_p^T(j_p, i_p) \\
=&\sum_{n=1}^{N} \phi_{1n}(i_1)\ldots\phi_{p-1,n}(i_{p-1})\phi_{p+1,n}(i_{p+1})\ldots\phi_{mn}(i_m)\sum_{i_p=1}^{D_p}\phi_{pn}(i_p)\phi_p^T(j_p, i_p) \\
=&\sum_{n=1}^{N} \phi_{1n}(i_1)\ldots\phi_{p-1,n}(i_{p-1})\phi_{p+1,n}(i_{p+1})\ldots\phi_{mn}(i_m)\mathbf{k}_{pn}(j_p)
\end{aligned}$$

Then the $(j_1, j_2, \ldots, j_m)$'th entry of $\mathcal{F}$ is

$$\begin{aligned}
\mathcal{F}(j_1, j_2, \ldots, j_m) =&(\mathcal{C} \times_1 \phi^T(X_1) \times_2 \phi^T(X_2)\ldots \times_m \phi^T(X_m))(j_1, j_2, \ldots, j_m) \\
=&\sum_{n=1}^{N} \mathbf{k}_{1n}(j_1)\mathbf{k}_{2n}(j_2)\ldots\mathbf{k}_{mn}(j_m).
\end{aligned} \tag{B.3}$$

By comparing (B.1) and (B.3), we complete the proof. □

References

Allen GI (2012) Sparse higher-order principal components analysis. In: International Conference on Artificial Intelligence and Statistics, pp 27–36

Bach FR, Jordan MI (2005) A probabilistic interpretation of canonical correlation analysis. Tech. Rep. 688, Department of Statistics, University of California, Berkeley

Belkin M, Niyogi P (2001) Laplacian eigenmaps and spectral techniques for embedding and clustering. In: Advances in Neural Information Processing Systems, pp 585–591

Benabdeslem K, Hindawi M (2014) Efficient semi-supervised feature selection: Constraint, relevance and redundancy. IEEE Transactions on Knowledge and Data Engineering 26(5):1131–1143

Blaschko MB, Lampert CH (2008) Correlational spectral clustering. In: IEEE Conference on Computer Vision and Pattern Recognition, pp 1–8

Blum A, Mitchell T (1998) Combining labeled and unlabeled data with co-training. In: Annual conference on Computational Learning Theory, pp 92–100

Chapelle O, Schölkopf B, Zien A (2006) Semi-supervised learning. MIT press Cambridge, MA

Chaudhuri K, Kakade SM, Livescu K, Sridharan K (2009) Multi-view clustering via canonical correlation analysis. In: International Conference on Machine Learning, pp 129–136

Chen N, Zhu J, Sun F, Xing EP (2012) Large-margin predictive latent subspace learning for multiview data analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence 34(12):2365–2378

Chua TS, Tang J, Hong R, Li H, Luo Z, Zheng Y (2009) NUS-WIDE: a real-world web image database from national university of singapore. In: International Conference on Image and Video Retrieval, pp 48:1–48:9

Comon P, Luciani X, De Almeida AL (2009) Tensor decompositions, alternating least squares and other tales. Journal of Chemometrics 23(7-8):393–405

De Lathauwer L, De Moor B, Vandewalle J (2000a) A multilinear singular value decomposition. SIAM journal on Matrix Analysis and Applications 21(4):1253–1278

De Lathauwer L, De Moor B, Vandewalle J (2000b) On the best rank-1 and rank-(r 1, r 2,..., rn) approximation of higher-order tensors. SIAM Journal on Matrix Analysis and Applications 21(4):1324–1342

Farquhar JDR, Hardoon D, Meng H, Shawe-taylor JS, Szedmak S (2005) Two view learning: SVM-2K, theory and practice. In: Advances in Neural Information Processing Systems, pp 355–362

Fisch D, Kalkowski E, Sick B (2014) Knowledge fusion for probabilistic generative classifiers with data mining applications. IEEE Transactions on Knowledge and Data Engineering 26(3):652–666

Foster DP, Johnson R, Zhang T (2008) Multi-view dimensionality reduction via canonical correlation analysis. Tech. Rep. TR-2009-5, TTI-Chicago

Guillaumin M, Mensink T, Verbeek J, Schmid C (2009) Tagprop: Discriminative metric learning in nearest neighbor models for image auto-annotation. In: International Conference on Computer Vision, pp 309–316

Han Y, Wu F, Tao D, Shao J, Zhuang Y, Jiang J (2012) Sparse unsupervised dimensionality reduction for multiple view data. IEEE Transactions on Circuits and Systems for Video Technology 22(10):1485–1496

Hardoon DR, Szedmak S, Shawe-Taylor J (2004) Canonical correlation analysis: An overview with application to learning methods. Neural Computation 16(12):2639–2664

Hou C, Zhang C, Wu Y, Nie F (2010) Multiple view semi-supervised dimensionality reduction. Pattern Recognition 43(3):720–730

Jenatton R, Audibert JY, Bach F (2011) Structured variable selection with sparsity-inducing norms. Journal of Machine Learning Research 12:2777–2824

Kakade SM, Foster DP (2007) Multi-view regression via canonical correlation analysis. In: Annual conference on Computational Learning Theory, pp 82–96

Kettenring JR (1971) Canonical analysis of several sets of variables. Biometrika 58(3):433–451

Kim TK, Cipolla R (2009) Canonical correlation analysis of video volume tensors for action categorization and detection. IEEE Transactions on Pattern Analysis and Machine Intelligence 31(8):1415–1428

Kroonenberg PM, De Leeuw J (1980) Principal component analysis of three-mode data by means of alternating least squares algorithms. Psychometrika 45(1):69–97

Kumar A, Rai P, Iii HD (2011) Co-regularized multi-view spectral clustering. In: Advances in Neural Information Processing Systems, pp 1413–1421

Kushmerick N (1999) Learning to remove internet advertisements. In: Proceedings of the third annual conference on Autonomous Agents, pp 175–181

Lanckriet G, Cristianini N, Bartlett P, Ghaoui L, Jordan M (2004) Learning the kernel matrix with semidefinite programming. Journal of Machine Learning Research 5:27–72

Lee SH, Choi S (2007) Two-dimensional canonical correlation analysis. IEEE Signal Processing Letters 14(10):735–738

Liu Q, Chen E, Xiong H, Ge Y, Li Z, Wu X (2014) A cocktail approach for travel package recommendation. IEEE Transactions on Knowledge and Data Engineering 26(2):278–293

Long B, Philip SY, Zhang ZM (2008) A general model for multiple view unsupervised learning. In: SDM, pp 822–833

Lowe DG (2004) Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision 60(2):91–110

Lu H (2013) Learning canonical correlations of paired tensor sets via tensor-to-vector projection. In: International Joint Conference on Artificial Intelligence, pp 1516–1522

McFee B, Lanckriet G (2011) Learning multi-modal similarity. Journal of Machine Learning Research 12:491–523

Oliva A, Torralba A (2001) Modeling the shape of the scene: A holistic representation of the spatial envelope. International Journal of Computer Vision 42(3):145–175

Scholkopf B, Smola A (2002) Learning with kernels: support vector machines, regularization, optimization, and beyond. the MIT Press

Shawe-Taylor J, Cristianini N (2004) Kernel methods for pattern analysis. Cambridge university press

Su H, Sun M, Fei-Fei L, Savarese S (2009) Learning a dense multi-view representation for detection, viewpoint classification and synthesis of object categories. In: International Conference on Computer Vision, pp 213–220

Vía J, Santamaría I, Pérez J (2007) A learning algorithm for adaptive canonical correlation analysis of several data sets. Neural Networks 20(1):139–152

Wang H (2010) Local two-dimensional canonical correlation analysis. IEEE Signal Processing Letters 17(11):921–924

White M, Zhang X, Schuurmans D, Yu Yl (2012) Convex multi-view subspace learning. In: Advances in Neural Information Processing Systems, pp 1682–1690

Wu J, Liu H, Xiong H, Cao J, Chen J (2015) K-means-based consensus clustering: A unified view. IEEE Transactions on Knowledge and Data Engineering 27(1):155–169

Xia T, Tao D, Mei T, Zhang Y (2010) Multiview spectral embedding. IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics 40(6):1438–1446

Xu B, Bu J, Chen C, Wang C, Cai D, He X (2015) Emr: A scalable graph-based ranking model for content-based image retrieval. IEEE Transactions on Knowledge and Data Engineering 27(1):102–114

Yan J, Zheng W, Zhou X, Zhao Z (2012) Sparse 2-d canonical correlation analysis via low rank matrix approximation for feature extraction. IEEE Signal Processing Letters 19(1):51–54

Yang S, Yi Z, Ye M, He X (2014) Convergence analysis of graph regularized non-negative matrix factorization. IEEE Transactions on Knowledge and Data Engineering 26(9):2151–2165

Zhu H, Xiong H, Ge Y, Chen E (2015) Discovery of ranking fraud for mobile apps. IEEE Transactions on Knowledge and Data Engineering 27(1):74–87