

Generalized Multi-view Embedding for Visual Recognition and Cross-modal Retrieval

Guanqun Cao, Alexandros Iosifidis, *Senior Member, IEEE*, Ke Chen and Moncef Gabbouj, *Fellow, IEEE*
 {guanqun.cao, ke.chen, moncef.gabbouj}@tut.fi, alexandros.iosifidis@eng.au.dk

Abstract—In this paper, the problem of multi-view embedding from different visual cues and modalities is considered. We propose a unified solution for subspace learning methods using the Rayleigh quotient, which is extensible for multiple views, supervised learning, and non-linear embeddings. Numerous methods including Canonical Correlation Analysis, Partial Least Square regression and Linear Discriminant Analysis are studied using specific intrinsic and penalty graphs within the same framework. Non-linear extensions based on kernels and (deep) neural networks are derived, achieving better performance than the linear ones. Moreover, a novel Multi-view Modular Discriminant Analysis (MvMDA) is proposed by taking the view difference into consideration. We demonstrate the effectiveness of the proposed multi-view embedding methods on visual object recognition and cross-modal image retrieval, and obtain superior results in both applications compared to related methods.

I. INTRODUCTION

People see the world differently, and objects are described from various point of views and modalities. Identifying an object can not only benefit from visual cues including color, texture and shape, but textual annotations from different observations and languages. Thanks to data enrichment from sensor technologies, the accuracy in image retrieval and recognition has been significantly improved by taking advantage of multi-view and cross-domain learning [1], [2]. Since matching the data samples across various feature spaces directly is infeasible, subspace learning approaches, which learn a common feature space from multi-view spaces, becomes an effective approach in solving the problem.

Numerous methods have been proposed in subspace learning. They can be grouped into three major categories based on the characteristics of machine learning: *two-view learning* and *multi-view learning*; *unsupervised learning* and *supervised learning*; and *linear learning* and *non-linear learning*. While traditional techniques in multivariate analysis take two inputs [3], multi-view methods have been proposed to find an optimal representation from more than two views [4], [5]. Compared to learning the feature transformation in an unsupervised manner, discriminative methods, such as Linear Discriminant Analysis

(LDA) have been extended to multi-view cases. Additionally, the transformation can also be kernel-based or learned by (deep) neural nets to exploit their non-linear properties.

Two-view learning and *multi-view learning*: One of the most popular methods in multivariate statistics is Canonical Correlation Analysis (CCA) [6]. It seeks to maximize the correlation between two sets of variables. Alternatively, its multi-view counterpart aims to obtain a common space from $V > 2$ views [4], [5], [7]. This is achieved either by scaling the cross-covariance matrices to incorporate the covariances from more than two views, or by finding the best rank-1 approximation of the data covariance tensor. A similar approach to find the common subspace is Partial Least Square Regressions [8]. It maximizes the cross-covariance from two views by regressing the data samples to the common space. Besides transformation and regression, Multi-view Fisher Discriminant Analysis (MFDA) [9] learns the transformation minimizing the difference between data samples of predicted labels. The Dropout regularization was introduced for the multi-view linear discriminant analysis in [10].

Unsupervised learning and *supervised learning*: In contrast to unsupervised transformations, including CCA and PLS, LDA [11], [12] exploits the class labels effectively by maximizing the between-class scatter while minimizing the within-class scatter simultaneously. CCA has been successfully combined with LDA to find a discriminative subspace in [13], [14], [15]. Coupled Spectral Regression (CSR) [16] projects two different inputs to the low-dimensional embedding of labels by PLS regressions. Consistent with the original LDA, a Multi-view Discriminant Analysis (MvDA) [17] finds a discriminant representation over V views. The between-class scatter is maximized regardless of the difference between inter-view and intra-view covariances, while the within-class scatter is minimized in the mean time. Generalized Multi-view Analysis (GMA) [18] was proposed to maximize the intra-view discriminant information. Recently, a semi-supervised alternative [19] was also proposed for multi-view learning, which adopts a non-negative matrix factorization method for view mapping and a robust sparse regression model for clustering the labeled samples. Moreover, a multi-view information bottleneck method [20] was proposed to retain its discrimination and robustness for multi-view learning.

Linear and *non-linear learning*: Many problems are not linearly separable and thereby kernel-based methods and learning representation by (deep) neural nets are introduced. By mapping the features to the high dimensional feature space using the kernel trick [21], kernel CCA [22] adopts a pre-

The authors are with the Laboratory of Signal Processing, Tampere University of Technology, Finland. A. Iosifidis is also with the Dept. of Engineering, Electrical and Computer Engineering, Aarhus University, DK-8200, Aarhus N, Denmark., Denmark.

This work was supported by the NSF-TEKES Center for Visual and Decision Informatics (CVDI), sponsored by Tieto Oy Finland. A. Iosifidis and K. Chen were supported from the Academy of Finland Postdoctoral Research Fellowships (No. 295854 and 298700, respectively).

defined kernel and limits its application on small datasets. Many linear multi-view methods subsequently made their kernel extension [23], [15], [24]. Kernel approximation [5] was adopted later to work on big data. Deep CCA [25] was proposed using neural nets to learn adaptive non-linear representations from two views, and uses the weights in the last layers to find the maximum correlation. A similar idea has been exploited on LDA [26]. PCANet [27] was introduced to adopt a cascade of linear transformation, followed by binary hashing and block histograms.

We make several contributions in this paper: First, we propose a unified multi-view subspace learning method for CCA, PLS and LDA techniques using the graph embedding framework [11]. We design both intrinsic and penalty graphs to characterize the intra-view and inter-view information, respectively. The intra-view and inter-view covariance matrices are scaled up to incorporate more than two views for numerous techniques by exploiting their specific intrinsic and penalty graphs. In our proposed Multi-view Modular Discriminant Analysis (MvMDA), the two graphs also characterize the within-class compactness and between-class separability. Based on the aforementioned characteristics of subspace learning algorithms, we propose a generalized objective function for multi-view subspace learning using Rayleigh quotient. This unified multi-view embedding approach can be solved as a generalized eigenvalue problem.

Second, we introduce a Multi-view Modular Discriminant Analysis (MvMDA) method by exploiting the distances between centers representing classes of different views. This is of particular interest since the resulting scatter encodes cross-view information, which empirically is shown to provide superior results. Third, we also extend the unified framework to the non-linear cases with kernels and (deep) neural networks. Kernel-based multi-view learning method is derived with an implicit kernel mapping. For larger datasets, we use the explicit kernel mapping [28] to approximate the kernel matrices. We also derive the formulation of stochastic gradient descent (SGD) for optimizing the objective function in the neural nets.

Last but not least, we demonstrate the effectiveness of the proposed embedding methods on visual object recognition and cross-modal image retrieval. Specifically, zero-shot recognition is evaluated by discovering novel object categories based on the underlying intermediate representation [29], [30], [31]. Its performance is heavily dependent on the representation in the latent space shared by visual and semantic cues. We integrate observations from *attributes* as a middle-level semantic property for the joint learning. Superior recognition results are achieved by exploiting the latent feature space with non-linear solutions learned from the multi-view representations. We also employ the proposed multi-view subspace learning methods for cross-modal image retrieval [1], [32], [?], [33]. This type of methods differs from the co-training methods for image classification [34] and web image reranking [35], [36]. In the experiments, we show promising retrieval results performed by embedding more modalities into the common feature space, and find that even conventional content-based image retrieval can be improved.



Fig. 1: Visualization of test images from the AwA dataset grouped by the features in the subspace. We highlight one of the representative classes “leopard” bounded in orange to show images of the same animal categories are positioned in their neighborhoods after multi-view embedding. Note the 2-dimensional t-SNE map [37] is generated from a near circular shape.

The rest of the paper is organized as follows. Section II reviews the related work. In Section III, we show the unified formulation to generalize the subspace learning methods. It is followed by the extension to multi-view techniques and derivation in kernels and neural nets. Then, in Section IV, we present the comparative results in zero-shot object recognition and cross-modal image retrieval on three popular multimedia datasets. Finally, Section V concludes the paper.

II. RELATED WORK

In this section, we first define the common notations used throughout the paper. Then, we will briefly review the related methods for multi-view subspace learning. Moreover, recent work on non-linear methods concerning kernels and (deep) neural networks are discussed.

A. Notations

We define the data matrix $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N]$, $\mathbf{x}_i \in \mathbb{R}^D$, where N is the number of samples and D is the feature dimension. We also define $\mathbf{X}_v \in \mathbb{R}^{D_v \times N}$, $v = 1, \dots, V$ for the feature vectors of the v th view, and discard the index in the single-view case for notation simplicity. Note that the dimensionality of the various feature spaces D_v may vary across the views. The covariance matrix is a statistics commonly used in CCA and PLS. We denote $\bar{\mathbf{X}}_v = \mathbf{X}_v - \frac{1}{N} \mathbf{X}_v \mathbf{e} \mathbf{e}^\top$ as the centered data matrix. The cross-view covariance matrix between view i and j is then expressed as $\Sigma_{ij} = \frac{1}{N} \bar{\mathbf{X}}_i \bar{\mathbf{X}}_j^\top =$

$\frac{1}{N}\mathbf{X}_i\left(\mathbf{I}-\frac{1}{N}\mathbf{e}\mathbf{e}^\top\right)\mathbf{X}_j^\top$, where $\mathbf{e}\in\mathbb{R}^N$ is a vector of ones and $\mathbf{I}\in\mathbb{R}^{N\times N}$ is the identity matrix. For the supervised learning problems, the class label of the sample \mathbf{x}_i is noted as $c_i\in\{1,2,\dots,C\}$, where C is the number of classes. We define the class vector $\mathbf{e}^c\in\mathbb{R}^N$ with $e_c(i)=1$, if $c_i=c$, and $e_c(i)=0$, otherwise. $\mathbf{W}_v\in\mathbb{R}^{D_v\times d}$, $v=1,\dots,V$ is the projection matrix for each view, d is the number of dimensions in the latent space. The feature dimension D_v in the original space of each view is usually high, which makes the distribution of the samples sparse, leading to several problems including the small sample size problem [38]. Therefore we want to project the samples to the latent space.

The generic projection function is defined to project $\mathbf{X}\in\mathbb{R}^{D\times N}$ to $\mathbf{Y}\in\mathbb{R}^{d\times N}$. We define the linear projection by $\mathbf{Y}=\mathbf{W}^\top\mathbf{X}$. In kernel methods, we map the data to a Hilbert space \mathcal{F} . Let us define $\phi(\cdot)$ as the non-linear function mapping $x_i\in\mathbb{R}^D$ to \mathcal{F} , and $\Phi=[\phi(\mathbf{x}_1),\dots,\phi(\mathbf{x}_N)]$ as the data matrix in \mathcal{F} . In multi-view cases, $\Phi=[\Phi_1^\top,\dots,\Phi_V^\top]^\top$. Since the dimensionality of \mathcal{F} is arbitrary, the kernel trick [39] is exploited in order to implicitly map the data to \mathcal{F} . The Gram matrix is given by

$$\mathbf{K}_v=\kappa(\mathbf{X}_v,\mathbf{X}_v)=\Phi_v^\top\cdot\Phi_v, \quad (1)$$

where $\kappa(\cdot,\cdot)$ is the so-called kernel function. The centered Gram matrix is $\bar{\mathbf{K}}_v=\mathbf{K}_v-\frac{1}{N}\mathbf{1}\mathbf{K}_v-\frac{1}{N}\mathbf{K}_v\mathbf{1}^\top+\frac{1}{N^2}\mathbf{1}\mathbf{K}_v\mathbf{1}$, where $\mathbf{1}\in\mathbb{R}^{N\times N}$ is an all-ones matrix. In order to find the optimal projection, we can express \mathbf{W}_v of each view as a linear combination of the training samples in the kernel space based on the Representer Theorem [21], [40]. This can be expressed by using a new weight matrix \mathbf{A}_v as

$$\mathbf{W}_v=\Phi_v\mathbf{A}_v. \quad (2)$$

In the case where a neural network with M layers is considered, β_j contains the weight parameters in the j th layer, $j=1,\dots,M$. The weights $\mathbf{B}=[\beta_1,\dots,\beta_M]$ are learned by applying stochastic gradient descent (SGD), and $h(\cdot;\mathbf{B})$ is a non-linear mapping function which maps \mathbf{X}_v to the representation of the last hidden layer \mathbf{H}_v , i.e.

$$\mathbf{H}_v=h(\mathbf{X}_v;\mathbf{B}_v), \quad (3)$$

where \mathbf{B}_v is the weight matrix trained by applying backpropagation in the v th network.

B. Canonical Correlation Analysis (CCA)

Canonical Correlation Analysis (CCA) [6], [41] is a conventional statistical technique which finds the maximum correlation between two sets of data samples $\mathbf{X}_1\in\mathbb{R}^{D_1\times N}$ and $\mathbf{X}_2\in\mathbb{R}^{D_2\times N}$ using the linear combination $\mathbf{Y}_1=\mathbf{W}_1^\top\mathbf{X}_1$ and $\mathbf{Y}_2=\mathbf{W}_2^\top\mathbf{X}_2$. \mathbf{W}_1 and \mathbf{W}_2 are determined by optimizing:

$$\mathcal{J}=\arg\max_{\mathbf{W}_1,\mathbf{W}_2}\text{corr}(\mathbf{W}_1^\top\mathbf{X}_1,\mathbf{W}_2^\top\mathbf{X}_2) \quad (4)$$

$$=\arg\max_{\mathbf{W}_1,\mathbf{W}_2}\frac{\mathbf{W}_1^\top\boldsymbol{\Sigma}_{12}\mathbf{W}_2}{\sqrt{\mathbf{W}_1^\top\boldsymbol{\Sigma}_{11}\mathbf{W}_1}\cdot\sqrt{\mathbf{W}_2^\top\boldsymbol{\Sigma}_{22}\mathbf{W}_2}}, \quad (5)$$

where

$$\boldsymbol{\Sigma}=\begin{bmatrix}\boldsymbol{\Sigma}_{11}&\boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21}&\boldsymbol{\Sigma}_{22}\end{bmatrix}=\frac{1}{N}\begin{bmatrix}\bar{\mathbf{X}}_1\bar{\mathbf{X}}_1^\top & \bar{\mathbf{X}}_1\bar{\mathbf{X}}_2^\top \\ \bar{\mathbf{X}}_2\bar{\mathbf{X}}_1^\top & \bar{\mathbf{X}}_2\bar{\mathbf{X}}_2^\top\end{bmatrix} \quad (6)$$

C. Kernel CCA

Kernel CCA finds the maximum correlation between two views after mapping them to the kernel space [22]. This is expressed by

$$\mathcal{J}=\arg\max_{\mathbf{W}_1,\mathbf{W}_2}\text{corr}(\mathbf{W}_1^\top\Phi_1,\mathbf{W}_2^\top\Phi_2) \quad (7)$$

We use the kernel trick [39] and the Representer Theorem in (2), and derive the objective function for the kernel CCA as

$$\mathcal{J}=\arg\max_{\mathbf{A}_1,\mathbf{A}_2}\frac{\mathbf{A}_1^\top\mathbf{K}_1\mathbf{K}_2\mathbf{A}_2}{\sqrt{\mathbf{A}_1^\top\mathbf{K}_1\mathbf{K}_1\mathbf{A}_1}\cdot\sqrt{\mathbf{A}_2^\top\mathbf{K}_2\mathbf{K}_2\mathbf{A}_2}}. \quad (8)$$

D. Deep CCA

Deep CCA maximizes the correlation between a pair of views by learning non-linear representations from the input data through multiple stacked layers of neurons [25], [42]. A linear CCA layer is added on top of both networks, and the inputs to the CCA layer depend on the network outputs \mathbf{H}_1 and \mathbf{H}_2 . Similar to the non-linear case in (8), a modified objective function $\min_{\mathbf{W}_1,\mathbf{W}_2}-\frac{1}{N}\text{Tr}(\mathbf{W}_1^\top\mathbf{H}_1\mathbf{H}_2^\top\mathbf{W}_2)$ is optimized, where $\mathbf{W}_1,\mathbf{W}_2$ are the projection matrices in the CCA layer, and the correlated outputs are $\mathbf{Y}_1=\mathbf{W}_1^\top\mathbf{H}_1$ and $\mathbf{Y}_2=\mathbf{W}_2^\top\mathbf{H}_2$. A modified SGD method is developed with respect to the inputs \mathbf{H}_1 and \mathbf{H}_2 to the linear layer, which are also the outputs from the two networks. The objective function is expressed as $\text{Tr}(\mathbf{W}_1^\top\mathbf{H}_1\mathbf{H}_2^\top\mathbf{W}_2)=\text{Tr}(\mathbf{T}^\top\mathbf{T})^{\frac{1}{2}}$, which describes the correlation as the sum of the top d singular vectors of $\mathbf{T}=\boldsymbol{\Sigma}_{11}^{-1/2}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1/2}$ whose definition can be found in [3].

E. Partial Least Squares (PLS) regression

Partial Least Squares (PLS) regression [8] is another dimensionality reduction technique derived from the linear combination of the input vectors \mathbf{X}_1 together with the target information which is considered as the second view \mathbf{X}_2 . PLS maximizes the between-view covariance by solving

$$\mathcal{J}=\arg\max_{\mathbf{W}_1,\mathbf{W}_2}[\text{Tr}(\mathbf{W}_1^\top\mathbf{X}_1\mathbf{X}_2^\top\mathbf{W}_2)], \quad (9)$$

$$\text{subject to } \mathbf{W}_1^\top\mathbf{W}_1=\mathbf{I},\mathbf{W}_2^\top\mathbf{W}_2=\mathbf{I}. \quad (10)$$

The non-linear extensions of PLS are obtained in the similar manner as the ones in CCA.

F. Generalized Multi-view Analysis (GMA)

GMA [18] is a generalized framework incorporating numerous dimensionality reduction methods. It maximizes the intra-view discriminant information, but ignores the inter-view information.

$$\mathcal{J} = \arg \max_{\mathbf{W}} \left[\text{Tr} \left(\sum_i^V \sum_{i < j}^V 2\lambda_{ij} \mathbf{W}_i^\top \mathbf{X}_i \mathbf{X}_j^\top \mathbf{W}_j + \sum_{i=1}^V \mu_i \mathbf{W}_i^\top \mathbf{P}_i \mathbf{W}_i \right) \right],$$

subject to $\sum_i^V \mathbf{W}_i^\top \mathbf{Q}_i \mathbf{W}_i = \mathbf{I}$. (11)

Here both \mathbf{P} and \mathbf{Q} are the intra-view covariance matrices. \mathbf{P} is a square matrix and \mathbf{Q} is a square symmetric definite matrix. We adopt Generalized Multiview Marginal Fisher Analysis (GMMFA) in this framework. The method is also kernelizable using the Representer Theorem and kernel trick.

G. Linear Discriminant Analysis (LDA)

Linear Discriminant Analysis (LDA) [11], [43] finds the projection by maximizing the ratio of the between-class scatter to the within-class scatter. Let us define by $\boldsymbol{\mu}_c$ the mean vector of the c 'th class, formed by N_c samples, and $\boldsymbol{\mu}$ the global mean. Then, LDA optimizes the following criterion:

$$\mathcal{J} = \arg \max_{\mathbf{W}} \frac{\text{Tr}(\mathbf{W}^\top \mathbf{P} \mathbf{W})}{\text{Tr}(\mathbf{W}^\top \mathbf{Q} \mathbf{W})}, \quad (12)$$

where

$$\mathbf{P} = \sum_{c=1}^C N_c (\boldsymbol{\mu}_c - \boldsymbol{\mu})(\boldsymbol{\mu}_c - \boldsymbol{\mu})^\top = \mathbf{X} \left(\sum_{c=1}^C \frac{1}{N_c} \mathbf{e}_c \mathbf{e}_c^\top - \frac{1}{N} \mathbf{e} \mathbf{e}^\top \right) \mathbf{X}^\top, \quad (13)$$

$$\mathbf{Q} = \sum_{i=1}^N (\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})^\top = \mathbf{X} \left(\mathbf{I} - \sum_{c=1}^C \frac{1}{N_c} \mathbf{e}_c \mathbf{e}_c^\top \right) \mathbf{X}^\top. \quad (14)$$

Non-linear extensions with kernels include KDA [44] and KRDA [45].

H. Multi-view Discriminant Analysis (MvDA)

MvDA [17] is the multi-view version of LDA which maximizes the ratio of the determinant of the between-class scatter matrix to that of the within-class scatter matrix. Its objective function is

$$\mathcal{J} = \arg \max_{\mathbf{W}} \frac{\text{Tr}(\mathbf{S}_B^M)}{\text{Tr}(\mathbf{S}_W^M)}, \quad (15)$$

where the between-class scatter matrix is

$$\mathbf{S}_B^M = \sum_{i=1}^V \sum_{j=1}^V \mathbf{W}_i^\top \mathbf{X}_i \left(\sum_{c=1}^C \frac{1}{N_c} \mathbf{e}_c \mathbf{e}_c^\top - \frac{1}{N} \mathbf{e} \mathbf{e}^\top \right) \mathbf{X}_j^\top \mathbf{W}_j, \quad (16)$$

and the within-class scatter matrix is

$$\mathbf{S}_W^M = \sum_{i=1}^V \sum_{j=1}^V \mathbf{W}_i^\top \mathbf{X}_i \left(\mathbf{I} - \sum_{c=1}^C \frac{1}{N_c} \mathbf{e}_c \mathbf{e}_c^\top \right) \mathbf{X}_j^\top \mathbf{W}_j. \quad (17)$$

\mathbf{W} contains the eigenvectors of the matrix $\mathbf{S} = \mathbf{S}_W^{M-1} \mathbf{S}_B^M$ corresponding to the leading d eigenvalues λ_i .

III. GENERALIZED MULTI-VIEW EMBEDDING

Here we propose a generalized expression of objective function for multi-view subspace learning. The generalized optimization problem is given by:

$$\mathcal{J} = \arg \max_{\mathbf{W}} \frac{\text{Tr}(\mathbf{W}^\top \mathbf{P} \mathbf{W})}{\text{Tr}(\mathbf{W}^\top \mathbf{Q} \mathbf{W})} \quad (18)$$

where \mathbf{P} and \mathbf{Q} are the matrices describing the inter-view and intra-view covariances, respectively. The above equation has the form of the Rayleigh quotient. Therefore, all subspace learning methods that maximize the criterion can be reduced to a generalized eigenvalue problem:

$$\mathbf{P} \mathbf{W} = \boldsymbol{\rho} \mathbf{Q} \mathbf{W}, \quad (19)$$

and the solution is given in (20) below:

$$\mathbf{W} = \begin{pmatrix} \mathbf{W}_1 \\ \vdots \\ \mathbf{W}_V \end{pmatrix} \text{ and } \boldsymbol{\rho} = \sum_{i=1}^d \lambda_i \quad (20)$$

are the generalized eigenvector and the sum of the top d generalized eigenvalues λ_i , respectively. \mathbf{W} contains the projection matrices of all views, and $\boldsymbol{\rho}$ is the value of Rayleigh quotient. We address the Rayleigh quotient as the uniform objective function, reaching out to all subspace learning methods in the paper. The non-linear multi-view embeddings can be achieved by kernel mappings, or (deep) neural networks optimized by SGD. Suppose we have a linear projection $\mathbf{Y} = \mathbf{W}^\top \mathbf{X}$, \mathbf{S}_{vij} is a similarity weight matrix which encodes the intra-view properties to be minimized, and \mathbf{S}'_{vij} is a penalty weight expressing the inter-view properties to be maximized. Then based on [11], [46], we can express the objective function as follows

$$\mathcal{J} = \arg \max_{\mathbf{W}^\top \mathbf{W} = \mathbf{I}} \frac{\sum_{v=0}^V \sum_{i=0}^N \sum_{j=0}^N \mathbf{S}'_{vij} \|\mathbf{W}_v^\top \mathbf{X}_{vi} - \mathbf{W}_v^\top \mathbf{X}_{vj}\|^2}{\sum_{v=0}^V \sum_{i=0}^N \sum_{j=0}^N \mathbf{S}_{vij} \|\mathbf{W}_v^\top \mathbf{X}_{vi} - \mathbf{W}_v^\top \mathbf{X}_{vj}\|^2} \quad (21)$$

$$= \arg \max_{\mathbf{W}^\top \mathbf{W} = \mathbf{I}} \frac{\text{Tr}(\mathbf{W}^\top \mathbf{X} \mathbf{L}' \mathbf{X}^\top \mathbf{W})}{\text{Tr}(\mathbf{W}^\top \mathbf{X} \mathbf{L} \mathbf{X}^\top \mathbf{W})}. \quad (22)$$

In the kernel case, we also have

$$\mathcal{J} = \arg \max_{\mathbf{A}^\top \mathbf{K} \mathbf{A} = \mathbf{I}} \frac{\text{Tr}(\mathbf{A}^\top \mathbf{K} \mathbf{L}' \mathbf{K} \mathbf{A})}{\text{Tr}(\mathbf{A}^\top \mathbf{K} \mathbf{L} \mathbf{K} \mathbf{A})}. \quad (23)$$

In the above, we define the diagonal matrix of each view pair as \mathbf{D}_{uv} whose i -th element is $[\mathbf{D}_{uv}]_{ii} = \sum_j [\mathbf{S}_{uv}]_{ij}$, and the total graph Laplacian matrix as $\mathbf{L} = \mathbf{D} - \mathbf{S}$. Similarly, we have \mathbf{D}' , \mathbf{S}' , \mathbf{L}' in the penalty graph.

For the non-linear mapping by neural networks, we deploy a linear embedding layer on top of the networks. This scheme is illustrated in Fig. 2. Since we have more than two input views, we train multiple neural networks whose outputs are connected to the linear layer and the objective is the same as in the linear case. By backpropagating the error of the weight matrix, we optimize the Rayleigh quotient criterion with respect to the non-linear feature representation from each view in the last hidden layer of the networks. The projection is found in the same way as in the linear case, and we will address the SGD formulation for the specific algorithms in the next section.

Fig. 3 illustrates the proposed framework graphically. We can extract different types of low-level features from images, texts, and intermediate representations. The multi-modal feature vectors are passed through linear or non-linear projec-

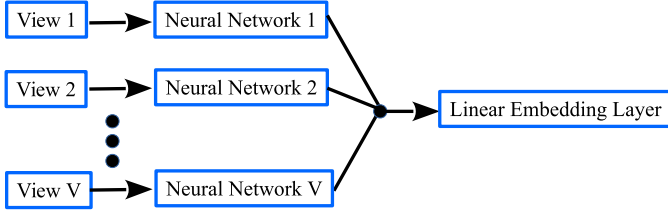


Fig. 2: An illustration of Multi-view (Deep) Embedding Neural Networks.

tions to the latent space. The projected features characterize the properties of the intra-view compactness and inter-view separability based on the proposed criterion. We show the scaled inter-view and intra-view matrices for each multi-view algorithm in the next section. Then, the projection matrices are presented with respect to their own intrinsic and penalty graph matrices and the optimization methods.

A. Scaling up the inter-view and intra-view covariance matrices

The idea behind multi-view CCA (MvCCA) is to maximize the correlation between all pairs of views. Its objective can be rephrased as maximizing the inter-view covariance while minimizing the intra-view covariance in the latent space. Therefore, we consider inter-view covariance matrices between different view representations in \mathbf{P} and the covariance matrices of each view in \mathbf{Q} . Multi-view PLS (MvPLS) maximizes the inter-view covariance directly. Since we also embed the target information for the subspace learning, the proposed MvPLS differs from MvCCA only in the intra-view minimization. Taking the class discrimination into consideration, the novel multi-view modular discriminant analysis (MvMDA) extends to separate the data of different classes between views while making the intra-class data compact. We illustrate the structure of \mathbf{P} and \mathbf{Q} for each method in Table I.

TABLE I: The matrices \mathbf{P} and \mathbf{Q} for the proposed multi-view CCA, PLS and MvMDA.

	\mathbf{P}	\mathbf{Q}
MvCCA	$\begin{bmatrix} \mathbf{0} & \Sigma_{12} & \cdots & \Sigma_{1V} \\ \Sigma_{21} & \mathbf{0} & \cdots & \Sigma_{2V} \\ \vdots & \vdots & \ddots & \vdots \\ \Sigma_{V1} & \Sigma_{V2} & \cdots & \mathbf{0} \end{bmatrix}$	$\begin{bmatrix} \Sigma_{11} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \Sigma_{22} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \Sigma_{VV} \end{bmatrix}$
MvPLS	$\begin{bmatrix} \mathbf{0} & \Sigma_{12} & \cdots & \Sigma_{1V} \\ \Sigma_{21} & \mathbf{0} & \cdots & \Sigma_{2V} \\ \vdots & \vdots & \ddots & \vdots \\ \Sigma_{V1} & \Sigma_{V2} & \cdots & \mathbf{0} \end{bmatrix}$	$\begin{bmatrix} \mathbf{I} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{I} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{I} \end{bmatrix}$
MvMDA	$\begin{bmatrix} \mathbf{P}_{11} & \mathbf{P}_{12} & \cdots & \mathbf{P}_{1V} \\ \mathbf{P}_{21} & \mathbf{P}_{22} & \cdots & \mathbf{P}_{2V} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{P}_{V1} & \mathbf{P}_{V2} & \cdots & \mathbf{P}_{VV} \end{bmatrix}$	$\begin{bmatrix} \mathbf{Q}_{11} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{Q}_{22} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{Q}_{VV} \end{bmatrix}$

B. Linear subspace learning

When the subspace projection is linear, we can obtain the latent feature vectors from each view as

$$\mathbf{Y}_v = \mathbf{W}_v^\top \mathbf{X}_v, \quad (24)$$

and the projection matrix is derived directly by solving the generalized eigenvalue problem in (19). As shown in Table I, multi-view CCA has the total covariance matrix $\Sigma = \mathbf{P} + \mathbf{Q}$, and we derive its projection matrix by fulfilling the criterion below

$$\mathcal{J} = \arg \max_{\mathbf{W}_v, v=1, \dots, V} \frac{\text{Tr} \left(\sum_{i=1}^V \sum_{\substack{j=1 \\ j \neq i}}^V \mathbf{W}_i^\top \mathbf{X}_i \mathbf{L} \mathbf{X}_j^\top \mathbf{W}_j \right)}{\text{Tr} \left(\sum_{i=1}^V \mathbf{W}_i^\top \mathbf{X}_i \mathbf{L} \mathbf{X}_i^\top \mathbf{W}_i \right)}, \quad (25)$$

where the Laplacian matrix $\mathbf{L} = \mathbf{I} - \frac{1}{N} \mathbf{e} \mathbf{e}^\top$.

Multi-view PLS has the same Laplacian matrix as the one in Multi-view CCA. We only optimize the Rayleigh quotient by maximizing the cross-covariance matrices between different views as

$$\mathcal{J} = \arg \max_{\mathbf{W}^\top \mathbf{W} = \mathbf{I}} \text{Tr} \left(\sum_{i=1}^V \sum_{\substack{j=1 \\ j \neq i}}^V \mathbf{W}_i^\top \mathbf{X}_i \mathbf{L} \mathbf{X}_j^\top \mathbf{W}_j \right), \quad (26)$$

whose solution is the projection matrix.

We propose two ways to determine the projection matrix in multi-view LDA. The first approach is the multi-view extension of the standard LDA, and its between-class scatter \mathbf{S}_B maximizes the distance between the class means from all views:

$$\begin{aligned} \mathbf{S}_B &= \sum_{i=1}^V \sum_{j=1}^V \sum_{\substack{p=1 \\ p \neq q}}^C \sum_{q=1}^C (\mathbf{m}_p^i - \mathbf{m}_q^j)(\mathbf{m}_p^i - \mathbf{m}_q^j)^\top \\ &= \sum_{i=1}^V \sum_{j=1}^V \mathbf{W}_i^\top \mathbf{X}_i \mathbf{L}_B \mathbf{X}_j^\top \mathbf{W}_j, \end{aligned} \quad (27)$$

where the between-class Laplacian matrix is

$$\mathbf{L}_B = \begin{cases} 2 \sum_{p=1}^C \sum_{\substack{q=1 \\ p \neq q}}^C \left(\frac{V}{N_p^2} \mathbf{e}_p \mathbf{e}_p^\top - \frac{1}{N_p N_q} \mathbf{e}_p \mathbf{e}_q^\top \right) & \text{if } i = j, \\ -2 \sum_{p=1}^C \sum_{\substack{q=1 \\ p \neq q}}^C \frac{1}{N_p N_q} \mathbf{e}_p \mathbf{e}_q^\top & \text{if } i \neq j. \end{cases} \quad (28)$$

\mathbf{m}_p^i denotes the mean from the i th view of the p th class in the latent space, and \mathbf{e}_p is the N -dimensional class vector, with N_p as the number of samples in the p th class. The class q is different from the class p .

Alternatively, we propose the between-class scatter matrix which maximizes the distance between different class centers across different views. Since it considers the samples from the class of the specific view origin, we call it Multi-view Modular Discriminant Analysis (MvMDA), and its formulation is

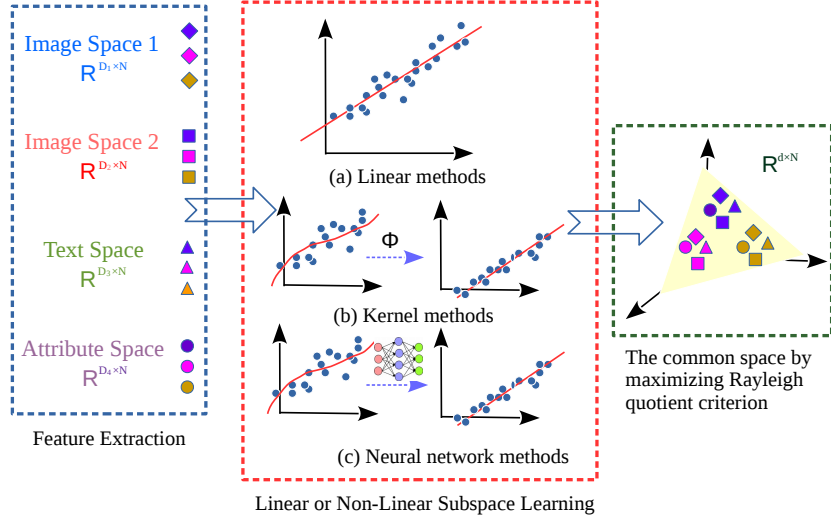


Fig. 3: Overview of the generalized multi-view embedding: Features from different modalities are extracted and either linearly or nonlinearly mapped into the common subspace by maximizing the Rayleigh quotient criterion.

$$\begin{aligned}
 \mathbf{S}'_B &= \sum_{i=1}^V \sum_{j=1}^V \sum_{p=1}^C \sum_{\substack{q=1 \\ p \neq q}}^C (\mathbf{m}_p^i - \mathbf{m}_q^i)(\mathbf{m}_p^j - \mathbf{m}_q^j)^\top \\
 &= \sum_{i=1}^V \sum_{j=1}^V \mathbf{W}_i^\top \mathbf{X}_i \mathbf{L}'_B \mathbf{X}_j^\top \mathbf{W}_j,
 \end{aligned} \quad (29)$$

$$\mathcal{J} = \arg \max_{\mathbf{W}_v, v=1, \dots, V} \frac{\text{Tr} \left(\sum_{i=1}^V \sum_{j=1}^V \mathbf{W}_i^\top \mathbf{X}_i \mathbf{L}'_B \mathbf{X}_j^\top \mathbf{W}_j \right)}{\text{Tr} \left(\sum_{i=1}^V \mathbf{W}_i^\top \mathbf{X}_i \mathbf{L}_W \mathbf{X}_i^\top \mathbf{W}_i \right)}, \quad (32)$$

where \mathbf{L}'_B is denoted as the Laplacian matrix of either \mathbf{L}_B or \mathbf{L}'_B .

and the Laplacian matrix is

$$\mathbf{L}'_B = 2 \sum_{p=1}^C \sum_{q=1}^C \left(\frac{1}{N_p^2} \mathbf{e}_p \mathbf{e}_p^\top - \frac{1}{N_p N_q} \mathbf{e}_p \mathbf{e}_q^\top \right). \quad (30)$$

The difference between the two approaches is that \mathbf{S}_B has $\frac{1}{N_c^2} (V-1) \sum_{i=1}^V \sum_{c=1}^C \mathbf{W}_i^\top \mathbf{X}_i \mathbf{e}_c \mathbf{e}_c^\top \mathbf{X}_i^\top \mathbf{W}_i$, while \mathbf{S}'_B has the term $\frac{1}{N_c^2} \sum_{i=1}^V \sum_{\substack{j=1 \\ j \neq i}}^V \sum_{c=1}^C \mathbf{W}_i^\top \mathbf{X}_i \mathbf{e}_c \mathbf{e}_c^\top \mathbf{X}_j^\top \mathbf{W}_j$ which suggests that

the first proposal only considers the maximum of the intra-view distances, while the second proposal can maximize the distance between different views. We also validate experimentally that the second proposal achieves better results. Detailed derivation of the two approaches of (27) and (29) are included in the supplementary material.

We extend the same formulation of within-class Laplacian matrix in the latent space as the single-view LDA, i.e.

$$\begin{aligned}
 \mathbf{S}_W &= \sum_{i=1}^V \mathbf{W}_i^\top \mathbf{X}_i \left(\mathbf{I} - \sum_{c=1}^C \frac{1}{N_c} \mathbf{e}_c \mathbf{e}_c^\top \right) \mathbf{X}_i^\top \mathbf{W}_i \\
 &= \sum_{i=1}^V \mathbf{W}_i^\top \mathbf{Q}_{ii} \mathbf{W}_i,
 \end{aligned} \quad (31)$$

where $\mathbf{Q}_{ii} = \mathbf{X}_i \mathbf{L}_W \mathbf{X}_i^\top$, and $\mathbf{L}_W = \mathbf{I} - \sum_{c=1}^C \frac{1}{N_c} \mathbf{e}_c \mathbf{e}_c^\top$. From (27) and (31), it is shown that the between-class and within-class scatters are equivalent to the projected inter-view and intra-view covariance, respectively. The projection matrix of the multi-view LDA is found by optimizing the following objective function

C. Kernel-based non-linear subspace learning

Exploiting the kernel trick in (1) and the Representer theorem in (2) and (24) can be expressed as follows

$$\mathbf{Y}_v = \mathbf{A}_v^\top \Phi_v^\top \Phi_v = \mathbf{A}_v^\top \mathbf{K}_v. \quad (33)$$

The criterion of kernel multi-view CCA is then,

$$\mathcal{J} = \arg \max_{\mathbf{K}_v, v=1, \dots, V} \frac{\text{Tr} \left(\sum_{i=1}^V \sum_{\substack{j=1 \\ j \neq i}}^V \mathbf{A}_i^\top \mathbf{K}_i \mathbf{L} \mathbf{K}_j \mathbf{A}_j \right)}{\text{Tr} \left(\sum_{i=1}^V \mathbf{A}_i^\top \mathbf{K}_i \mathbf{L} \mathbf{K}_i \mathbf{A}_i \right)}. \quad (34)$$

It can be easily shown that the solution for \mathbf{A}_v is the same as (19).

Kernel multi-view PLS maximizes the covariance between pairs of feature vectors in the kernel space and therefore the objective function is

$$\mathcal{J} = \arg \max_{\mathbf{K}_v, v=1, \dots, V} \text{Tr} \left(\sum_{i=1}^V \sum_{\substack{j=1 \\ j \neq i}}^V \mathbf{A}_i^\top \mathbf{K}_i \mathbf{L} \mathbf{K}_j \mathbf{A}_j \right). \quad (35)$$

The criterion for kernel multi-view discriminant analysis is

$$\mathcal{J} = \arg \max_{\mathbf{K}_v, v=1, \dots, V} \frac{\text{Tr} \left(\sum_{i=1}^V \sum_{j=1}^V \mathbf{A}_i^\top \mathbf{K}_i \mathbf{L}'_B \mathbf{K}_j \mathbf{A}_j \right)}{\text{Tr} \left(\sum_{i=1}^V \mathbf{A}_i^\top \mathbf{K}_i \mathbf{L}_W \mathbf{K}_i \mathbf{A}_i \right)} \quad (36)$$

D. Non-linear subspace learning using (deep) neural networks

Exploiting the non-linear mapping using neural networks by (3), (24) can be expressed as

$$\mathbf{Y}_v = \mathbf{W}_v^\top h(\mathbf{X}_v; \mathbf{B}_v) = \mathbf{W}_v^\top \mathbf{H}_v. \quad (37)$$

Since the network outputs are combined by a linear layer as shown in Fig. 2, the parameters \mathbf{B}_v of each network are jointly trained to reach the optimal criterion value. After the transformation by neural networks, the projection becomes the same as the multi-view linear subspace learning with respect to \mathbf{H}_v . Therefore, we need an additional optimization solved by SGD. We experimented with SGD without variance constraints, and found that we could obtain much better results with the projections constrained to have the unit variance, i.e. in Deep Multi-view CCA (DMvCCA), we have

$$\sum_{i=1}^V \mathbf{W}_i^\top \mathbf{H}_i \mathbf{L} \mathbf{H}_i^\top \mathbf{W}_i = \mathbf{I}. \quad (38)$$

Without intra-view minimization, the optimization of Deep Multi-view PLS (DMvPLS) is constrained to have unit variance $\sum_{i=1}^V \mathbf{W}_i^\top \mathbf{W}_i = \mathbf{I}$, while in Deep Multi-view Modular Discriminant Analysis (DMvMDA), we project the within-class scatter into unit, i.e.

$$\sum_{i=1}^V \mathbf{W}_i^\top \mathbf{H}_i \mathbf{L}_W \mathbf{H}_i^\top \mathbf{W}_i = \mathbf{I} \quad (39)$$

With the variance constraint, the expressions of the gradients in DMvCCA and DMvPLS are the same as

$$\begin{aligned} \frac{\partial \mathcal{J}}{\partial \mathbf{H}_i} &= \frac{\partial}{\partial \mathbf{H}_i} \text{Tr} \left(\sum_{i=1}^V \sum_{\substack{j \neq i \\ j=1}}^V \mathbf{W}_i^\top \mathbf{H}_i \mathbf{L} \mathbf{H}_j^\top \mathbf{W}_j \right) \\ &= \sum_{i=1}^V \sum_{\substack{j \neq i \\ j=1}}^V \mathbf{W}_i \mathbf{W}_j^\top \mathbf{H}_j \mathbf{L}, \end{aligned} \quad (40)$$

and the gradient of DMvMDA is computed as

$$\begin{aligned} \frac{\partial \mathcal{J}}{\partial \mathbf{H}_i} &= \frac{\partial}{\partial \mathbf{H}_i} \text{Tr} \left(\sum_{i=1}^V \sum_{j=1}^V \mathbf{W}_i^\top \mathbf{H}_i \mathbf{L}_B^* \mathbf{H}_j^\top \mathbf{W}_j \right) \\ &= \sum_{i=1}^V \sum_{j=1}^V \mathbf{W}_i \mathbf{W}_j^\top \mathbf{H}_j \mathbf{L}_B^*, \end{aligned} \quad (41)$$

Detailed derivation of (40) and (41) can be found in the supplementary material.

IV. EXPERIMENTS

In this section, we evaluate the multi-view methods on two important multimedia applications: zero-shot recognition on the Animal with Attribute (AwA) dataset, and cross-modal image retrieval on the Wikipedia and Microsoft-COCO datasets.

A. Experimental Setup

We conduct the experiments on three popular multimedia datasets. One common property in these datasets is that multi-

modal feature representations can be generated. The Animal with Attribute (AwA) dataset consists of 50 animal classes with 30,475 images in total, and 85 class-level attributes. We follow the same setup as in [31] by splitting 40 classes (24,295 images) to train the categorical model while the rest 10 classes with 6,180 images for testing. Sample images from the test set are shown in Fig. 1. Each animal class contains more than one positive attribute, and the attributes are shared across classes which enables zero-shot recognition. The detailed class labels and attributes are provided in [31].

Wikipedia is a cross-modal dataset collected from the ‘‘Wikipedia featured articles’’ [1]. The dataset is organized in 10 categories and consists of 2,866 documents. Each document is a short paragraph with a median text length of 200 words, and is associated with a single image. We follow the train/test split in [1] who use 2,173 training and 693 test pairs of images and documents.

The third dataset we use is the Microsoft COCO 2014 Dataset [47] (abbreviated as COCO in latter paragraphs). We collect the images belonging to at least one fine-grained category, which amounts to 82,081 training images, and 40,137 validation images. More than 5 human-annotated different captions are associated to each image. We follow the same definition in [47] to use 12 super classes as the class labels, and 91 fine-grained categories as the attributes. The class names and attributes are presented in Table II. The classes that the images belong to are highly semantic, and the same image can have multiple class labels. Meanwhile, similar images may belong to several different classes.

TABLE II: The class labels and attributes on the COCO dataset.

Classes
outdoor, food, indoor, appliance, sports, person, animal, vehicle, furniture, accessory, electronic, kitchen
Attributes
person, bicycle, car, motorcycle, airplane, bus, train, truck, boat, traffic, light, fire, hydrant, stop, sign, parking, meter, bench, bird, cat, dog, horse, sheep, cow, elephant, bear, zebra, giraffe, backpack, umbrella, handbag, tie, suitcase, frisbee, skis, snowboard, sports, ball, kite, bat, baseball, glove, skateboard, surfboard, tennis, racket, bottle, wine, glass, cup, fork, knife, spoon, bowl, banana, apple, sandwich, orange, broccoli, carrot, hot dog, pizza, donut, cake, chair, couch, potted, plant, bed, dining, table, toilet, tv, laptop, mouse, remote, keyboard, cell phone, microwave, oven, toaster, sink, refrigerator, book, clock, vase, scissors, teddy, bear, hair, drier, toothbrush

We use the following feature representations in the experiments:

- **Image feature by CNN models:** We employ the off-the-shelf CNN models as stated in [48] and [?] on all image datasets — Visual features are extracted by adopting two powerful pre-trained models. We rescale the size of the input images to 224×224 , and generate the features from the outputs of the *fc8* layer in a VGGNet with 16 weight

layers [49] (denoted as *VGG-16* in latter sections), and the *loss3/classifier* layer from a GoogleNet [50]. Both models produce 1000-dimension feature vectors.

- **Class label encoding:** Since each image corresponds to one class label on the AwA and Wikipedia dataset, we can describe the image category using the textual feature mapped from the image feature. Specifically, we firstly train a 100-dimension skip-gram model [51] on the entire English Wikipedia articles composed of 2.9 billion words. Then we can extract a separate set of word vectors from class labels of our datasets. In order to correlate the labels with the image contents, we train a ridge regressor with 10-fold cross-validation to map the *VGG-16* image features to each dimension of the word vectors respectively. The regressor outputs are used as the class label features.
- **Attribute encoding:** We also adopt another important modality from visual attributes on the AwA and COCO datasets. On the AwA dataset, we use the 50×85 class-attribute matrix in [52], [53] which specifies attribute probabilities of each class, while on the COCO dataset, we develop a 91-bin feature vector as attributes for each image of which 1's denote the image has the fine-grained tag and 0's otherwise. Then, we train a ridge regressor between the *VGG-16* image feature and formulated attribute probabilities. The predicted probabilities associated with each image are used as the attribute feature.
- **Sentence encoding:** A vital feature of cross-modal retrieval system is that we make use of textual features directly. We can find a paragraph of text describing each image on the Wikipedia dataset, while on the COCO dataset, a similar paragraph can be developed by concatenating all captions from the annotators which are associated to each image. We generated the sentence vectors from the paragraphs by the pre-trained skip-thoughts model [54]. The model was trained over the MovieBook and BookCorpus dataset [55]. On the Wikipedia, we employ the *combined-skip* vector of 4800 dimensions, while due to the large size of COCO dataset, we only use the *uni-skip* vector of 2400 dimensions.

The Experiment protocol and performance metrics are described below:

- **Zero-shot recognition on the AwA dataset:** We follow a similar experiment pipeline as in [56], and the comparative results show the performance of the proposed multi-view embedding methods. We project the multi-view representations to the latent space. Zero-shot recognition is achieved by semi-supervised label propagation on a transductive hypergraph in the latent space. Specifically, the cross-domain knowledge learned from the common semantic space is transferred to the target space of 10 test animal classes via attributes. The prediction of target classes is undertaken on a hypergraph to better integrate different views. We replace the multi-view linear CCA for joint embedding in [56] with the generalized embedding methods. Since the same hypergraph is used, the recognition results indicate the different performance by

the multi-view methods in this paper. For the evaluation metric, we use the average classification accuracy which is also employed in [31], [56].

- **Cross-modal retrieval on the Wikipedia and COCO datasets:** We perform two tasks in cross-modal retrieval, i.e. text query for image retrieval and image query for text retrieval. Moreover, a conventional content-based image retrieval system is evaluated in Section IV-C4. We first extract the test features in their own domains. A latent space is jointly learned from the image features, intermediate feature and sentence feature in the training set. Test features are then projected to the latent space by the trained model. The semantic matching from [1] is performed by training a logistic regressor over the embedded features from all of the ground truth samples which maps the projected features of both queries and to-be-retrieved images/texts towards the class labels. The feature vectors generated from the ground truth class labels are essentially the class vectors, whose dimensionality is the number of classes. We use the class probabilities from the regressor outputs for matching between modalities.

We present the results using 11-point interpolated precision-recall (PR) curves. The Mean Average Precision (MAP) score, which is the average precision at the ranks where recall changes, can be computed based on the Precision Recall curves. The Average Precision (AP) measures the relevance between a query and retrieved items [57], and the MAP score calculates the mean AP by querying all items in the test set.

B. Parameter Settings

The dimensionality d in the latent space is a pre-defined parameter. We will evaluate the effects of different d values in the following section. In the experiment, we use $d = 50$ for linear projections on all datasets. On the Wikipedia and AwA dataset, we choose $d = 150$ for kernel mappings, and $d = 200$ for the COCO dataset. For computational efficiency on the AwA and COCO dataset, an approximated RBF kernel mapping is adopted for the non-linear mappings. We set σ in the RBF kernel as the average distance between samples from different views/modalities, which is the natural scaling factor for each dataset. In all of the experiments, the original training set is further partitioned into a 80% training split and a 20% validation split.

The topology of neural networks has more variabilities, and we chose the optimal one according to the held-out validation set. We refer to [58], [59] for a detailed discussion on topologies. On the AwA dataset, we took 3 hidden layer, each with 1,024 neurons with the *relu* activation before the 50-dimensional linear embedding layer. We only adopted the linear and kernel-based embeddings on the Wikipedia dataset in view of its small size. On the COCO dataset, we chose a single hidden layer with 1500 *relu* neurons, and the dimensionality of the final linear layer is also 1500. We experimented both with the whole batch and multiple mini-batches for SGD, and adopted a batch size of 200 which achieves a superior

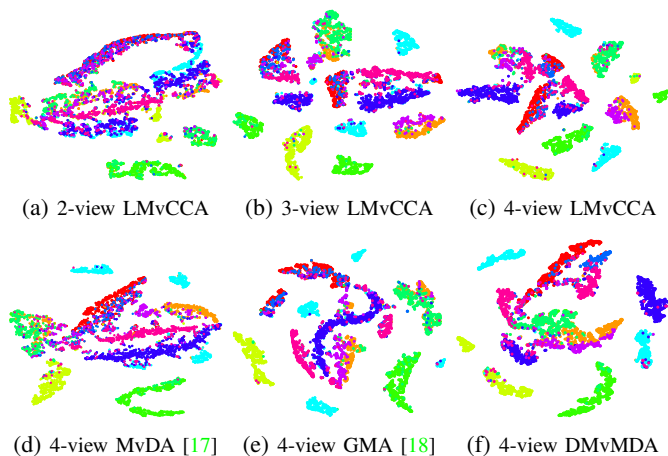


Fig. 4: The first row shows the 2-D visualization of embeddings by LMvCCA with an increasing number of views on the AwA dataset. The second row presents the embedding maps by different methods all with 4 views on the same dataset. The samples from different classes are denoted in different colors.

performance. The number of epochs is set to 50 empirically.

C. Experimental Results

The abbreviations of the numerous methods are shown in Table III.

1) *Results on zero-shot recognition:* We visualize the embedded space in Fig. 4. We use the VGG-16 feature and class label encoding for two views, and augment attribute and GoogleNet encodings as the additional views. In the first row, it is shown with the increasing number of views in MvCCA, the latent feature vector progresses from being distributed incoherently to showing more distinct groups. In the second row, we compare different methods with 4 views. It is clearly shown we obtain a set of more compact and separable features by the proposed DMvMDA.

Recognition accuracy of different methods is compared quantitatively in Table IV. The first group contains the linear projection results, the second uses the kernel methods, the third are the results by deep neural nets, and the last category includes several comparative results in the literature. The linear methods perform favorably in general while the leading recognition rates can be found in the non-linear methods using neural nets with 4 views. The kernel approximation does not provide superior results compared to linear methods due to the information loss in sampling [28]. Above all, the 4-view DMvMDA is reported to be the best method for zero-shot recognition. The results are also organized by the number of views in columns, and it is shown for all methods that we consistently obtain a better accuracy with more views. Specifically, the proposed LMvPLS achieves the highest accuracy with two input views, while the novel LMvMDA has a more discriminant representation in the latent space leading to a better recognition when more views are presented.

2) *Cross-modal retrieval results on the Wikipedia Dataset:* Due to the limited number of samples, we use PCA before performing the subspace learning. We use the VGG-16 and

sentence features for two views, and augment attribute and GoogleNet encodings as the additional modalities. It is shown that a better MAP score is obtained when enriching the latent feature with more modalities as shown in Table V. We also observe that the supervised methods perform better than the unsupervised counterparts, and non-linear projections by kernel methods are superior. KMvMDA achieves the best retrieval results with supervision and non-linearity.

We present more detailed results in the form of PR curves in Fig. 5. For image queries, KMvMDA consistently outperforms the other methods across all views, which can be explained by its utilization of class labels and kernel-based representations. For text queries, the supervised and non-linear methods also outperform their linear counterparts. KMvCCA and KMvMDA are the leading methods in this category, which shows the strength of cross-modal retrieval by making use of view difference.

3) *Cross-modal retrieval results on the COCO Dataset:* The COCO dataset is much larger than the Wikipedia dataset, and we pay more attention to the non-linear methods especially the ones using neural networks. Many images have more than one class labels, and therefore we focus on the unsupervised learning algorithms. Similar to the experiments above, the MAP scores in Table VI show that a gain of retrieval accuracy can be obtained by embedding additional modalities into the latent space. DCCA2 [25] achieves a superior performance with 2 views thanks to its non-linear projection which makes the latent feature more discriminant for retrieval. However, its formulation limits the algorithm to 2 views, and DMvCCA and DMvPLS based on the proposed framework can improve the state-of-the-art method by increasing the number of modalities. From the PR curves in Fig. 6, we compare the methods using the proposed objective function with DCCA2 which contains two views. For image queries, KapMvCCA obtains the best retrieval result with 2 views, but it is further improved by the methods using neural networks benefitted by attributes and GoogleNet features. For text queries, it also suggests more modalities and neural network-based representations contribute to the retrieval performance. The cross-modal retrieval by the 4-view DMvCCA achieves the overall highest precision score on this dataset.

4) *Content-based Image Retrieval (CBIR) Performance on the COCO dataset:* We also show the effectiveness of multi-view embedding method on the conventional CBIR task in Fig. 7. We randomly pick two image-to-text pairs as queries, to perform image-to-image retrieval using both the VGG-16 visual feature and the projected visual feature by the 4-view DCCA. We also perform text-to-image retrieval by querying the corresponding captions of the query image used in CBIR in the last column. We observe the CBIR performance can be further improved by incorporating the semantic information. In Table VII, we present the quantitative results of CBIR by the projected visual features. “RAW” in the Table shows the retrieval results by visual features directly, while the rest are the multi-view embedding results. It is shown that more modalities and non-linear projections yield a discriminant latent visual feature, which improves the retrieval performance.

TABLE III: List of Abbreviations

LMvCCA / KmVCCA / KapMvCCA / DMvCCA	Linear / Kernel / Approximate Kernel / Deep Multi-view Canonical Correlation Analysis
LMvPLS / KmVPLS / KapMvPLS / DMvPLS	Linear / Kernel / Approximate Kernel / Deep Multi-view Partial Least Square Regression
SLMvDA / SKMvDA	Standard Linear / Kernel Multi-view Discriminant Analysis using (28)
LMvMDA / KmVMDA / KapMvMDA / DMvMDA	Linear / Kernel / Approximate Kernel / Deep Multi-view Modular Discriminant Analysis using (30)
MULDA / KMuDA [15]	Multi-view Uncorrelated Linear / Kernel Discriminant Analysis
MvDA [17]	Multi-view Discriminant Analysis
GMA [18]	Generalized Multi-view Analysis
DCCA2 [25]	Deep Canonical Correlation Analysis

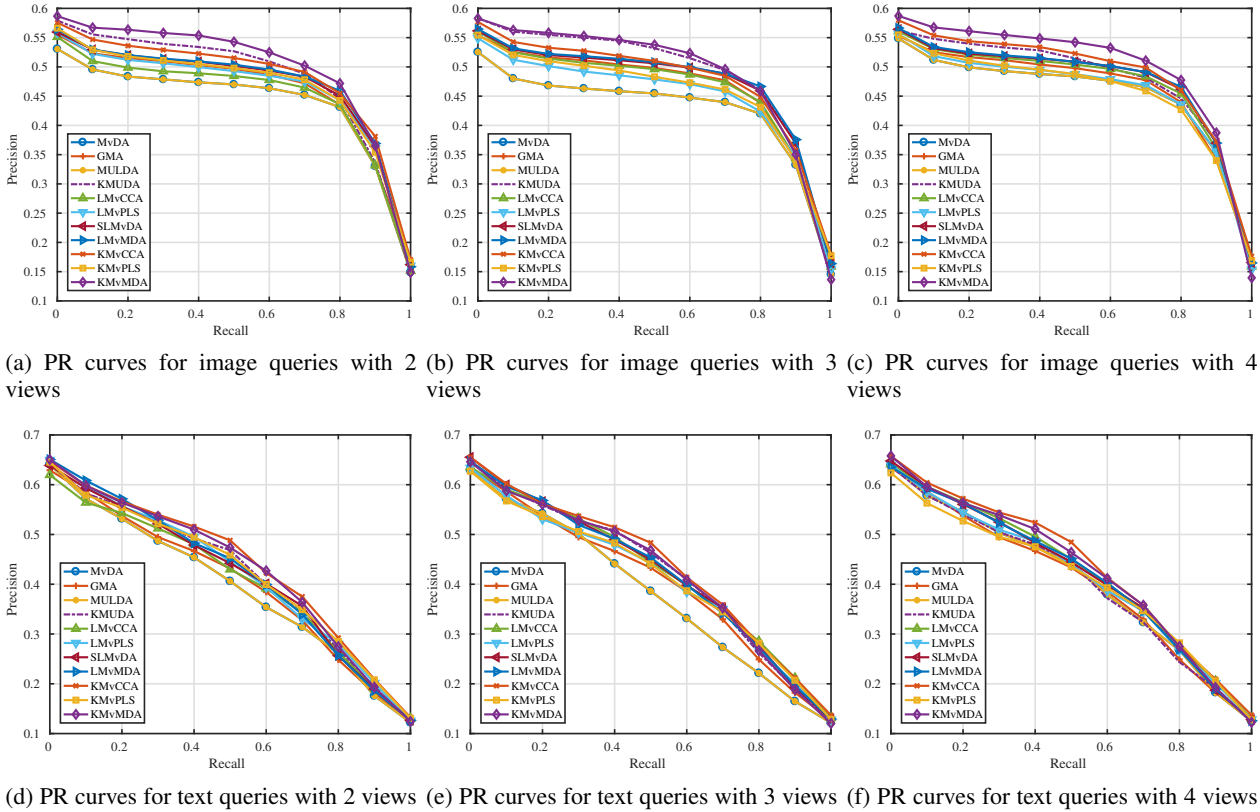


Fig. 5: PR curves across different number of views on the Wikipedia dataset for the Image-to-Text retrieval and the Text-to-Image retrieval.

TABLE IV: RECOGNITION ACCURACY (%) on the AWA DATASET

Method	2 views	3 views	4 views
Proposed LMvCCA	55.86	75.88	82.01
Proposed LMvPLS	58.52	73.59	77.09
Proposed LMvMDA	55.85	77.64	82.88
Proposed SLMvDA	54.58	69.02	70.56
Proposed KapMvCCA	56.41	73.40	74.76
Proposed KapMvPLS	55.58	74.40	75.05
Proposed KapMvMDA	57.19	71.64	75.63
Proposed DMvCCA	51.25	71.12	82.27
Proposed DMvPLS	43.28	68.81	74.63
Proposed DMvMDA	53.87	75.61	83.66
MvDA [17]	49.95	68.55	70.00
GMA [18]	52.12	73.49	78.46
MULDA [15]	55.46	74.13	74.88
TMV-HLP [56]	-	73.50	80.50
DCCA2 [25]	50.47	-	-

D. Parameter sensitivity analysis of dimension d in linear and kernel cases

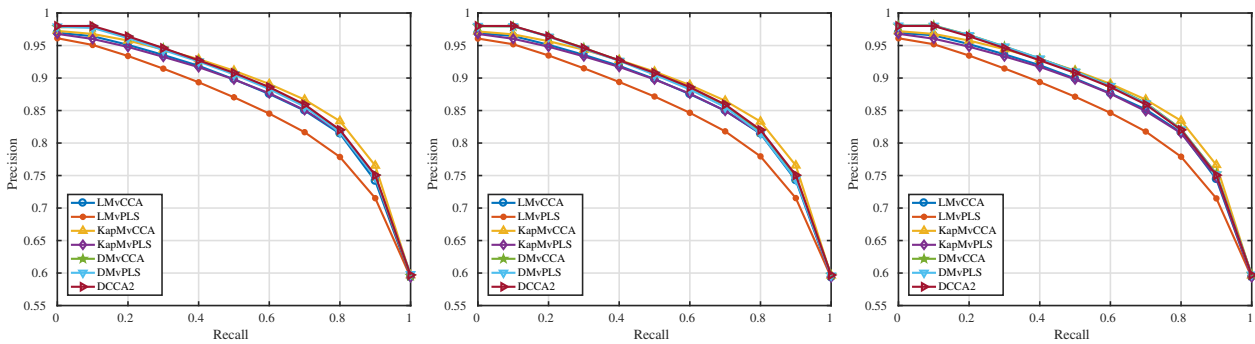
The number of dimension of the feature vectors in the latent space is determined by the top d eigenvectors in the projection matrix, and it is pre-defined in the former experiments. Therefore in this section, we investigate the effect by the variation of d shown in Fig. 8 and 9, ranging from $\{10, 20, 50, 100, 150, 200\}$. The performance on the Wikipedia dataset is reported with both text queries on images and image queries on texts. The results on different number of views are also recorded. In general, we obtain a better retrieval performance when d is between 50 and 150. It can be explained by the fact that the most informative eigenvectors are included within the range. Therefore, $d = 50$ was chosen for the multi-view linear embeddings in the experiments. Except LMvPLS and KmVPLS, we find the majority of the methods are robust to the dimensionality changes in the subspace.

TABLE V: MAP Scores (%) on the Wikipedia

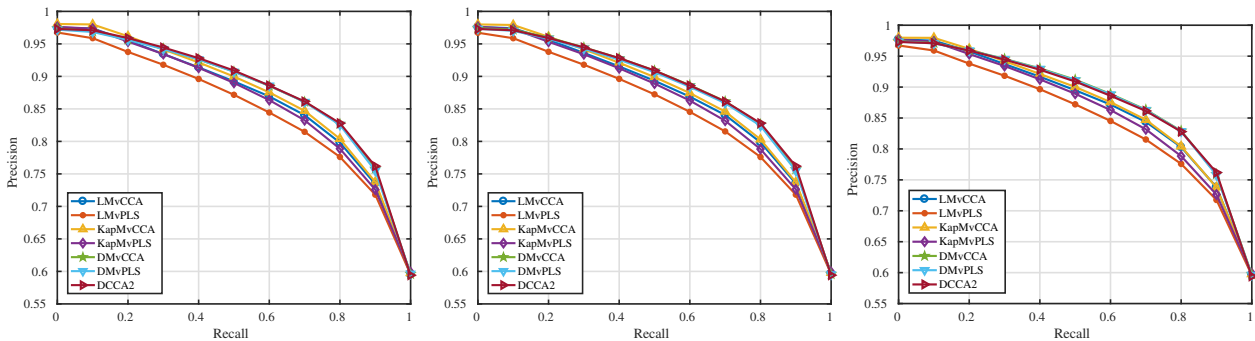
	2 views			3 views			4 views		
	img. query	txt. query	avg.	img. query	txt. query	avg.	img. query	txt. query	avg.
MvDA [17]	39.73	37.14	38.43	39.34	35.04	37.19	41.07	39.21	40.14
GMA [18]	41.91	38.55	40.23	42.26	38.66	40.46	42.26	38.67	40.47
MULDA [15]	43.04	39.87	41.46	43.45	40.68	42.07	43.79	40.32	42.06
Proposed LMvCCA	41.37	39.07	40.22	42.10	39.64	40.87	42.53	39.98	41.26
Proposed LMvPLS	42.49	40.42	41.46	41.29	39.34	40.31	41.86	39.74	40.80
Proposed SLMvDA	43.20	40.07	41.64	43.14	39.86	41.50	43.77	40.24	41.80
Proposed LMvMDA	43.38	40.32	41.85	43.74	40.46	42.10	43.90	40.23	42.07
KMUDA [15]	44.38	39.52	41.95	45.40	39.96	42.68	44.29	38.12	41.20
Proposed KMvCCA	44.78	41.83	43.30	44.06	41.41	42.73	45.13	41.66	43.40
Proposed KMvPLS	42.94	40.46	41.70	42.03	39.40	40.71	41.94	38.84	40.39
Proposed SKMvDA	45.52	38.39	41.96	44.66	38.47	41.57	42.94	39.32	41.13
Proposed KMvMDA	46.01	40.96	43.49	45.40	40.16	42.78	46.48	40.73	43.61

TABLE VI: MAP Scores (%) on the COCO dataset

	2 views			3 views			4 views		
	img. query	txt. query	avg.	img. query	txt. query	avg.	img. query	txt. query	avg.
Proposed LMvCCA	87.18	86.92	87.05	87.20	87.01	87.11	87.31	87.22	87.27
Proposed LMvPLS	84.76	85.05	84.91	84.83	85.07	84.95	84.82	85.05	84.94
Proposed KapMvCCA	88.42	87.58	88.00	88.35	87.52	87.94	88.45	87.60	88.03
Proposed KapMvPLS	87.16	86.58	86.87	87.14	86.56	86.85	87.14	86.56	86.85
Proposed DMvCCA	88.14	88.10	88.12	88.20	88.26	88.23	88.49	88.40	88.45
Proposed DMvPLS	88.01	88.03	88.02	88.06	88.03	88.05	88.45	88.34	88.40
DCCA2 [25]	88.30	88.27	88.29	-	-	-	-	-	-



(a) PR curves for image queries with 2 views (b) PR curves for image queries with 3 views (c) PR curves for image queries with 4 views



(d) PR curves for text queries with 2 views (e) PR curves for text queries with 3 views (f) PR curves for text queries with 4 views

Fig. 6: PR curves across different number of views on the COCO dataset for the Image-to-Text retrieval and the Text-to-Image retrieval. Note the curve by DCCA2 [25] is presented across all numbers of views.

V. CONCLUSION

In this paper, we proposed a generalized multi-view embedding method using the graph embedding framework. We showed multi-view CCA, PLS and LDA can be characterized

by their specific intrinsic and penalty graph matrices within the same framework. A novel discriminant analysis method named MvMDA was introduced by exploiting the distances between class centers of different views. Meanwhile, we also


Image Query 	Text Query 1. A very big building with many windows and a clock on it. 2. A very old tall building with a large clock tower sticking out of it. 3. The clock tower stands high above the city. 4. A clock that is on the side of a large building. 5. The bridge is in front of a huge building with a clock tower in the middle of it.		
Precision: 53.33%	Precision: 86.67%	Precision: 100%	
			
(a) Query by original image feature	(b) Query by projected image feature	(c) Query by text	
Image Query 	Text Query 1. An open laptop sits on a desk in front of a window. 2. An Apple laptop sitting on a wooden desk. 3. An Apple laptop sitting on a wooden desk in an office. 4. An Apple laptop on a desk in an office. 5. A desk with a laptop sitting on top of it.		
Precision: 60.00%	Precision: 86.67%	Precision: 66.67%	
			
(a) Query by original image feature	(b) Query by projected image feature	(c) Query by text	

Fig. 7: Sample retrieval results on the COCO dataset. The first row of each table presents the query image and text, and the second row shows the retrieved images by different query types. False positive results are bounded in red.

TABLE VII: MAP(%) scores of CBIR on the COCO dataset

Method	2 views	3 views	4 views
Raw		83.77	
Proposed LMvCCA	85.64	85.76	85.93
Proposed LMvPLS	84.30	84.30	84.32
Proposed KapMvCCA	85.43	85.47	85.49
Proposed KapMvPLS	84.56	84.57	84.58
Proposed DMvCCA	89.33	89.62	89.84
Proposed DMvPLS	89.50	89.34	89.79
DCCA2 [25]	89.71	-	-

studied non-linear embeddings, and found implicit and explicit kernel mappings for multi-view learning. A unified scheme for learning by neural networks was developed which combined the learned representations with a linear embedding layer. We thereby formulated the expression of stochastic gradient descent for optimizing the proposed objective function.

We validated the formulation by conducting experiments in zero-shot visual object recognition and cross-modal image retrieval. It was shown that supervised and non-linear subspace learning outperformed the unsupervised and linear methods when large amount of images and texts were available. Moreover, the recognition or retrieval performance were consis-

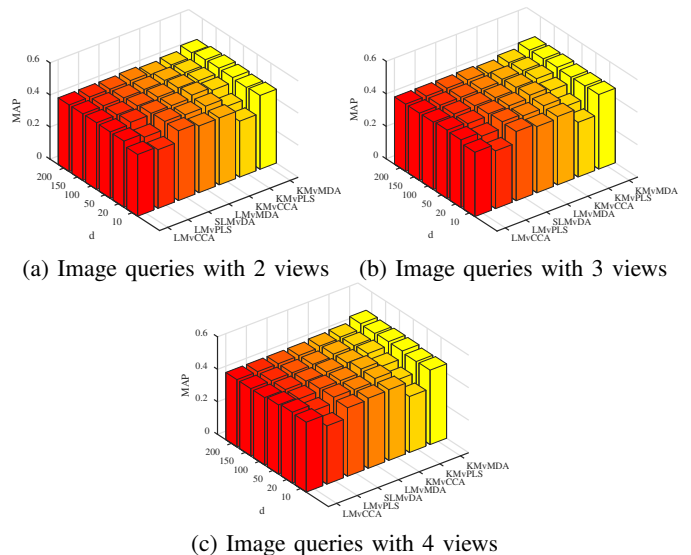


Fig. 8: Performance variation for image queries on texts of Wikipedia dataset with respect to the different dimension d .

tently improved by embedding more views/modalities into the latent feature space. We also performed the traditional CBIR experiments where the multi-view embeddings can contribute

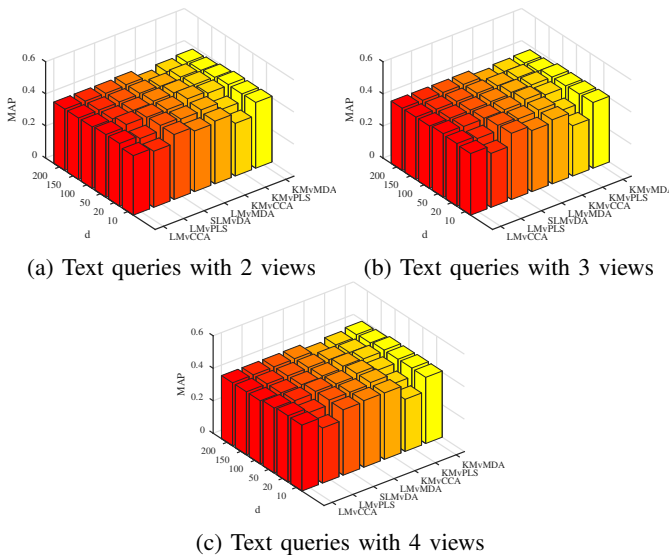


Fig. 9: Performance variation for text queries on images of Wikipedia dataset with respect to the different dimension d .

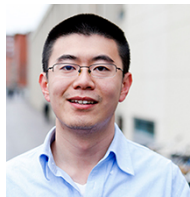
to the performance gain.

Interesting future research directions include learning from the raw data to achieve an end-to-end solution for multi-view learning. We should further reduce the computational cost for kernel methods to cope with large scale of images. In addition, learning from incomplete and unlabeled multi-view data should be studied for video analysis.

REFERENCES

- [1] J. Costa Pereira, E. Coviello, G. Doyle, N. Rasiwasia, G. R. Lanckriet, R. Levy, and N. Vasconcelos, "On the role of correlation and abstraction in cross-modal multimedia retrieval," *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)*, vol. 36, no. 3, pp. 521–535, 2014.
- [2] C. Xu, D. Tao, and C. Xu, "A survey on multi-view learning," *arXiv preprint arXiv:1304.5634*, 2013.
- [3] K. V. Mardia, J. T. Kent, and J. M. Bibby, *Multivariate analysis*. Academic press, 1980, ch. 10 Canonical Correlation Analysis, pp. 281–290.
- [4] A. A. Nielsen, "Multiset canonical correlations analysis and multispectral, truly multitemporal remote sensing data," *IEEE Transactions on Image Processing (TIP)*, vol. 11, no. 3, pp. 293–305, 2002.
- [5] Y. Gong, Q. Ke, M. Isard, and S. Lazebnik, "A multi-view embedding space for modeling internet images, tags, and their semantics," *International Journal of Computer Vision*, vol. 106, no. 2, pp. 210–233, 2014.
- [6] H. Hotelling, "Relations between two sets of variates," *Biometrika*, pp. 321–377, 1936.
- [7] 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, Nov 2015.
- [8] S. Wold, A. Ruhe, H. Wold, and W. Dunn, III, "The collinearity problem in linear regression. the partial least squares (PLS) approach to generalized inverses," *SIAM Journal on Scientific and Statistical Computing*, vol. 5, no. 3, pp. 735–743, 1984.
- [9] T. Diethe, D. R. Hardoon, and J. Shawe-Taylor, "Multiview fisher discriminant analysis," in *NIPS workshop on learning from multiple sources*, 2008.
- [10] G. Cao, M. A. Waris, A. Iosifidis, and M. Gabbouj, "Multi-modal subspace learning with dropout regularization for cross-modal recognition and retrieval," in *2016 Sixth International Conference on Image Processing Theory, Tools and Applications (IPTA)*, Dec 2016, pp. 1–6.
- [11] S. Yan, D. Xu, B. Zhang, H.-J. Zhang, Q. Yang, and S. Lin, "Graph embedding and extensions: a general framework for dimensionality reduction," *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)*, vol. 29, no. 1, pp. 40–51, 2007.
- [12] A. Iosifidis, A. Tefas, and I. Pitas, "On the optimal class representation in linear discriminant analysis," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 24, no. 9, pp. 1491–1497, Sept 2013.
- [13] T. Sun, S. Chen, J. Yang, and P. Shi, "A novel method of combined feature extraction for recognition," in *In Proceedings of IEEE International Conference on Data Mining, (ICDM)*. IEEE, 2008, pp. 1043–1048.
- [14] Y. Ma, S. Lao, E. Takikawa, and M. Kawade, "Discriminant analysis in correlation similarity measure space," in *Proceedings of the 24th international conference on Machine learning (ICML)*. ACM, 2007, pp. 577–584.
- [15] S. Sun, X. Xie, and M. Yang, "Multiview uncorrelated discriminant analysis," *IEEE Transactions on Cybernetics*, vol. 46, no. 12, pp. 3272–3284, Dec 2016.
- [16] Z. Lei and S. Z. Li, "Coupled spectral regression for matching heterogeneous faces," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2009, pp. 1123–1128.
- [17] M. Kan, S. Shan, H. Zhang, S. Lao, and X. Chen, "Multi-view discriminant analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)*, vol. 38, no. 1, pp. 188–194, Jan 2016.
- [18] A. Sharma, A. Kumar, H. Daume III, and D. W. Jacobs, "Generalized multiview analysis: A discriminative latent space," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2012, pp. 2160–2167.
- [19] J. Liu, Y. Jiang, Z. Li, Z. H. Zhou, and H. Lu, "Partially shared latent factor learning with multiview data," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 26, no. 6, pp. 1233–1246, June 2015.
- [20] C. Xu, D. Tao, and C. Xu, "Large-margin multi-view information bottleneck," *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)*, vol. 36, no. 8, pp. 1559–1572, Aug 2014.
- [21] B. Schölkopf, S. Mika, C. J. Burges, P. Knirsch, K.-R. Müller, G. Rätsch, and A. J. Smola, "Input space versus feature space in kernel-based methods," *IEEE Transactions on Neural Networks*, vol. 10, no. 5, pp. 1000–1017, 1999.
- [22] 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.
- [23] T. Sun, S. Chen, Z. Jin, and J. Yang, "Kernelized discriminative canonical correlation analysis," in *International Conference on Wavelet Analysis and Pattern Recognition*, vol. 3. IEEE, 2007, pp. 1283–1287.
- [24] A. Iosifidis and M. Gabbouj, "Scaling up class-specific kernel discriminant analysis for large-scale face verification," *IEEE Transactions on Information Forensics and Security*, vol. PP, no. 99, pp. 1–1, 2016.
- [25] G. Andrew, R. Arora, J. Bilmes, and K. Livescu, "Deep canonical correlation analysis," in *Proceedings of the 30th International Conference on Machine Learning*, 2013, pp. 1247–1255.
- [26] M. Dorfer, R. Kelz, and G. Widmer, "Deep linear discriminant analysis," *International Conference on Learning Representations (ICLR)*, 2016.
- [27] T.-H. Chan, K. Jia, S. Gao, J. Lu, Z. Zeng, and Y. Ma, "Pcnet: A simple deep learning baseline for image classification?" *IEEE Transactions on Image Processing (TIP)*, vol. 24, no. 12, pp. 5017–5032, Dec 2015.
- [28] A. Rahimi and B. Recht, "Random features for large-scale kernel machines," in *Advances in neural information processing systems*, 2007, pp. 1177–1184.
- [29] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth, "Describing objects by their attributes," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2009, pp. 1778–1785.
- [30] C. H. Lampert, H. Nickisch, and S. Harmeling, "Learning to detect unseen object classes by between-class attribute transfer," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2009, pp. 951–958.
- [31] —, "Attribute-based classification for zero-shot visual object categorization," *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)*, vol. 36, no. 3, pp. 453–465, 2014.
- [32] C. Kang, S. Xiang, S. Liao, C. Xu, and C. Pan, "Learning consistent feature representation for cross-modal multimedia retrieval," *IEEE Transactions on Multimedia*, vol. 17, no. 3, pp. 370–381, March 2015.
- [33] R. He, M. Zhang, L. Wang, Y. Ji, and Q. Yin, "Cross-modal subspace learning via pairwise constraints," *IEEE Transactions on Image Processing*, vol. 24, no. 12, pp. 5543–5556, Dec 2015.
- [34] J. Yu, Y. Rui, Y. Y. Tang, and D. Tao, "High-order distance-based multi-view stochastic learning in image classification," *IEEE Transactions on Cybernetics*, vol. 44, no. 12, pp. 2431–2442, Dec 2014.
- [35] J. Yu, Y. Rui, and D. Tao, "Click prediction for web image reranking using multimodal sparse coding," *IEEE Transactions on Image Processing*, vol. 23, no. 5, pp. 2019–2032, May 2014.

- [36] J. Yu, X. Yang, F. Gao, and D. Tao, "Deep multimodal distance metric learning using click constraints for image ranking," *IEEE Transactions on Cybernetics*, 2017, to be published.
- [37] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of Machine Learning Research*, vol. 9, no. 2579-2605, p. 85, 2008.
- [38] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*. Springer, 2013, ch. 18. High-Dimensional Problems: $p \gg N$, pp. 649–694.
- [39] K. R. Muller, S. Mika, G. Ratsch, K. Tsuda, and B. Scholkopf, "An introduction to kernel-based learning algorithms," *IEEE Transactions on Neural Networks*, vol. 12, no. 2, pp. 181–201, Mar 2001.
- [40] B. Schölkopf, R. Herbrich, and A. J. Smola, "A generalized representer theorem," in *Proceedings of Annual Conference of Computational Learning Theory*. Springer, Heidelberg, Germany, 2001, pp. 416–426.
- [41] M. Borga, "Canonical correlation: a tutorial," <http://people.imt.liu.se/~magnus/cca/tutorial/tutorial.pdf>, 2001.
- [42] W. Wang, R. Arora, K. Livescu, and J. A. Bilmes, "Unsupervised learning of acoustic features via deep canonical correlation analysis," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 4590–4594.
- [43] A. M. Martínez and A. C. Kak, "PCA versus LDA," *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)*, vol. 23, no. 2, pp. 228–233, 2001.
- [44] G. Baudat and F. Anouar, "Generalized discriminant analysis using a kernel approach," *Neural computation*, vol. 12, no. 10, pp. 2385–2404, 2000.
- [45] A. Iosifidis, A. Tefas, and I. Pitas, "Kernel reference discriminant analysis," *Pattern Recognition Letters*, vol. 49, pp. 85–91, 2014.
- [46] Y. Jia, F. Nie, and C. Zhang, "Trace ratio problem revisited," *IEEE Transactions on Neural Networks*, vol. 20, no. 4, pp. 729–735, 2009.
- [47] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *European Conference on Computer Vision*. Springer, 2014, pp. 740–755. [Online]. Available: <http://mscoco.org/home/>
- [48] A. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "CNN features off-the-shelf: an astounding baseline for recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2014, pp. 806–813.
- [49] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *International Conference on Learning Representations (ICLR)*, 2015.
- [50] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 1–9.
- [51] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in neural information processing systems (NIPS)*, 2013, pp. 3111–3119.
- [52] C. Kemp, J. B. Tenenbaum, T. L. Griffiths, T. Yamada, and N. Ueda, "Learning systems of concepts with an infinite relational model," in *AAAI*, vol. 3, 2006, p. 5.
- [53] D. N. Osherson, J. Stern, O. Wilkie, M. Stob, and E. E. Smith, "Default probability," *Cognitive Science*, vol. 15, no. 2, pp. 251–269, 1991.
- [54] R. Kiros, Y. Zhu, R. R. Salakhutdinov, R. Zemel, R. Urtasun, A. Torralba, and S. Fidler, "Skip-thought vectors," in *Advances in Neural Information Processing Systems*, 2015, pp. 3276–3284.
- [55] Y. Zhu, R. Kiros, R. Zemel, R. Salakhutdinov, R. Urtasun, A. Torralba, and S. Fidler, "Aligning books and movies: Towards story-like visual explanations by watching movies and reading books," *Cai2007*, 2015.
- [56] Y. Fu, T. Hospedales, T. Xiang, and S. Gong, "Transductive multi-view zero-shot learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)*, vol. 37, no. 11, pp. 2332–2345, Nov 2015.
- [57] C. D. Manning, P. Raghavan, H. Schütze *et al.*, *Introduction to Information Retrieval*. Cambridge university press Cambridge, 2008, vol. 1, no. 1, ch. 8. Evaluation in information retrieval, pp. 188–210.
- [58] X. Yao, "Evolving artificial neural networks," *Proceedings of the IEEE*, vol. 87, no. 9, pp. 1423–1447, 1999.
- [59] S. Kiranyaz, T. Ince, A. Yildirim, and M. Gabbouj, "Evolutionary artificial neural networks by multi-dimensional particle swarm optimization," *Neural Networks*, vol. 22, no. 10, pp. 1448 – 1462, 2009. [Online]. Available: <http://www.sciencedirect.com/science/article/B6T08-4WGF117-1/2/600b51dc41c51f5fc4c9427b352c7e6a>



Guanqun Cao received the double B.Eng. degree in Electronic and Information/Computer Engineering from Huazhong University of Science and Technology, China and University of Birmingham, UK. He also received the M.Sc degree from the joint Erasmus Mundus programme in Color Informatics and Media Technology. He is currently a PhD student at the Multimedia Research Group, Tampere University of Technology, Finland. His research interests include multimedia retrieval and machine learning with a focus on multi-view data analysis.



Alexandros Iosifidis (SM'16) received the Diploma and the M.Eng. degrees in Electrical & Computer Engineering from Democritus University of Thrace in 2008 and 2010, respectively. He received the Ph.D. in Informatics from Aristotle University of Thessaloniki in 2014. Currently, he is an Assistant Professor in the Department of Engineering, Electrical and Computer Engineering, Aarhus University, Denmark. Before that, he was a postdoctoral researcher at the Multimedia Research Group of the Department of Signal Processing in Tampere

University of Technology, holding an Academy Postdoctoral Research Fellow position.

Dr. Iosifidis is a Senior member of IEEE. He has (co-)authored more than 100 journal and conference papers in his areas of expertise. His research interests are in the areas of pattern recognition and machine learning, with applications mainly in images/videos and time series.



Ke Chen was born in Wuxi, China in 1985. He received Ph.D major in computer vision under the supervision of Prof. Shaogang Gong and Prof. Tao Xiang at School of Electronic Engineering and Computer Science, Queen Mary, University of London, UK. He received his B.E. major in automation and M.E. major in software engineering supervised by Prof. Yunong Zhang at Sun Yat-sen University, China in 2007 and 2009, respectively.

Dr. Chen is currently the Academy of Finland post-doctoral research fellow at the Department of Signal Processing, Tampere University of Technology. His research interests include computer vision, pattern recognition, neural dynamic modelling, and robotic inverse kinematics. He has published more than forty peer-reviewed conference and journal papers in computer vision, neural networks and robotics.



Moncef Gabbouj (F'11) received his BS degree in electrical engineering in 1985 from Oklahoma State University, and his MS and PhD degrees in electrical engineering from Purdue University, in 1986 and 1989, respectively. Dr. Gabbouj is a Professor of Signal Processing at the Department of Signal Processing, Tampere University of Technology, Tampere, Finland. He was Academy of Finland Professor during 2011–2015. His research interests include multimedia content-based analysis, indexing and retrieval, machine learning, nonlinear signal and

image processing and analysis, voice conversion, and video processing and coding. Dr. Gabbouj is a Fellow of the IEEE and member of the Academia Europaea and the Finnish Academy of Science and Letters. He is the past Chairman of the IEEE CAS TC on DSP and committee member of the IEEE Fourier Award for Signal Processing. He served as associate editor and guest editor of many IEEE, and international journals and Distinguished Lecturer for the IEEE CASS. He organized several tutorials and special sessions for major IEEE conferences and EUSIPCO. Dr. Gabbouj guided 46 PhD students and published 700 papers.