IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received September 25, 2017, accepted October 23, 2017, date of publication October 31, 2017,
date of current version November 28, 2017.

Digital Object Identifier 10.1109/ACCESS.2017.2767818

Retargeted Multi-View Feature Learning With Separate and Shared Subspace Uncovering

GUO-SEN XIE[1,2], XIAO-BO JIN[3], ZHENG ZHANG[4], ZHONGHUA LIU[1,2], XIAOWEI XUE[5], AND JIEXIN PU[1,2]

[1]Department of Information Engineering College, Henan University of Science and Technology, Luoyang 471023, China
[2]Henan Joint International Research Laboratory of Image Processing and Intelligent Detection, Henan University of Science and Technology, Luoyang 471023, China
[3]School of Information Science and Engineering, Henan University of Technology, Zhengzhou 450001, China
[4]Bio-Computing Research Center, Shenzhen Graduate School, Harbin Institute of Technology, Shenzhen 518055, China
[5]College of Computer Science, Zhejiang University, Hangzhou 310027, China

Corresponding author: Guo-Sen Xie (gsxiehm@gmail.com)

This work was supported in part by the National Natural Science Foundation of China under Grant 61103138 and Grant 61702163, and in part by the Henan International Cooperation Project under Grant 152102410036.

ABSTRACT Multi-view feature learning aims at improving the performances of learning tasks, by fusing various kinds of features (views), such as heterogeneous features and/or homogeneous features. Current leading multi-view feature learning approaches usually learn features in each view separately while not uncovering shared information from multiple views. In this paper, we propose a multi-view feature learning framework, which can simultaneously learn separate subspace for each view and shared subspace for all the views, respectively; specifically, the separate subspace for each view can preserve the particular information within this view, meanwhile, the shared subspace can capture feature correlation among multiple views. Both the particularity and communality are essential for classification. Furthermore, we relax the labels of training samples within the concatenated subspaces, thus resulting in the retargeted least square regression (LSR) classifier. The transformation matrices tailored for each subspace within the corresponding view and the label relaxed LSR classifier are jointly learned in a unified framework, based on an efficient alternative optimization manner. Extensive experiments on four benchmark data sets well demonstrate the superiority of the proposed method, which has led to better performances than compared counterpart methods.

INDEX TERMS Feature learning, multi-view, feature fusion, subspace learning.

I. INTRODUCTION

Rencent years have witnessed great progress in multi-view feature analysis, due to the emerging of large scale data sources, e.g., millions or billions of images, videos, and texts are produced from the internet day and night; To cope with all these various kinds of data, researchers have developed useful machine learning algorithms to conduct image, video and text analysis. As the first step of data analysis, various kinds of features (views) should be extracted, such as the SIFT features [1] from image, dense trajectory features [2] from videos, and bag of words [3] features from texts. To cope with these various kinds of features, multi-view feature learning [4] is proposed. From the perspective of feature fusing, multi-view feature learning can also be recognized as one kind of feature fusing strategy, however, the main difference between multi-view feature learning and original feature fusing lies in that 1) multi-view feature learning address more on discovering the correlations from different features such as CCA [5], and 2) multi-view learning are successively utilized to specific machine learning tasks such as multi-view clustering [6], [7], multi-view dimensionality reduction [8], [9], and multi-view semi-supervised learning [10], [11].

Taking content based image retrival as an example, to enhance the retrival correct rate, multiple visual features for an input image should be extracted, e.g., the histogram of oriented gradients (HOG) feature [12], scale invariant feature transform (SIFT) feature [1], mid-level bag of visual word feature (BOW) [13], and current leading convolutional neural network feature (CNN) [14]–[16]; different features can describe images from different aspects, and reveal more inherent properties from the images than single view features in some extent; then multi-view feature learning will

be carried out based on these proposed features, which can further improve the image retrieval accuracy tremendously, compared with corresponding single view methods. Specifically, multi-view feature learning can be classified into three categories [4], [17], [18], i.e., 1) co-training based algorithms, 2) multiple kernel learning based algorithms, and 3) subspace learning based algorithms.

Co-training [19], as a semi-supervised learning scheme, considers the situation that each sample only has two independent views; co-training first trains two respective classifier using labeled data from each view, then the two classifiers are enhanced by maximizing the mutual agreement on the unlabeled data from these two distinct views [17]. To handle more complex multi-view learning tasks, co-training has been incorporated into many learning algorithms, thus leading to variants of co-training; when expectation-maximization (EM) is carried out in co-training manner, we get the co-EM algorithm [20]; co-EM version of support vector machines (SVM) is further proposed in [21]; by introducing active learning into multi-view learning, co-testing is presented in [22]; furthermore, bayesian co-training [23] is proposed by introducing the bayesian undirected graphical model into co-training; [24] advocates to treat co-training as combinative label propagation over two views; through combining the simplicity of k-means clustering and linear discriminant analysis in a co-training manner, a multi-view clustering algorithm is presented in [25]. To successfully accomplish the above co-training induced algorithms, the following three assumptions are essential, i.e., 1) sufficiency: each view is sufficient for classifying its own data, 2) compatibility: the classifiers of both views can output the same labels for co-occurring features with a high probability, and 3) conditional independence: data from each view is conditional independent [17]. In multiple kernel learning (MKL), each kernel is corresponding to one view, thus MKL is treated as one series of multi-view learning. MKL learns multiple kernels linearly or non-linearly; a number of MKL learning algorithms [17], [26]–[39], have been proposed and demonstrated their superiority than previous single kernel learning methods. However, MLK based multi-view learning algorithms still fail to capture the correlation between different views.

To discover feature correlation from each view, subspace based feature learning algorithms have been proposed. Canonical correlation analysis (CCA) [40], maximizing the correlation from two views, seeks a latent subspace in an unsupervised manner; kernel CCA [41], pursuing maximally correlated nonlinear projections from two views, is further proposed by extending CCA into its kernel subspace; other extensions of CCA include: sparse CCA [42], [43], bayesian CCA [44], deep CCA [45], multi-view CCA [46], and tensor CCA [47]; CCA relevant approaches have been applied to multi-view clustering [48], regression [49], and dimensionality reduction [50], [51]; Moreover, by incorporating Fisher discriminative analysis into the multi-view subspace learning, multi-view discriminant analysis is presented in [52]. There still exist some other methods to find the latent subspace for multi-view data, such as Markov network method [53], low-rank method [8] and Gaussian process method [54]. However, the above methods, which can learn a latent common subspace with various kinds of regularization, usually not consider the separate subspace within each views.

In this paper, we propose a novel multi-view feature learning framework, which can simutaneously learn one separate subspace for each view and one shared subspace for all the views. In this way, the learnt subspace for each view can preserve particularity and communality, which are essential for improving the performance of classification task. Then the samples in the separate subspace from each view and sample from the shared subspace are concatenated together, thus leading to the final representation for afterward classifier training and prediction. Meanwhile, motivated from recently proposed modified least square regression algorithm [55], we relax the labels of the training data and utilize retargeted least square regression as our classifier. Particularly, the transformation matrices (w.r.t. each view) for constructing the final feature representation and classifier parameters are jointly learnt until a local optima is reached. By modeling label relaxed classifier learning, separate subspace learning, and shared subspace learning into a joint framework, the performances for multi-view learning can be further improved greatly. The contributions of this paper are summarized as follows:

1. We proposed a retargeted multi-view feature learning framework (termed as RMVFL) which can jointly learn classifier parameters and transformation matrix for each view; the transformation matrix can transform the original feature space into a new subspace, in which the particularity for each view and commonality from all views are preserved.

2. To obtain the initial subspace for each view, PCA and generalized Fisher discriminative analysis (FDA) are utilized, thus leading to two kinds of RMVFL algorithms; we call them as PCA initialized RMVFL (PCA RMVFL) and FDA initialized RMVFL (FDA RMVFL), respectively.

3. In the proposed RMVFL frameworks, we use label relaxed least square regression classifier, which can directly learn the regression target from the training data rather than using the fixed regression labels. It can be deemed as retargeting (relaxing) of the labels, which can measure the classification error more accurately.

4. The RMVFL algorithms are solved elegantly and efficiently, with theoretically guaranteed convergence. Extensive experiments on four real world datasets well demonstrate the effectiveness and superior performance of RMVFL algorithms.

The rest of this paper is organized as follows. Section II describes the proposed retargeted muti-view feature learning framework (RMVFL) detailly. Optimization of the RMVFL algorithm is illustrated in Section III, followed by convergence analysis of RMVFL algorithm in Section IV. Experimental results are presented in Section V. Finally, concluding remarks are offered in Section VI.

II. THE PROPOSED RETARGETED MULTI-VIEW FEATURE LEARNING FRAMEWORK

In this section, we first detailly show our proposed retargeted multi-view feature learning framework, followed by the description of classification procedure for testing samples.

A. THE RMVFL FRAMEWORK

Throughout the whole paper, we use bold uppercase letters to represent matrices, meanwhile, bold lowercase letters are used to denote vectors. Specially, suppose $\mathbf{A} \in \mathbb{R}^{m\times s}$ is an arbitrary matrix; its ith row and jth column are $\mathbf{a}^i$ and $\mathbf{a}_j$, respectively; A_{ij} is the element in $\mathbf{A}$, that is located at the position of the ith row and jth column; $\|\mathbf{A}\|_F^2 = \sum_{i=1}^{m}\sum_{j=1}^{s} A_{ij}^2 = Tr(\mathbf{A}^T\mathbf{A}) = Tr(\mathbf{A}\mathbf{A}^T)$ is the Frobenius norm of $\mathbf{A}$.

As shown in Figure 1, suppose that there exist K views for each input sample, i.e., we have K groups of data $\mathbf{X}_i = \left[\mathbf{x}_1^i, \mathbf{x}_2^i, \cdots, \mathbf{x}_n^i\right] \in \mathbb{R}^{d_i\times n}$, $i = 1, 2, \cdots, K$ w.r.t. these K views. Here, d_i is the dimensionality of samples within the ith view, n is the number of training samples, which is the same for all these K views. Suppose that the number of categories for the processed dataset is C, then the fixed label matrix can be denoted as $\mathbf{Y} \in \mathbb{R}^{n\times C}$; suppose that the ith sample is from the j class, then the elements in the ith row of $\mathbf{Y}$ are $\mathbf{y}^i = [0, \cdots, 0, 1, 0, \cdots, 0]$, in which element 1 is located at the j position in $\mathbf{y}^i$; we utilize y_i to represent the ground truth label of the ith sample in $\mathbf{X}_q (q = 1, 2, \cdots, K)$, and labels of the ith sample are the same for these K views. To uncover particularity from each view, and meanwhile, to get more accurate classifier parameters, we propose the following optimization objective:

$$\begin{aligned}\min_{(\mathbf{W},\mathbf{b}),(\mathbf{P}_i,\mathbf{b}_i,\mathbf{Z}_i),\mathbf{T}} & \sum_{i=1}^{K}\left(\left\|\mathbf{X}_i^T\mathbf{P}_i + \mathbf{1}\mathbf{b}_i^T - \mathbf{Z}_i\right\|_F^2 + \lambda_1\|\mathbf{P}_i\|_F^2\right) \\ & + \gamma\left\|\left[\mathbf{Z}_1 \ \cdots \ \mathbf{Z}_K\right]\mathbf{W} + \mathbf{1}\mathbf{b}^T - \mathbf{T}\right\|_F^2 \\ & + \lambda_2\|\mathbf{W}\|_F^2 \\ s.t.\ T_{iy_i} - \max_{j\neq y_i} T_{ij} \geq 1, & \ i = 1, 2, \cdots, n, \end{aligned} \tag{1}$$

where $\mathbf{Z}_i \in \mathbb{R}^{n\times d}$ is the learnt separate subspace w.r.t. $\mathbf{X}_i$, $\mathbf{P}_i \in \mathbb{R}^{d_i\times d}$ is the transformation matrix w.r.t. $\mathbf{Z}_i$, and $\mathbf{1} = [1, 1, \cdots, 1]^T \in \mathbb{R}^n$ is a vector, whose elements all equal to 1; $\mathbf{T} \in \mathbb{R}^{n\times C}$ is the regression target which is learnt with the constraint that for each sample (in the concatenated subspace), the margin between the true target and the false target should be larger than 1, i.e., $T_{iy_i} - \max_{j\neq y_i} T_{ij} \geq 1, i = 1, 2, \cdots, n$; furthermore, $\mathbf{W} \in \mathbb{R}^{(\sum_{i=1}^{K} d_i)\times C}$ is the classifier parameter, and $\mathbf{b} \in \mathbb{R}^C, \mathbf{b}_i \in \mathbb{R}^d, i = 1, 2, \cdots, K$ are the bias term w.r.t. the classifier term and the subspace learning term in Eqn. 1, respectively; λ_1 and λ_2 are used to avoid the overfitting during the learning of Eqn. 1, and γ is a tradeoff parameter, which can balance the importantance of classifier learning and the subspace learning. Note that in this paper we assume that the dimensionalities of the learnt K separate subspace w.r.t. K views are all d ($\min_{1\leq i\leq K} d_i$). By setting each separate subspace to have the same dimensionalities, we can tremendously reduce the time consumption of tuning the dimensionality of each subspace, meanwhile, the improved performance is still large. To further consider the shared information from each view, we reformulate Eqn. 1 as the following compact model:

$$\begin{aligned}\min_{(\mathbf{W},\mathbf{b}),(\mathbf{P}_i,\mathbf{b}_i,\mathbf{Z}_i),\mathbf{Z},\mathbf{T}} & \sum_{i=1}^{K}\left(\left\|\mathbf{X}_i^T\mathbf{P}_i + \mathbf{1}\mathbf{b}_i^T - [\mathbf{Z}\mathbf{Z}_i]\right\|_F^2 + \lambda_1\|\mathbf{P}_i\|_F^2\right) \\ & + \gamma\left\|\left[\mathbf{Z}\mathbf{Z}_1 \ \cdots \ \mathbf{Z}_K\right]\mathbf{W} + \mathbf{1}\mathbf{b}^T - \mathbf{T}\right\|_F^2 \\ & + \lambda_2\|\mathbf{W}\|_F^2 \\ s.t.\ T_{iy_i} - \max_{j\neq y_i} T_{ij} \geq 1, & \ i = 1, 2, \cdots, n, \end{aligned} \tag{2}$$

where $\mathbf{Z} \in \mathbb{R}^{n\times d_s}$ is the shared subspace for all the views, and the dimensionality of this subspace is d_s. In Eqn. 2, the dimensionalities of some matrices should be changed accordingly, i.e., $\mathbf{P}_i \in \mathbb{R}^{d_i\times(d_s+d)}$, $\mathbf{W} \in \mathbb{R}^{(\sum_{i=1}^{K} d_i+d_s)\times C}$, and $\mathbf{b}_i \in \mathbb{R}^{d_s+d}$.

B. CLASSIFICATION FOR TESTING SAMPLES

In this subsection, we will present the classification scheme for testing samples, given the trained parameters in Eqn. 2. Specifically, for testing sample $\mathbf{x}_t$ with K feature representations $\mathbf{x}_t^i, i = 1, 2, \cdots, K$, corresponding to K views. We first get the separate and shared subspace representation for each $\mathbf{x}_t^i$, as follow:

$$[\,\mathbf{z}_s^i \ \ \mathbf{z}^i\,] = \left(\mathbf{x}_t^i\right)^T\mathbf{P}_i, \tag{3}$$

where $\mathbf{z}_s^i \in \mathbb{R}^{1\times d_s}$, and $\mathbf{z}^i \in \mathbb{R}^{1\times d}$. We further calculate the averaged shared subspace representation as follows:

$$\bar{\mathbf{z}}_s = \frac{1}{K}\sum_{i=1}^{K}\mathbf{z}_s^i. \tag{4}$$

Then, the concatenated representation for testing sample $\mathbf{z}_t$ is

$$\mathbf{z}_{concat} = [\bar{\mathbf{z}}_s\mathbf{z}^1\mathbf{z}^2\cdots\mathbf{z}^K] \in {}^{1\times(Kd+d_s)}. \tag{5}$$

Based on $\mathbf{z}_{concat}$, we get the C dimensional predicted vector $\mathbf{y}^{predict} = \mathbf{W}^T\mathbf{z}_{concat}^T + \mathbf{b} \in \mathbb{R}^C$. Finally, the predicted label for $\mathbf{z}_t$ is achieved by

$$y_{label} = arg\max_{1\leq i\leq K} y_i^{predict}. \tag{6}$$

In this paper, we report the mAP of testing samples, which is more reasonable than classification accuracy, and mAP has been utilized in many literatures [56], [57] for multi-view feature learning.

III. OPTIMIZATION OF THE RMVFL FRAMEWORK

In this section, we first describe the alternative optimization procedure for solving RMVFL algorithm, followed by the discussion about the initialization of the separate and shared subspaces.

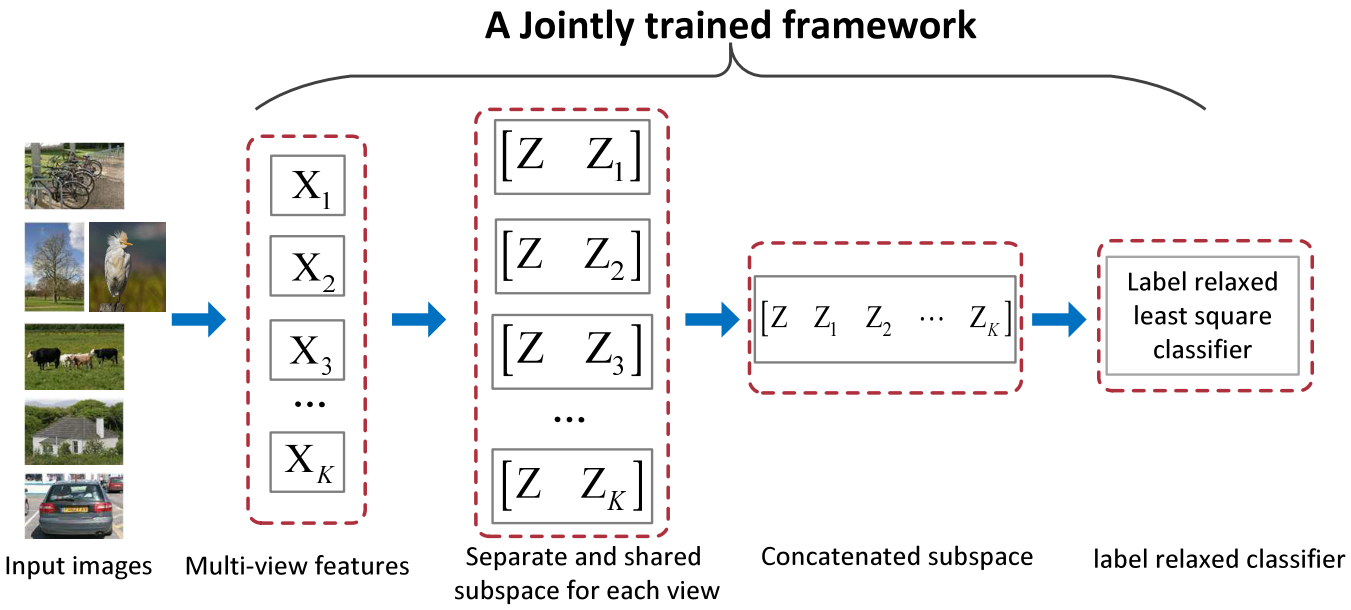


FIGURE 1. The flowchat of the proposed RMVFL framework. Best viewed in color.

A. OPTIMIZATION OF RMVFL ALGORITHM

The most difficult parts for solving Eqn. (2) are due to 1) the concatenated separate and shared subspace representation; and 2) the inequality constraints for each sample; To efficiently solve Eqn. (2), we utilize an alternative optimization procedure, wherein the optimum solution of each subproblem is achieved through gradient decent algorithm.

Update $\mathbf{b}_i$, $\mathbf{b}$*:* By taking derivative of the objective in Eqn. (2) w.r.t. $\mathbf{b}_i$ and $\mathbf{b}$, and setting the derivative to zero, we get

$$\mathbf{b}_i = \frac{1}{n}\left([\mathbf{Z}\ \mathbf{Z}_i]^T\mathbf{1} - \mathbf{P}_i{}^T\mathbf{X}_i\mathbf{1}\right). \tag{7}$$

$$\mathbf{b} = \frac{1}{n}\left(\mathbf{T}^T\mathbf{1} - \mathbf{W}^T[\mathbf{Z}\ \mathbf{Z}_1\ \mathbf{Z}_2 \cdots \mathbf{Z}_K]^T\mathbf{1}\right). \tag{8}$$

For each iteration, $\mathbf{b}_i$ and $\mathbf{b}$ can be updated based Eqn. (7) and Eqn. (8), respectively.

Update $\mathbf{P}_i$*:* Substituting Eqn. (7) and Eqn. (8) into Eqn. (2), the optimization problem of Eqn. (2) is changed as

$$\begin{aligned}\min_{\mathbf{W},(\mathbf{P}_i,\mathbf{Z}_i),\mathbf{Z},\mathbf{T}} \sum_{i=1}^{K}\Bigg(&\Big\|(\mathbf{I}-\frac{1}{n}\mathbf{1}\mathbf{1}^T)\mathbf{X}_i{}^T\mathbf{P}_i \\ &-(\mathbf{I}-\frac{1}{n}\mathbf{1}\mathbf{1}^T)[\mathbf{Z}\mathbf{Z}_i]\Big\|_F^2+\lambda_1\|\mathbf{P}_i\|_F^2\Bigg) \\ &+\gamma\left\|(\mathbf{I}-\frac{1}{n}\mathbf{1}\mathbf{1}^T)[\mathbf{Z}\mathbf{Z}_1\cdots\mathbf{Z}_K]\right. \\ &\left.\mathbf{W}-(\mathbf{I}-\frac{1}{n}\mathbf{1}\mathbf{1}^T)\mathbf{T}\right\|_F^2+\lambda_2\|\mathbf{W}\|_F^2 \\ s.t.\ T_{iy_i}-\max_{j\neq y_i}T_{ij}\geq 1,&\quad i=1,2,\cdots,n,\end{aligned} \tag{9}$$

where $\mathbf{I}\in\mathbb{R}^{n\times n}$ is an identity matrix; let $\mathbf{H}=\mathbf{I}-\frac{1}{n}\mathbf{1}\mathbf{1}^T$, then Eqn. (9) becomes

$$\begin{aligned}\min_{\mathbf{W},(\mathbf{P}_i,\mathbf{Z}_i),\mathbf{Z},\mathbf{T}} \sum_{i=1}^{K}\left(\left\|\mathbf{H}\mathbf{X}_i{}^T\mathbf{P}_i-\mathbf{H}[\mathbf{Z}\mathbf{Z}_i]\right\|_F^2+\lambda_1\|\mathbf{P}_i\|_F^2\right) \\ +\gamma\left\|\mathbf{H}\left[\mathbf{Z}\mathbf{Z}_1\ \cdots\ \mathbf{Z}_K\right]\mathbf{W}-\mathbf{H}\mathbf{T}\right\|_F^2 \\ +\lambda_2\|\mathbf{W}\|_F^2 \\ s.t.\ T_{iy_i}-\max_{j\neq y_i}T_{ij}\geq 1,\quad i=1,2,\cdots,n.\end{aligned} \tag{10}$$

By setting the derivative of objective in Eqn. (10) w.r.t. $\mathbf{P}_i$ to zero, we get

$$\mathbf{P}_i = \left(\mathbf{X}_i\mathbf{H}\mathbf{X}_i{}^T+\lambda_1\mathbf{I}\right)^{-1}\mathbf{X}_i\mathbf{H}[\mathbf{Z}\ \mathbf{Z}_i], \tag{11}$$

where $\mathbf{I}\in\mathbb{R}^{d_i\times d_i}$ is an identity matrix.

Update $\mathbf{W}$*:* Fix other parameters which are irrelevant with $\mathbf{W}$, then setting the derivative of objective in Eqn. (10) w.r.t. $\mathbf{W}$ to zero, we achieve the updating formula of $\mathbf{W}$ as follows

$$\begin{aligned}\mathbf{W} = \Big([\mathbf{Z}\ \mathbf{Z}_1\ \mathbf{Z}_2\cdots\ \mathbf{Z}_K]^T\mathbf{H}[\mathbf{Z}\ \mathbf{Z}_1\ \mathbf{Z}_2\cdots\ \mathbf{Z}_K]+\frac{\lambda_2}{\gamma}\mathbf{I}\Big)^{-1} \\ [\mathbf{Z}\ \mathbf{Z}_1\ \mathbf{Z}_2\cdots\ \mathbf{Z}_K]^T\mathbf{H}\mathbf{T}.\end{aligned} \tag{12}$$

Update $[\mathbf{Z}\ \mathbf{Z}_1\ \mathbf{Z}_2\cdots\ \mathbf{Z}_K]$*:* Taking other parameters as constants, Eqn. (10) is equivalent to

$$\begin{aligned}\min_{\mathbf{Z}_i,\mathbf{Z}} \sum_{i=1}^{K}\left\|\mathbf{H}\mathbf{X}_i{}^T\mathbf{P}_i-\mathbf{H}[\mathbf{Z}\mathbf{Z}_i]\right\|_F^2 \\ +\gamma\left\|\mathbf{H}\left[\mathbf{Z}\mathbf{Z}_1\ \cdots\ \mathbf{Z}_K\right]\mathbf{W}-\mathbf{H}\mathbf{T}\right\|_F^2.\end{aligned} \tag{13}$$

Let $\mathbf{D} = [\mathbf{Z}\ \mathbf{Z}_1\ \mathbf{Z}_2 \cdots\ \mathbf{Z}_K]$, and denote

$$f(\mathbf{D}) = \sum_{i=1}^{K} \left\| \mathbf{H}\mathbf{X}_i{}^T\mathbf{P}_i - \mathbf{H}[\mathbf{Z}\mathbf{Z}_i] \right\|_F^2 + \gamma \left\| \mathbf{H}\left[\mathbf{Z}\mathbf{Z}_1\ \cdots\ \mathbf{Z}_K\right]\mathbf{W} - \mathbf{H}\mathbf{T} \right\|_F^2. \quad (14)$$

Let $\mathbf{P}_i = [\mathbf{P}_{i1}\mathbf{P}_{i2}]$, where $\mathbf{P}_{i1} \in \mathbb{R}^{d_i \times d_s}$, $\mathbf{P}_{i2} \in \mathbb{R}^{d_i \times d}$; by setting the derivative of $f(\mathbf{D})$ w.r.t. $\mathbf{D}$ to zero, we get

$$\begin{aligned} &\gamma\mathbf{H}\mathbf{T}\mathbf{W}^T + [\textstyle\sum_{i=1}^{K} \mathbf{H}\mathbf{X}_i{}^T\mathbf{P}_{i1}\ , \mathbf{H}\mathbf{X}_1{}^T\mathbf{P}_{12}\ , \mathbf{H}\mathbf{X}_2{}^T\mathbf{P}_{22}, \\ &\qquad \cdots, \mathbf{H}\mathbf{X}_K{}^T\mathbf{P}_{K2}] \\ &= \gamma\mathbf{H}\mathbf{D}\mathbf{W}\mathbf{W}^T + \mathbf{H}\mathbf{D}\begin{bmatrix} K\mathbf{I}_{d_s} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_d & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{I}_d \end{bmatrix} \end{aligned} \quad (15)$$

where, $\mathbf{I}_{d_s} \in \mathbb{R}^{d_s \times d_s}$ and $\mathbf{I}_d \in \mathbb{R}^{d \times d}$ are two identity matrices. From Eqn. (15), we further obtain the final representation of $\mathbf{D}$ (Eqn. (16)), which is used for gradient updating.

$$\begin{aligned} \mathbf{D} = &\Big(\gamma\mathbf{T}\mathbf{W}^T + [\textstyle\sum_{i=1}^{K} \mathbf{X}_i{}^T\mathbf{P}_{i1}\ , \mathbf{X}_1{}^T\mathbf{P}_{12}, \\ &\mathbf{X}_2{}^T\mathbf{P}_{22}, \cdots, \mathbf{X}_K{}^T\mathbf{P}_{K2}]\Big) \\ &\left(\gamma\mathbf{W}\mathbf{W}^T + \begin{bmatrix} K\mathbf{I}_{d_s} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_d & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{I}_d \end{bmatrix}\right)^{-1} \end{aligned} \quad (16)$$

Update $\mathbf{T}$*:* Fix other parameters, and taking them as constant variables except $\mathbf{T}$, then Eqn. (2) is simplified as follows

$$\begin{aligned} &\min_{\mathbf{T}} \left\| \left[\mathbf{Z}\mathbf{Z}_1\ \cdots\ \mathbf{Z}_K\right]\mathbf{W} + \mathbf{1}\mathbf{b}^T - \mathbf{T} \right\|_F^2 = \|\mathbf{Q} - \mathbf{T}\|_F^2 \\ &s.t.\ T_{iy_i} - \max_{j \neq y_i} T_{ij} \geq 1, \quad i = 1, 2, \cdots, n. \end{aligned} \quad (17)$$

In Eqn. (17), $\mathbf{Q} = [\mathbf{Z}\mathbf{Z}_1 \cdots \mathbf{Z}_K]\mathbf{W} + \mathbf{1}\mathbf{b}^T \in \mathbb{R}^{n \times C}$ is the predition result of the objective; as we know, Eqn. (17) is a convex constrained quandratic programming (QP) problem, thus it can be recognized as n subproblems, e.g., the ith subproblem is as follows

$$\begin{aligned} &\min_{\mathbf{t}^i} \left\| \mathbf{q}^i - \mathbf{t}^i \right\|_2^2 = \sum_{j=1}^{C} \left(Q_{ij} - T_{ij}\right)^2 \\ &s.t.\ T_{iy_i} - \max_{j \neq y_i} T_{ij} \geq 1, \end{aligned} \quad (18)$$

where $\mathbf{q}^i = [Q_{i1}, Q_{i2}, \cdots, Q_{iC}]$ and $\mathbf{t}^i = [T_{i1}, T_{i2}, \cdots, T_{iC}]$ are the ith row from $\mathbf{Q}$ and $\mathbf{T}$, respectively; the label of the ith sample (corresponding to the ith row in $\mathbf{Q}$ and $\mathbf{T}$) is denoted as y_i; we need to solve n subproblems like Eqn. (18); Specifically, to solve Eqn. (18), we introduce vector $\boldsymbol{\rho} = [\rho_1, \rho_2, \cdots, \rho_C] \in \mathbb{R}^{1 \times C}$, and denote $\rho_j = 1 + T_{ij} - T_{iy_i}$. Then if $\rho_j \leq 0$, it indicates that the learnt intermediate varible T_{ij} satisfies the constraint in Eqn. (18); otherwise, if $\rho_j > 0$, it means that T_{ij} violates the constraint; suppose that the optimal value T_{iy_i} for the true class (y_i) equals to a small modification of the predicted result (Q_{iy_i}), i.e., $T_{iy_i} = Q_{iy_i} + \delta$, where δ needs to be learnt. To optimize $\{T_{ij} | j \neq y_i, j = 1, 2, \cdots, C\}$ w.r.t. the false classes, by fixing $T_{iy_i} = Q_{iy_i} + \delta$, Eqn. (18) can be modified as the following $C-1$ subproblems

$$\begin{aligned} &\min_{T_{ij}} \left(Q_{ij} - T_{ij}\right)^2 \\ &s.t.\ Q_{iy_i} + \delta - T_{ij} \geq 1, \quad \forall\, j \neq y_i. \end{aligned} \quad (19)$$

By solving the above constrained QP problems, we get $T_{ij} = Q_{ij} + \min\ (\delta - \rho_j, 0), \forall j \neq y_i$. Through the above discussions, we summarize the optimal solution of Eqn. (18) as

$$T_{ij} = \begin{cases} Q_{ij} + \delta, & if\ j = y_i \\ Q_{ij} + \min\left(\delta - \rho_j, 0\right), & if\ j \neq y_i. \end{cases} \quad (20)$$

Based on Eqn. (20), Eqn. (18) is changed as

$$\min_{\delta}\ h(\delta) = \delta^2 + \sum_{j \neq y_i} \left(min\left(\delta - \rho_j, 0\right)\right)^2. \quad (21)$$

Calculate the derivative of $h(\delta)$ w.r.t. δ, we get

$$h'(\delta) = 2\delta + 2\sum_{j \neq y_i} min\left(\delta - \rho_j, 0\right). \quad (22)$$

By setting $h'(\delta) = 0$, we can obtain the optimal value of δ, as follows

$$\delta = \frac{\sum\limits_{j \neq y_i} \rho_j \Pi\left(h'(\rho_j) > 0\right)}{1 + \sum\limits_{j \neq y_i} \Pi\left(h'(\rho_j) > 0\right)}, \quad (23)$$

where $\Pi(x)$ is an indicator function, with $\Pi(x) = 1$, if x is true; otherwise, $\Pi(x) = 0$. The optimal solution of Eqn. (18) can be achieved through Eqn. (20]) and Eqn. (23]); then the optimal solution for Eqn. (17) can be obtained by optimizing n subproblems, with the same procedure as optimizing Eqn. (18).

After deducing the formulas (used for gradient updating) w.r.t. all the unknown variables, an alternative optimization algorithm, listed in Algorithm 1, is proposed for solving the RMVFL framework.

B. PROCEDURES FOR INITIALIZING THE SEPARATE AND SHARED SUBSPACES

From the objective in Eqn. (2) and Algorithm 1, it can be seen that the initializing strategies of the separate and shared subspaces are important for subsequent convergence of the proposed algorithm; therefore, in this part, we illustrate the initializing strategies for $\mathbf{Z}$ and $\mathbf{Z}_i, i = 1, 2, \cdots, K$. Particularly, two kinds of initialization approaches are utilized, i.e., PCA [58] based initialization and FDA based initialization. The detailed procedure for PCA based initialization is shown in Figure 2.

Note that, to extract more than $C - 1$ useful components for FDA, we utilize a regularized version of FDA [59], [60].

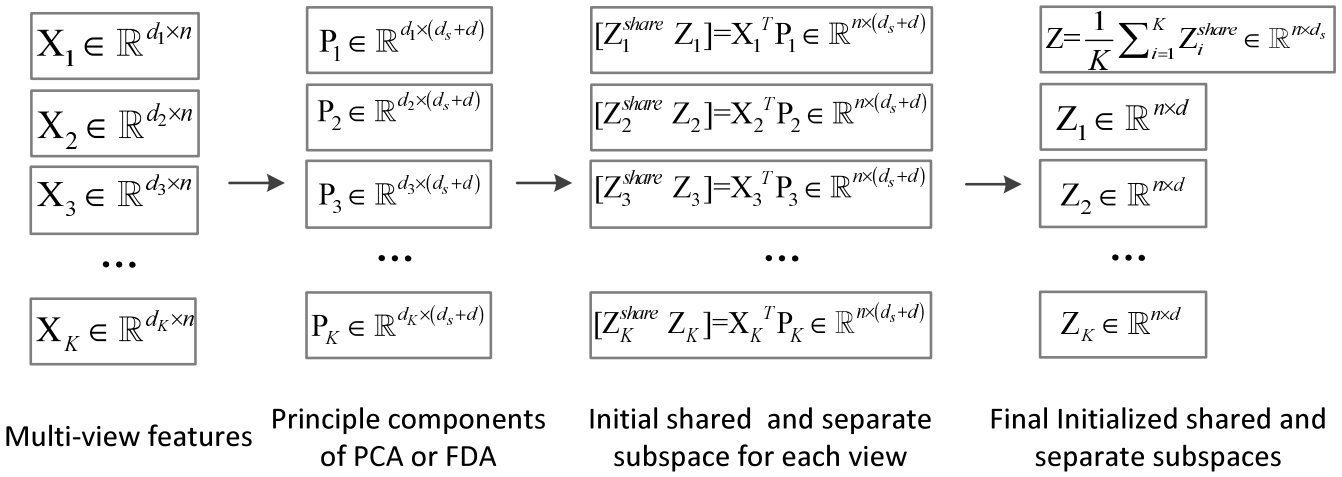


FIGURE 2. The procedure for PCA or FDA based initialization of the subspaces.

Algorithm 1 Optimization of RMVFL

Input: Training data from K views: $\mathbf{X}_v$, $v = 1, 2, \cdots, K$, labels: $\mathbf{Y}$, and λ_1, λ_2, γ, and *IterNum*.

Output: Label relaxed LSR parameters $(\mathbf{W}, \mathbf{b})$, K transformation matrices $\{\mathbf{P}_i | i = 1, 2, \cdots, K\}$ for K views, bias terms $\{\mathbf{b}_i | i = 1, 2, \cdots, K\}$ for K views, shared subspace $\mathbf{Z}$, separate subspace $\{\mathbf{Z}_i | i = 1, 2, \cdots, K\}$ for K views, and the label relaxed (retargeted) label matrix $\mathbf{T}$.

Initialization: initialize $\mathbf{Z}$, $\{\mathbf{Z}_v\}_{v=1}^K$, based on PCA or FDA (See Figure 2), initialize $\mathbf{T}$ as $\mathbf{Y}$, and denote $\mathbf{H} = I_n - \frac{1}{n} 1_n 1_n^T$.

1: **while** $s \leq$ *IterNum* **do**
2: Update $\mathbf{P}_i$ by Eqn. (11).
3: Update $\mathbf{W}$ by Eqn. (12).
4: Update $[\mathbf{Z}, \mathbf{Z}_1, \mathbf{Z}_2, \cdots, \mathbf{Z}_K]$ using Eqn. (16).
5: Update $\mathbf{T}$ using the procedure for optimizing Eqn. (17).
6: Update $\mathbf{b}_i$ using Eqn. (7).
7: Update $\mathbf{b}$ using Eqn. (8).
8: $s \leftarrow s + 1$
9: **end while**

IV. CONVERGENCE ANALYSIS

In this section, we present the detailed analysis on the convergence of Algorithm 1; for ease of description, we denote the objective function in Eqn. (2) as $\Gamma\ ((\mathbf{W}, \mathbf{b}), (\mathbf{P}_i, \mathbf{b}_i, \mathbf{Z}_i)|_{i=1}^K, \mathbf{Z}, \mathbf{T})$. The convergence of Eqn. (2) is equivalent to the following theorem.

Theorem: During the iterating of Algorithm 1, the value of $\Gamma\ ((\mathbf{W}, \mathbf{b}), (\mathbf{P}_i, \mathbf{b}_i, \mathbf{Z}_i)|_{i=1}^K, \mathbf{Z}, \mathbf{T})$ monotonically decreases, and it can converge to local minima.

Proof: Suppose that the qth iteration of Algorithm 1 has been complished; then the current value of the objective function in Eqn. (2) is taken as $\Gamma\ ((\mathbf{W}^q, \mathbf{b}^q), (\mathbf{P}_i^q, \mathbf{b}_i^q, \mathbf{Z}_i^q)|_{i=1}^K, \mathbf{Z}^q, \mathbf{T}^q)$. For the next $(q+1)$th iteration of Algorithm 1, we fix $\mathbf{T}$ as $\mathbf{T}^q$, and Eqn. (2) becomes

$$\min_{(\mathbf{W},\mathbf{b}),(\mathbf{P}_i,\mathbf{b}_i,\mathbf{Z}_i)|_{i=1}^K,\mathbf{Z}} \Gamma\ ((\mathbf{W}, \mathbf{b}), (\mathbf{P}_i, \mathbf{b}_i, \mathbf{Z}_i)|_{i=1}^K, \mathbf{Z}, \mathbf{T}^q). \tag{24}$$

To solve Eqn. (24) efficiently, we first fix $\mathbf{Z}_i|_{i=1}^K$ and $\mathbf{Z}$ as $\mathbf{Z}_i^q|_{i=1}^K$ and $\mathbf{Z}^q$, respectively; then, $\Gamma\ ((\mathbf{W}, \mathbf{b}), (\mathbf{P}_i, \mathbf{b}_i, \mathbf{Z}_i^q)|_{i=1}^K, \mathbf{Z}^q, \mathbf{T}^q)$ is convex w.r.t. $(\mathbf{W}, \mathbf{b})$ and $(\mathbf{P}_i, \mathbf{b}_i)|_{i=1}^K$, respectively; therefore the following inequality is obtained

$$\begin{aligned}&\Gamma\ ((\mathbf{W}^{q+1}, \mathbf{b}^{q+1}), (\mathbf{P}_i^{q+1}, \mathbf{b}_i^{q+1}, \mathbf{Z}_i^q)|_{i=1}^K, \mathbf{Z}^q, \mathbf{T}^q)\\ &\quad\leq \Gamma\ ((\mathbf{W}^q, \mathbf{b}^q), (\mathbf{P}_i^q, \mathbf{b}_i^q, \mathbf{Z}_i^q)|_{i=1}^K, \mathbf{Z}^q, \mathbf{T}^q).\end{aligned} \tag{25}$$

Similarly, by fixing $(\mathbf{W}, \mathbf{b}), (\mathbf{P}_i, \mathbf{b}_i)|_{i=1}^K$ as the current optimal solutions, i.e., $(\mathbf{W}^{q+1}, \mathbf{b}^{q+1}), (\mathbf{P}_i^{q+1}, \mathbf{b}_i^{q+1})|_{i=1}^K$; $\Gamma\ ((\mathbf{W}^{q+1}, \mathbf{b}^{q+1}), (\mathbf{P}_i^{q+1}, \mathbf{b}_i^{q+1}, \mathbf{Z}_i)|_{i=1}^K, \mathbf{Z}, \mathbf{T}^q)$ is convex w.r.t. $(\mathbf{Z}_i)|_{i=1}^K, \mathbf{Z}$, thus leading to

$$\begin{aligned}&\Gamma\ ((\mathbf{W}^{q+1}, \mathbf{b}^{q+1}), (\mathbf{P}_i^{q+1}, \mathbf{b}_i^{q+1}, \mathbf{Z}_i^{q+1})|_{i=1}^K, \mathbf{Z}^{q+1}, \mathbf{T}^q)\\ &\quad\leq \Gamma\ ((\mathbf{W}^{q+1}, \mathbf{b}^{q+1}), (\mathbf{P}_i^{q+1}, \mathbf{b}_i^{q+1}, \mathbf{Z}_i^q)|_{i=1}^K, \mathbf{Z}^q, \mathbf{T}^q).\end{aligned} \tag{26}$$

Furthermore, let $(\mathbf{W}, \mathbf{b}), (\mathbf{P}_i, \mathbf{b}_i, \mathbf{Z}_i)|_{i=1}^K, \mathbf{Z}$ take the current optimal solutions, i.e., $(\mathbf{W}^{q+1}, \mathbf{b}^{q+1}), (\mathbf{P}_i^{q+1}, \mathbf{b}_i^{q+1}, \mathbf{Z}_i^{q+1})|_{i=1}^K, \mathbf{Z}^{q+1}$; then $\Gamma\ ((\mathbf{W}^{q+1}, \mathbf{b}^{q+1}), (\mathbf{P}_i^{q+1}, \mathbf{b}_i^{q+1}, \mathbf{Z}_i^{q+1})|_{i=1}^K, \mathbf{Z}^{q+1}, \mathbf{T})$ is convex w.r.t. $\mathbf{T}$, thus yielding the following inequality

$$\begin{aligned}&\Gamma\ ((\mathbf{W}^{q+1}, \mathbf{b}^{q+1}), (\mathbf{P}_i^{q+1}, \mathbf{b}_i^{q+1}, \mathbf{Z}_i^{q+1})|_{i=1}^K, \mathbf{Z}^{q+1}, \mathbf{T}^{q+1})\\ &\quad\leq \Gamma\ ((\mathbf{W}^{q+1}, \mathbf{b}^{q+1}), (\mathbf{P}_i^{q+1}, \mathbf{b}_i^{q+1}, \mathbf{Z}_i^q)|_{i=1}^K, \mathbf{Z}^{q+1}, \mathbf{T}^q).\end{aligned} \tag{27}$$

Integrating Eqn. (25),(26), and (27) together, we have

$$\begin{aligned}&\Gamma\ ((\mathbf{W}^{q+1}, \mathbf{b}^{q+1}), (\mathbf{P}_i^{q+1}, \mathbf{b}_i^{q+1}, \mathbf{Z}_i^{q+1})|_{i=1}^K, \mathbf{Z}^{q+1}, \mathbf{T}^{q+1})\\ &\quad\leq \Gamma\ ((\mathbf{W}^q, \mathbf{b}^q), (\mathbf{P}_i^q, \mathbf{b}_i^q, \mathbf{Z}_i^q)|_{i=1}^K, \mathbf{Z}^q, \mathbf{T}^q).\end{aligned} \tag{28}$$

From Eqn. (28), we conclude that 1) the value of the objective function in Eqn. (2) is monotonically decreased; and

2) it will reach at a local minima after we stop the iteration of Algorithm 1.

Overall, Algorithm 1 is an alternative optimization procedure, which indicates that we can not get the global optimal solution of Eqn. (2); however, through the experiments in Section V, we find that the achieved local minima is enough to get the satisfactory performances, within a small range of iterations.

V. EXPERIMENTS

In this section, to evaluate the performance of the proposed RMVFL framework, we address the classification task; Four publicly available datasets, NUS-Wide-Object [61], Outdoor scene [62], MSRC-v1 [63] and Handwritten digits [1] [64], are utilized to conduct the experiments. The above four datasets are extensively used in the community of multi-view feature analysis; we give a detailed description of them as follows.

Detailed Description of the Used Datasets: Specifically, NUS-Wide-Object is a subset of the NUS-Wide dataset, and it consists of 30,000 object images with 31 categories. In the experiments w.r.t. this dataset, the official training and test partition is used, i.e., 17,927 training images and 12,073 testing images are randomly selected. Outdoor scene dataset is composed of 2,688 images, belonging to 8 outdoor scene categories, i.e., coast, mountain, forest, open country, street, inside city, tall buildings and highways. Four fifths images per category are randomly selected and taken as training set, the rest one fifth images are taken as testing set. MSRC-v1 is a scene dataset, which contains 210 images with totally seven categories (each category has 30 images). These seven categories are building, tree, airplane, cow, face, car, and bicycle. For the partition of training and test set, 105 images are used for training, the rest images are used for testing. Finally, for Handwritten digits dataset, there are totally 2,000 images of ten digits (0,1,2,3,4,5,6,7,8,9), the training and test partition is the same as the partition of Outdoor scene dataset. For the above for datasets, we report the average mAP and standard deviation for running five rounds. The detailed descriptions about the used four datasets are listed in Table 1.

TABLE 1. Descriptions of the used four multi-view datasets.

Feature ids	NUS-Wide-Object	Outdoor scene	MSRC-v1	Handwritten digits
1	Color histogram (64D)	GIST (512D)	Color moment (48D)	FOU (76D)
2	Color correlogram (144D)	Color moment (432D)	LBP (256D)	FAC (216D)
3	Edge direction histogram (73D)	HOG (256D)	HOG (100D)	KAR (64D)
4	Wavelet texture)(128D)	LBP (48D)	SIFT (1230D)	PIX (240D)
5	Block-wise color moments (225D)	-	GIST (512D)	ZER (47D)
6	BoW SIFT (500D)	-	CENTRIST (1320D)	-
Class number	31	8	7	10
Number of total samples	30,000	2688	210	2000

Parameter Settings: There are three parameters (λ_1,λ_2, γ, d, d_s) in Algorithm 1. λ_1 is used for avoiding overfitting problem w.r.t. $\mathbf{P}_i$, and it is taken as the same values for $\mathbf{P}_i, i = 1, 2, \cdots, K$. λ_1 is empirically tuned from $\{10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 10^{0}, 10^{1}, 10^{2}, 10^{3}\}$. Similarly, λ_2 is selected from $\{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 10^{0}, 10^{1}, 10^{2}, 10^{3}, 10^{4}, 10^{5}\}$. To balance the importance of subspace learning and label relaxed classifier learning, γ is optimized from $\{10^{-3}, 10^{-2}, 10^{-1}, 10^{0}, 10^{1}, 10^{2}, 10^{3}, 10^{4}, 10^{5}, 10^{6}\}$; Moreover, d is chosen from $\{1, 5, 10, 15, \cdots, \min_{1\leqslant i\leqslant K} d_i\}$, and d_s is chosen from $\{1, 2, 3, \cdots, C\}$ with C being the number of class.

The Compared Methods: To extensively evaluate the performance of the proposed RMVFL algorithm, we first compare RMVFL with the counterpart single view method and the the method by concatenating of multiple view features. Specifically, SVM classifier is utilized to obtain the classification results, given the single view and the concatenated multi-view features. Furthermore, several multiple kernel learning (MKL) methods are taken as the compared counterpart methods; the compared MKL approaches include

[1] https://archive.ics.uci.edu/ml/datasets/Multiple+Features

TABLE 2. mAP and their standard deviation (std) of different methods on the four datasets.

Methods	NUS-Wide-Object	Outdoor scene	MSRC-v1	Handwritten digits
SVM (View 1)	0.161±0.016	0.830±0.018	0.786±0.026	0.964±0.023
SVM (View 2)	0.152±0.018	0.743±0.015	0.774±0.022	0.764±0.021
SVM (View 3)	0.144±0.020	0.665±0.017	0.794±0.021	0.923±0.018
SVM (View 4)	0.153±0.019	0.581±0.019	0.798±0.019	0.958±0.023
SVM (View 5)	0.142±0.021	-	0.781±0.018	0.798±0.026
SVM (View 6)	0.147±0.017	-	0.799±0.025	-
SVM (All views)	0.187±0.021	0.846±0.014	0.802±0.018	0.969±0.022
SVM $l1$ MKL [26]	0.215±0.026	0.848±0.017	0.824±0.018	0.968±0.023
SVM l_2 MKL [27]	0.212±0.024	0.847±0.018	0.801±0.022	0.966±0.022
SVM l_∞ MKL [30]	0.223±0.019	0.852±0.021	0.829±0.021	0.975±0.018
LSSVM l_1 MKL [66]	0.198±0.021	0.837±0.019	0.812±0.026	0.971±0.019
LSSVM l_2 MKL [38]	0.192±0.022	0.840±0.014	0.819±0.019	0.967±0.022
LSSVM l_∞ MKL [67]	0.211±0.020	0.835±0.021	0.795±0.024	0.969±0.020
GP method [73]	0.190±0.019	0.835±0.018	0.826±0.017	0.969±0.025
LPboost-β [68]	0.229±0.017	0.859±0.021	0.818±0.022	0.972±0.017
LPboost-$\mathbf{B}$ [68]	0.227±0.014	0.861±0.023	0.810±0.022	0.970±0.014
Multi-view CCA [74]	0.236±0.025	0.874±0.028	0.832±0.021	0.975±0.023
Multirelational Classification [70]	0.268±0.022	0.894±0.024	0.865±0.013	0.987±0.012
MAPGG [72]	0.258±0.024	0.917±0.022	0.908± 0.019	0.993±0.007
KI$_2$SCA [71]	0.274±0.023	0.923±0.022	0.902± 0.025	0.992±0.011
MVCS (w/o shared) [57]	0.297±0.011	0.911±0.017	0.918±0.12	0.983±0.003
MVCS (w shared) [57]	0.309±0.008	0.929±0.013	0.928±0.013	0.991±0.002
MVCL (w/o shared) [56]	0.272±0.021	0.919±0.018	0.891±0.017	0.990±0.009
MVCL (w shared) [56]	0.299±0.021	0.941±0.014	0.932±0.019	0.995±0.005
PCA RMVFL (w/o shared)	0.340±0.004	0.939±0.011	0.971±0.012	0.994±0.003
PCA RMVFL (w shared)	0.341±0.004	0.945±0.003	0.977±0.010	0.996±0.001
FDA RMVFL (w/o shared)	0.347±0.004	0.946±0.002	0.980±0.006	0.997±0.002
FDA RMVFL (w shared)	0.351±0.006	0.950±0.002	0.986±0.004	0.996±0.002

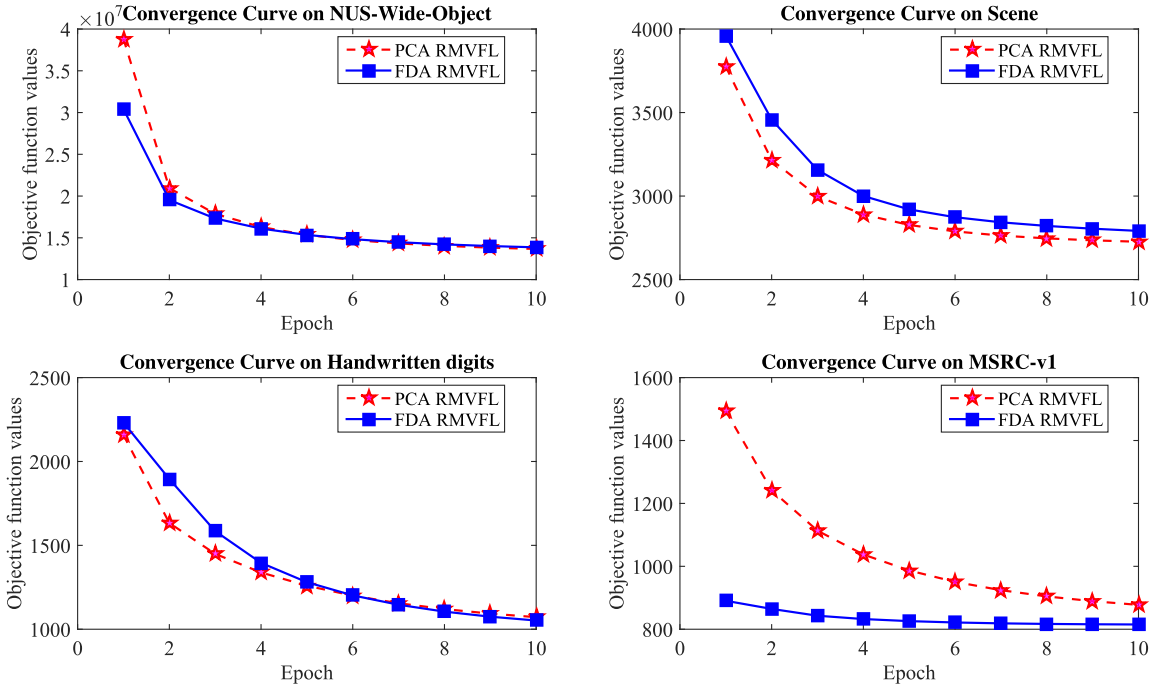


FIGURE 3. The convergent curves for PCA RMVFL and FDA RMVFL on the used four datasets.

1) SVM l_∞ MKL [65], 2) SVM l_2 MKL [27], 3) SVM l_1 MKL [65], 4) LSSVM l_1 MKL [66], 5) LSSVM l_2 MKL [38], and 6) LSSVM l_∞ MKL [67]; two LPboost methods are also used for comparisions, i.e., 1) LPboost-β [68], and 2) LPboost-$\mathbf{B}$ [68]; other methods, by discovering feature correlations among different views, are also compared; these

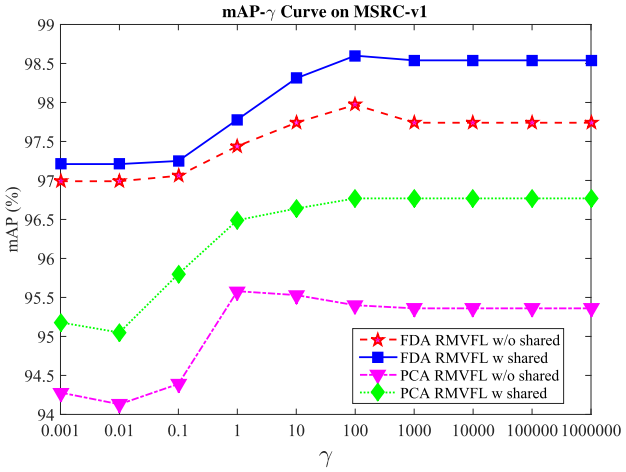


FIGURE 4. The mAP curves w.r.t. different values of γ on MSRC-v1 dataset. Here other parameters are fixed.

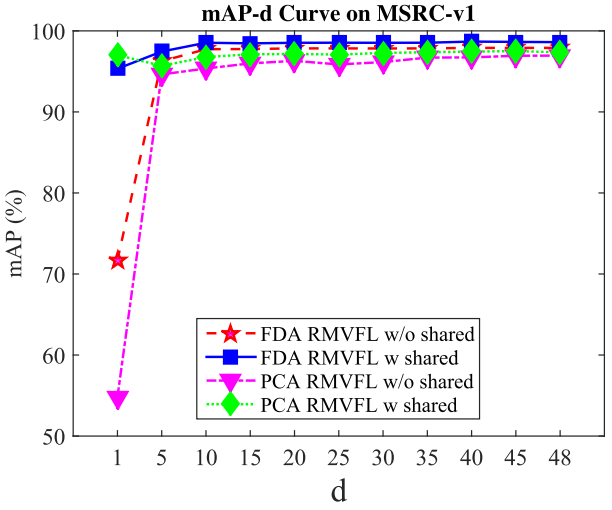


FIGURE 5. The mAP curves w.r.t. different values of d on MSRC-v1 dataset. Here other parameters are fixed.

methods are 1) multi-view CCA [69], 2) Multi-relational classification [70], 3) intra-view and inter-view correlation analysis [71], 4) manifold alignment [72], 5) Multi-view correlated learning with shared information (MVCS) [57], and 6) Multi-view feature learning with structure sparsity (MVCL) [56].

For SVM and MKL methods, Gaussian kernel is utilized for each two kinds of features, which is denoted as follows

$$\kappa\left(\mathbf{x}_i, \mathbf{x}_j\right) = \exp\left(-\sigma \left\|\mathbf{x}_i - \mathbf{x}_j\right\|_2^2\right), \tag{29}$$

where σ is selected from $\{10^{-6}, 10^{-4}, 10^{-2}, 10^{0}, 10^{2}, 10^{4}, 10^{6}]\}$, the same as literatures [56], [57]. In our experiments, we use the available implementation of MKL methods from [38]; Specifically, for the LSSVM l_∞ and the LSSVM l_2 approach, regularization parameter λ is estimated as the kernel coefficient of an identity matrix; similarly, for the LSSVM l_1 method, λ is set to 1. For LPboost methods, the publicly

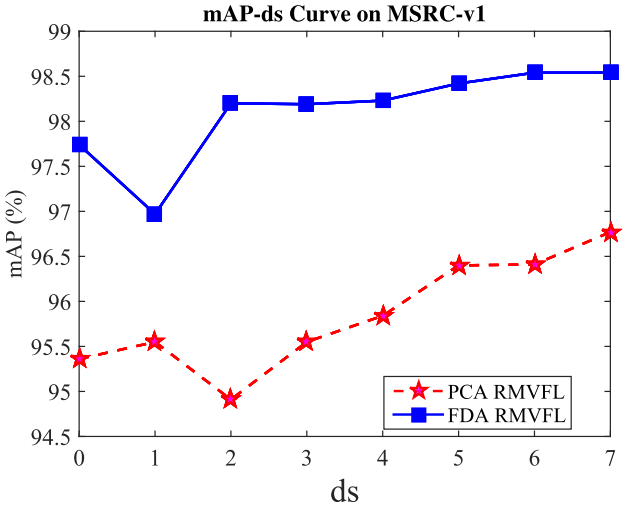


FIGURE 6. The mAP curves w.r.t. different values of d_s on MSRC-v1 dataset. Here other parameters are fixed.

available codes [2] [38] are used to reproduce the numbers on the four datasets; As for the regularization parameter C in other SVM based methods, we tune this parameter in the same range as the parameter γ in Algorithm 1.

A. EXPERIMENTAL RESULTS

In this part, we show all the compared results (mean average precision (mAP) with standard deviation) in Table 2.

From Table 2, we can conclude that 1) the proposed RMVFL frameworks (PCA RMVFL and FDA RMVFL) can consistently outperform the compared counterpart methods, 2) FDA RMVFL method usually achieves better results than PCA RMVFL ones, due to the introducing of discriminative information during the initialization of shared/separate subspace in Algorithm 1, 3) PCA/FDA RMVFL methods with shared subspace constraint can achieve better results than the methods without shared subspace constraint, and 4) all the multi-view methods are better than the methods using single view feature, e.g., the numbers of SVM MKL methods(Table 2) are much better than SVM methods trained using a single type of feature, by a large margin.

B. CONVERGENCE ANALYSIS BY EXPERIMENTS

Beside conducting the above experiments, we also illustrate the changing tendency of the objective function values in Eqn. (2), from which we can observe the convergence of Algorithm 1. It can be seen from Figure 3 that Algorithm 1 can reach to a stable state within 10 iterations, which further verify the theoretical proof of the convergence for Algorithm 1, in the perspective of experimental evaluation.

C. PARAMETER ANALYSIS

In this section, taking MSRC-v1 as an example, we conduct experiments to observe the changing tendency of performances w.r.t. the key parameters, i.e., γ, the

[2] http://files.is.tue.mpg.de/pgehler/projects/iccv09/

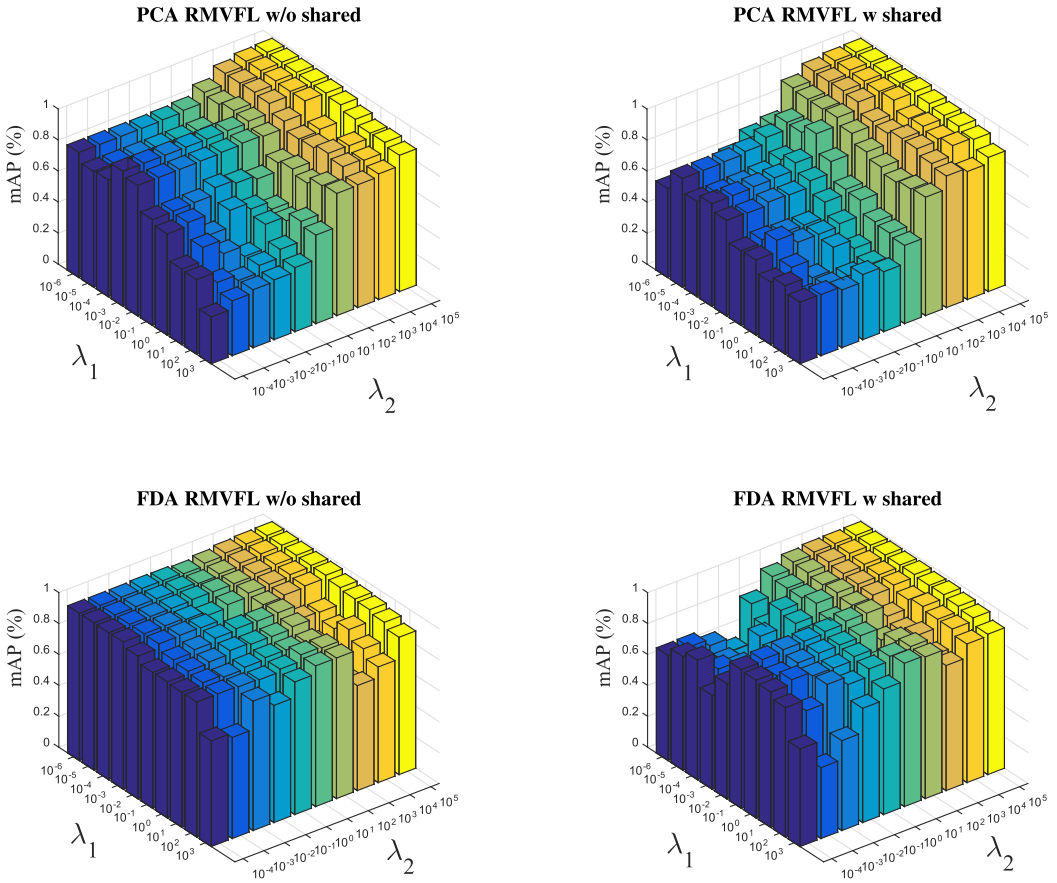


FIGURE 7. The mAP w.r.t. different values of λ_1 and λ_2 on MSRC-v1 dataset. Here other parameters are fixed.

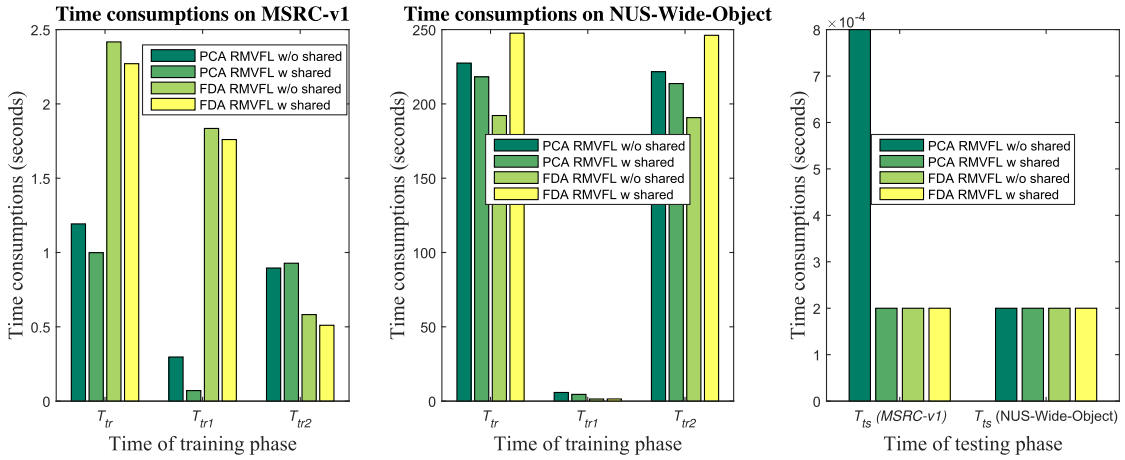


FIGURE 8. The time consumptions for both training and testing phase on MSRC-v1 and NUS-Wide-Object datasets.

dimensionality of the shared subspace d_s, the dimensionality of the separate subspace d, and λ_1, λ_2, which are used for avoiding overfitting problems during the model training. By fixing other parameters, and varying γ in $\{0.001, 0.01, 0.1, 1, 10, 100, 1000, 10000, 100000, 1000000\}$, we can get the mAP-$\gamma$ curve as shown in Figure 4 on MSRC-v1 dataset; Similarly, by varying d in $\{1, 5, 10, 15, \cdots, \min_{1 \leqslant i \leqslant K} d_i\}$, and d_s in $\{1, 2, 3, \cdots, C\}$ (C is the number of category), we obtain the mAP-d_s (Figure 6) curve and mAP-d curve (Figure 5), respectively. Moreover, we draw the parameter map (Figure 7) w.r.t. λ_1 and λ_2, which take values from $\{10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 10^{0}, 10^{1}, 10^{2}, 10^{3}\}$ and $\{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 10^{0}, 10^{1}, 10^{2}, 10^{3}, 10^{4}, 10^{5}\}$, respectively.

It can be concluded from Figure 5, Figure 6 and Figure 7 that 1) the mAP w.r.t. different values of γ are becoming stable while increasing the value of γ, 2) as the d_s or d is increased, the mAP is gradually increased as well, and 3) best mAP is achieved while λ_1 takes small values and λ_2 takes large values.

D. TIME COMPLEXITY ANALYSIS

In this subsection, we analyze the time consumption of Algorithm 1. As can be seen from Algorithm 1, the total time consumption T_{tr} for training phase consists of two parts, i.e., 1) the time consumption (T_{tr1}) for initializing $\mathbf{Z}$, $\{\mathbf{Z}_v\}|_{v=1}^{K}$; and 2) the time consumption (T_{tr2}) for parameter learning (step 2-7 in Algorithm 1); after learning all these parameters, we further conduct prediction; the average time consumption for each sample is denoted as T_{ts}. we take NUS-Wide-Object and MSRC-v1 as examples to illustrate the specific numbers of $T_{tr}(= T_{tr1} + T_{tr2}), T_{tr1}, T_{tr2}, T_{ts}$ (Figure 8).

VI. CONCLUSION AND FUTURE WORKS

In this paper, a retargeted multi-view feature learning framework, termed as RMVFL, is proposed; in the process of RMVFL learning, one separate subspace for each view, one shared subspace for all the views, and the label relaxed (retargeted) classifier are jointly learnt, through the alternative optimization manner; the theoretical convergence of RMVFL is gauranteed and proofed. Furthermore, two novel initialization procedures (PCA and FDA) for separate and shared subspaces are also presented, leading to the PCA RMVFL framework and FDA RMVFL framework, respectively. Experimental results on four benchmark datasets (for validating multi-view classification task) well demonstrate the effectiveness of our proposed RMVFL frameworks.

As convolutional neural network models are current leading visual recognition system, and there are no publicly available multi-view datasets, which contain both CNN features and traditional features; therefore, in the future, we will consider applying our framework to fuse CNN features and traditional features.

REFERENCES

[1] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.

[2] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu, "Action recognition by dense trajectories," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2011, pp. 3169–3176.

[3] G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray, "Visual categorization with bags of keypoints," in *Proc. Workshop Statist. Learn. Comput. Vis. (ECCV)*, Prague, Czech Republic, 2004, vol. 1. nos. 1–22. pp. 1–2.

[4] S. Sun, "A survey of multi-view machine learning," *Neural Comput. Appl.*, vol. 23, nos. 7–8, pp. 2031–2038, 2013.

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

[6] C. Zhang, H. Fu, S. Liu, G. Liu, and X. Cao, "Low-rank tensor constrained multiview subspace clustering," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 1582–1590.

[7] Y. Li, F. Nie, H. Huang, and J. Huang, "Large-scale multi-view spectral clustering via bipartite graph," in *Proc. 29th AAAI Conf. Artif. Intell.*, 2015, pp. 2750–2756.

[8] Z. Ding and Y. Fu, "Low-rank common subspace for multi-view learning," in *Proc. IEEE Int. Conf. Data Mining (ICDM)*, Dec. 2014, pp. 110–119.

[9] M. Gönen, G. B. Gönen, and F. Gürgen, "Bayesian multiview dimensionality reduction for learning predictive subspaces," in *Proc. 21st Eur. Conf. Artif. Intell.*, 2014, pp. 387–392.

[10] S. Sun, "Multi-view Laplacian support vector machines," in *Proc. Int. Conf. Adv. Data Mining Appl.*, 2011, pp. 209–222.

[11] X. Xie and S. Sun, "Multi-view Laplacian twin support vector machines," *Appl. Intell.*, vol. 41, no. 4, pp. 1059–1068, Dec. 2014.

[12] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2005, vol. 1, pp. 886–893.

[13] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong, "Locality-constrained linear coding for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2010, pp. 3360–3367.

[14] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[15] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "CNN features off-the-shelf: An astounding baseline for recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2014, pp. 806–813.

[16] Y. Jia *et al.*, "Caffe: Convolutional architecture for fast feature embedding," in *Proc. 22nd ACM Int. Conf. Multimedia*, 2014, pp. 675–678.

[17] C. Xu, D. Tao, and C. Xu. (2013). "A survey on multi-view learning." [Online]. Available: https://arxiv.org/abs/1304.5634

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

[19] A. Blum and T. Mitchell, "Combining labeled and unlabeled data with co-training," in *Proc. 11th Annu. Conf. Comput. Learn. Theory*, 1998, pp. 92–100.

[20] K. Nigam and R. Ghani, "Analyzing the effectiveness and applicability of co-training," in *Proc. 9th Int. Conf. Inf. Knowl. Manage.*, 2000, pp. 86–93.

[21] U. Brefeld and T. Scheffer, "Co-EM support vector learning," in *Proc. 21st Int. Conf. Mach. Learn.*, 2004, p. 16.

[22] I. Muslea, S. Minton, and C. A. Knoblock, "Active + semi-supervised learning = robust multi-view learning," in *Proc. ICML*, 2002, pp. 435–442.

[23] S. Yu, B. Krishnapuram, R. Rosales, and R. B. Rao, "Bayesian co-training," *J. Mach. Learn. Res.*, vol. 12, no. 1, pp. 2649–2680, Sep. 2011.

[24] W. Wang and Z.-H. Zhou, "A new analysis of co-training," in *Proc. 27th Int. Conf. Mach. Learn. (ICML)*, 2010, pp. 1135–1142.

[25] X. Zhao, N. Evans, and J.-L. Dugelay, "A subspace co-training framework for multi-view clustering," *Pattern Recognit. Lett.*, vol. 41, pp. 73–82, May 2014.

[26] G. R. G. Lanckriet, N. Cristianini, P. Bartlett, L. El Ghaoui, and M. I. Jordan, "Learning the kernel matrix with semidefinite programming," *J. Mach. Learn. Res.*, vol. 5, pp. 27–72, Jan. 2004.

[27] M. Kloft, U. Brefeld, P. Laskov, and S. Sonnenburg, "Non-sparse multiple kernel learning," in *Proc. NIPS Workshop Kernel Learn., Autom. Select. Kernels*, 2008.

[28] F. R. Bach, G. R. G. Lanckriet, and M. I. Jordan, "Multiple kernel learning, conic duality, and the SMO algorithm," in *Proc. 21st Int. Conf. Mach. Learn.*, 2004, p. 6.

[29] S. Sonnenburg, G. Rätsch, and C. Schäfer, "A general and efficient multiple kernel learning algorithm," in *Proc. Adv. Neural Inf. Process. Syst.*, 2006, pp. 1273–1280.

[30] S. Sonnenburg, G. Rätsch, C. Schäfer, and B. Schölkopf, "Large scale multiple kernel learning," *J. Mach. Learn. Res.*, vol. 7, pp. 1531–1565, Jul. 2006.

[31] A. Rakotomamonjy, F. Bach, S. Canu, and Y. Grandvalet, "More efficiency in multiple kernel learning," in *Proc. 24th Int. Conf. Mach. Learn.*, 2007, pp. 775–782.

[32] A. Rakotomamonjy, F. Bach, S. Canu, and Y. Grandvalet, "SimpleMKL," *J. Mach. Learn. Res.*, vol. 9, pp. 2491–2521, Nov. 2008.

[33] M. Szafranski, Y. Grandvalet, and A. Rakotomamonjy, "Composite kernel learning," *Mach. Learn.*, vol. 79, nos. 1–2, pp. 73–103, 2010.

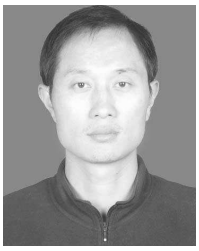
[34] C. Cortes, M. Mohri, and A. Rostamizadeh, "Learning non-linear combinations of kernels," in *Proc. Adv. Neural Inf. Process. Syst.*, 2009, pp. 396–404.

[35] N. Subrahmanya and Y. C. Shin, "Sparse multiple kernel learning for signal processing applications," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 5, pp. 788–798, May 2010.
[36] M. Varma and B. R. Babu, "More generality in efficient multiple kernel learning," in *Proc. 26th Annu. Int. Conf. Mach. Learn.*, 2009, pp. 1065–1072.
[37] Z. Xu, R. Jin, I. King, and M. Lyu, "An extended level method for efficient multiple kernel learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2009, pp. 1825–1832.
[38] S. Yu *et al.*, "L_2-norm multiple kernel learning and its application to biomedical data fusion," *BMC Bioinform.*, vol. 11, no. 1, p. 309, 2010.
[39] S. Wang, X. Chang, X. Li, G. Long, L. Yao, and Q. Z. Sheng, "Diagnosis code assignment using sparsity-based disease correlation embedding," *IEEE Trans. Knowl. Data Eng.*, vol. 28, no. 12, pp. 3191–3202, Dec. 2016.
[40] H. Hotelling, "Relations between two sets of variates," *Biometrika*, vol. 28, nos. 3–4, pp. 321–377, 1936.
[41] P. L. Lai and C. Fyfe, "Kernel and nonlinear canonical correlation analysis," *Int. J. Neural Syst.*, vol. 10, no. 5, pp. 365–377, Oct. 2000.
[42] X. Chen, H. Liu, and J. G. Carbonell, "Structured sparse canonical correlation analysis," in *Proc. Int. Conf. Artif. Intell. Statist.*, 2012, pp. 199–207.
[43] D. R. Hardoon and J. Shawe-Taylor, "Sparse canonical correlation analysis," *Mach. Learn.*, vol. 83, no. 3, pp. 331–353, 2011.
[44] A. Klami, S. Virtanen, and S. Kaski, "Bayesian canonical correlation analysis," *J. Mach. Learn. Res.*, vol. 14, pp. 965–1003, Apr. 2013.
[45] G. Andrew, R. Arora, J. Bilmes, and K. Livescu, "Deep canonical correlation analysis," in *Proc. Int. Conf. Mach. Learn.*, 2013, pp. 1247–1255.
[46] J. Vía, I. Santamaría, and J. Pérez, "A learning algorithm for adaptive canonical correlation analysis of several data sets," *Neural Netw.*, vol. 20, no. 1, pp. 139–152, 2007.
[47] Y. Luo, D. Tao, K. Ramamohanarao, C. Xu, and Y. Wen, "Tensor canonical correlation analysis for multi-view dimension reduction," *IEEE Trans. Knowl. Data Eng.*, vol. 27, no. 11, pp. 3111–3124, Nov. 2015.
[48] K. Chaudhuri, S. M. Kakade, K. Livescu, and K. Sridharan, "Multi-view clustering via canonical correlation analysis," in *Proc. 26th Annu. Int. Conf. Mach. Learn.*, 2009, pp. 129–136.
[49] S. M. Kakade and D. P. Foster, "Multi-view regression via canonical correlation analysis," in *Proc. Int. Conf. Comput. Learn. Theory*, 2007, pp. 82–96.
[50] B. Long, P. S. Yu, and Z. Zhang, "A general model for multiple view unsupervised learning," in *Proc. SIAM Int. Conf. Data Mining*, 2008, pp. 822–833.
[51] N. Chen, J. Zhu, F. Sun, and E. P. Xing, "Large-margin predictive latent subspace learning for multiview data analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 12, pp. 2365–2378, Dec. 2012.
[52] M. Kan, S. Shan, H. Zhang, S. Lao, and X. Chen, "Multi-view discriminant analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 1, pp. 188–194, Jan. 2016.
[53] N. Chen, J. Zhu, and E. P. Xing, "Predictive subspace learning for multi-view data: A large margin approach," in *Proc. Adv. Neural Inf. Process. Syst.*, 2010, pp. 361–369.
[54] L. Sigal, R. Memisevic, and D. J. Fleet, "Shared kernel information embedding for discriminative inference," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2009, pp. 2852–2859.
[55] X. Zhang, L. Wang, S. Xiang, and C. Liu, "Retargeted least squares regression algorithm," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 26, no. 9, pp. 2206–2213, Sep. 2015.
[56] Y.-S. Chang, F. Nie, and M.-Y. Wang, "Multiview feature analysis via structured sparsity and shared subspace discovery," *Neural Comput.*, vol. 29, no. 7, pp. 1986–2003, 2017.
[57] X. Xue, F. Nie, S. Wang, X. Chang, B. Stantic, and M. Yao, "Multi-view correlated feature learning by uncovering shared component," in *Proc. 31st AAAI Conf. Artif. Intell.*, 2017, pp. 2810–2816.
[58] S. Wold, K. Esbensen, and P. Geladi, "Principal component analysis," *Chemometrics Intell. Lab. Syst.*, vol. 2, nos. 1–3, pp. 37–52, 1987.
[59] D. Cai, X. He, and J. Han, "SRDA: An efficient algorithm for large-scale discriminant analysis," *IEEE Trans. Knowl. Data Eng.*, vol. 20, no. 1, pp. 1–12, Jan. 2008.
[60] D. Cai, *Spectral Regression: A Regression Framework for Efficient Regularized Subspace Learning*. Champaign, IL, USA: Univ. of Illinois at Urbana–Champaign, 2009.
[61] T.-S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y. Zheng, "NUS-WIDE: A real-world Web image database from National University of Singapore," in *Proc. ACM Int. Conf. Image Video Retr.*, 2009, p. 48.
[62] A. Monadjemi, B. T. Thomas, and M. Mirmehdi, "Experiments on high resolution images towards outdoor scene classification," Dept. Comput. Sci., Univ. Bristol, Bristol, U.K., Tech. Rep., 2002.
[63] K. Grauman and T. Darrell, "Unsupervised learning of categories from sets of partially matching image features," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 1. Jun. 2006, pp. 19–25.
[64] A. Asuncion and D. Newman, "UCI machine learning repository," Dept. Inf. Comput. Sci., Irvine, Univ. California, Irvine, CA, USA, 2007. [Online.] Available: http://www.ics.uci.edu/~mlearn/MLRepository.html
[65] M. Kloft, U. Brefeld, S. Sonnenburg, and A. Zien, "l_p-norm multiple kernel learning," *J. Mach. Learn. Res.*, vol. 12, pp. 953–997, Mar. 2011.
[66] J. A. K. Suykens, T. Van Gestel, and J. De Brabanter, *Least Squares Support Vector Machines*. Singapore: World Scientific, 2002.
[67] J. Ye, S. Ji, and J. Chen, "Multi-class discriminant kernel learning via convex programming," *J. Mach. Learn. Res.*, vol. 9, pp. 719–758, Apr. 2008.
[68] P. Gehler and S. Nowozin, "On feature combination for multiclass object classification," in *Proc. IEEE 12th Int. Conf. Comput. Vis.*, Sep./Oct. 2009, pp. 221–228.
[69] J. Rupnik and J. Shawe-Taylor, "Multi-view canonical correlation analysis," in *Proc. Conf. Data Mining Data Warehouses (SiKDD)*, 2010, pp. 1–4.
[70] H. Guo and H. L. Viktor, "Multirelational classification: A multiple view approach," *Knowl. Inf. Syst.*, vol. 17, no. 3, pp. 287–312, 2008.
[71] X.-Y. Jing, R.-M. Hu, Y.-P. Zhu, S.-S. Wu, C. Liang, and J.-Y. Yang, "Intra-view and inter-view supervised correlation analysis for multi-view feature learning," in *Proc. 28th AAAI Conf. Artif. Intell.*, 2014, pp. 1882–1889.
[72] C. Wang and S. Mahadevan, "Manifold alignment preserving global geometry," in *Proc. IJCAI*, 2013, pp. 1743–1749.
[73] A. Kapoor, K. Grauman, R. Urtasun, and T. Darrell, "Gaussian processes for object categorization," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 169–188, 2010.
[74] D. P. Foster, S. M. Kakade, and T. Zhang, "Multi-view dimensionality reduction via canonical correlation analysis," Tech. Rep. TR-2008-4, 2008.

GUO-SEN XIE received the Ph.D. degree in pattern recognition and intelligent systems from the Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2016. He is currently an Assistant Professor with the Department of Information Engineering College, Henan University of Science and Technology. His research interests include machine learning, deep learning, and their applications to object recognition. He received the Best Student Paper Awards from MMM'16.

XIAO-BO JIN received the Ph.D. degree in pattern recognition and intelligent systems from the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2009. He is currently an Associate Professor with the School of Information Science and Engineering, Henan University of Technology. His research interests include Web mining and machine learning. His work has appeared in *Pattern Recognition* and *Neurocomputing*.

ZHENG ZHANG received the B.S. degree from the Henan University of Science and Technology and the M.S. degree from the Shenzhen Graduate School, Harbin Institute of Technology, Shenzhen, China, in 2012 and 2014, respectively, where he is currently pursuing the Ph.D. degree in computer science and technology. His current research interests include pattern recognition, machine learning, and computer vision.



ZHONGHUA LIU received the B.S. degree in computer engineering from the First Aeronautical Institute of the Air Force, the M.S. degree in computer software and theory from Xihua University, and the Ph.D. degree in pattern recognition and intelligent systems from the Nanjing University of Science and Technology in 1998, 2005, and 2011, respectively. He is currently an Associate Professor with the Information Engineering College, Henan University of Science and Technology. His current research interests include pattern recognition, face recognition, image processing, and scene matching.

XIAOWEI XUE received the B.Sc. degree from Northeastern University in 2011. He is currently pursuing the Ph.D. degree with the College of Computer Science and Technology, Zhejiang University, Hangzhou, China. His research fields are computational intelligence, brain–machine interface, and cognitive computing model.

JIEXIN PU is currently a Professor with the Department of Information Engineering College, Henan University of Science and Technology. His research fields include pattern recognition, image processing, and computer vision.