LETTER **Communicated by Sen Wang**

Multiview Feature Analysis via Structured Sparsity and Shared Subspace Discovery

Yan-Shuo Chang
changyanshou@foxmail.com
School of Computer Science and Technology, Xidian University, Software Park, and Institute for Silk Road Research, Xi'an 71027, China

Feiping Nie
feipingnie@gmail.com
OPTIMA, Northwestern Polytechnical University, Xi'an 71027, China

Ming-Yu Wang
mingyuwang@gmail.com
School of Computer Science and Technology, Xidian University and Software Park, Xi'an 71027, China

Since combining features from heterogeneous data sources can significantly boost classification performance in many applications, it has attracted much research attention over the past few years. Most of the existing multiview feature analysis approaches separately learn features in each view, ignoring knowledge shared by multiple views. Different views of features may have some intrinsic correlations that might be beneficial to feature learning. Therefore, it is assumed that multiviews share subspaces from which common knowledge can be discovered. In this letter, we propose a new multiview feature learning algorithm, aiming to exploit common features shared by different views. To achieve this goal, we propose a feature learning algorithm in a batch mode, by which the correlations among different views are taken into account. Multiple transformation matrices for different views are simultaneously learned in a joint framework. In this way, our algorithm can exploit potential correlations among views as supplementary information that further improves the performance result. Since the proposed objective function is nonsmooth and difficult to solve directly, we propose an iterative algorithm for effective optimization. Extensive experiments have been conducted on a number of real-world data sets. Experimental results demonstrate superior performance in terms of classification against all the compared approaches. Also, the convergence guarantee has been validated in the experiment.

Neural Computation **29, 1986–2003** (2017)
doi:10.1162/NECO_a_00977

1 Introduction

With the proliferation of social networks, such as Facebook, Twitter, and Flickr, the volume of multimedia data exponentially increases, inevitably resulting in challenges to effective and efficient management of big media data (Yang et al., 2012; Chang, Nie, Wang et al., 2016; Chang, Nie, Yang, Zhang, & Huang, 2016; Fan, Chang, & Tao, 2017). Content-based categorization using visual features has always been focused on as a practical solution to the problem. In computer vision areas, a number of low-level visual features have been invented to describe visual information in a compact way. Meanwhile, researchers on machine learning have proposed many feature analysis approaches to understanding the features further. One direction is to utilize multiple heterogeneous features from different views (Chang, Yu, Yang, & Xing, 2016a, 2016b; Wang, Chang, Li, Long et al., 2016; Xue et al., 2017). Intuitively, an object can be characterized in various ways regarding different perspectives. For example, to describe visual objects in images, histogram of oriented gradients (HOG) (Dalal & Triggs, 2005), speeded-up robust features (SURF) (Bay, Tuytelaars, & Gool, 2006), and scale-invariant feature transform (SIFT) (Lowe, 2004) have been widely used to extract local visual features. Features in various views describe the data from different perspectives to meet the particular properties; for example, SIFT is robust to noise, changes of illumination, and rotation. If the heterogeneous features are properly integrated into a well-designed algorithm, better performance can be expected. Multiview feature learning therefore stems from this motivation. Previous work (Wang et al., 2014; Cai, Nie, & Huang, 2013; Chen et al., 2012; Conrad & Mester, 2013; Luo et al., 2017) has demonstrated that multiview feature learning can reduce noise and improve statistical significance. A number of multiple kernel learning (MKL) algorithms (Sonnenburg, Rätsch, Schäfer, & Schölkopf, 2006; Suykens & Vandewalle, 1999; Kloft, Brefeld, Laskov, & Sonnenburg, 2008; Yu et al., 2010; Ye, Ji, & Chen, 2008; Lanckriet, Cristianini, Bartlett, Ghaoui, & Jordan, 2004; Wang, Chang, Li, Sheng, & Chen, 2016) also learn different multiple kernels from heterogeneous features in multiview learning frameworks. Normally, MKL-based approaches learn an ensemble of kernels for a certain application to achieve better performance.

Although multiview learning approaches have achieved good performance in various applications, they neglect two problems that may be essential to improve the performance of multiview feature analysis further. First, features from each view are separately learned, which makes the learning framework unable to exploit the shared knowledge across multiple views. For example, texture-based and shape-based visual features should have some intrinsic correlations when capturing a certain object (e.g., the sun in images). It is presumed that exploiting shared knowledge is beneficial to feature learning. Second, most of the existing multiview feature analysis methods simply assume that features in all the views are equally

important for different classes. They cannot balance the weights of heterogeneous features when combining them together.

In this letter, we propose a novel multiview feature learning algorithm that exploits correlations among different views for visual-based applications. Our solution to the first problem mentioned above is to assume that there exist shared subspaces that contain shared knowledge across multiviews. Based on this assumption, our objective function squeezes out common components in the shared subspaces as much as possible in iterations until convergence is reached. Regarding the second challenge, we propose to use group ℓ_1-norm (G_1-norm) for regularization (Wang, Nie, & Huang, 2013; Chang, Nie, Yang, & Huang, 2014). From its definition, G_1-norm uses ℓ_2-norm within each view and ℓ_1-norm among different views. Hence, once a specific view of features is recognized as not informative for certain groups of objects, the algorithm will downgrade its importance by assigning zero weight to features in that view. Otherwise, large weights will be given. Consequently, the G_1-norm is capable of capturing relationships among different views. In this letter, we name our proposed algorithm multiview correlation feature learning (MVCL).

Existing multiview manifold learning algorithms (e.g., Wang & Mahadevan, 2013; Ham, Lee, & Saul, 2005) construct mapping functions that project data instances from different input domains to a new lower-dimensional space. However, they fail to take the shared component into consideration. In addition, Ham et al. (2005) design their approach in a semisupervised fashion.

The main contributions of this work can be summarized as follows:

1. We propose a novel multiview feature learning algorithm that takes the correlations among different views into consideration to improve subsequent classification performance. Based on the shared subspace learning, our algorithm can automatically exploit shared knowledge deeply buried across the multiviews.
2. In our algorithm, the importance of features in different views is not equally treated. The weights of features are dynamically adapted according to the sparsity condition across views. Those nonzero weights reflect their corresponding views as more informative than their counterparts with zero weight. In this way, our algorithm has more flexibility when learning multiview features than others.
3. Although the proposed objective function is nonsmooth, we derive an iterative algorithm to optimize the objective function effectively and efficiently. The convergence is theoretically guaranteed and empirically tested. From the experimental results, the proposed algorithm converges within 10 iterations in most of the cases.
4. Extensive experiments are conducted on several real-world data sets. The evaluation results show that our algorithm yields superior performance against all other compared multiview learning approaches.

The rest of this letter is organized as follows. We describe in detail the proposed feature learning framework in section 2, followed by a corresponding optimization algorithm in section 3 and convergence analysis in section 4. The experimental results are reported and discussed in section 5. We conclude section 6.

2 Multiview Correlated Feature Learning Framework

We begin by summarizing the notations and definitions used in this letter. Matrices and vectors are written as boldface uppercase letters and boldface lowercase letters, respectively. For an arbitrary matrix $\boldsymbol{A}$, its ith row and jth column are denoted as $\boldsymbol{a}^i$ and $\boldsymbol{a}_j$, respectively. Given n training samples $\boldsymbol{x}_i^j|_{i=1}^n \in \mathbb{R}^{d_j}$ in the jth view, the training data matrix is represented by $\boldsymbol{X}_j = \left[\boldsymbol{x}_1^j, \ldots, \boldsymbol{x}_n^j\right]$. d_j is the feature dimension in the jth view, $j = 1, \cdots, k$. k is the number of views. The label matrix is denoted as $\boldsymbol{Y} \in \mathbb{R}^{n\times c}$, and c is the number of classes.

Basically, we propose to learn the features by minimizing an objective function in the framework as follows:

$$\min_{\boldsymbol{W},\boldsymbol{P}_i,\boldsymbol{b},\boldsymbol{b}_i,\boldsymbol{Z}_i} \sum_{i=1}^{k} \|\boldsymbol{X}_i^T\boldsymbol{P}_i + \boldsymbol{1}\boldsymbol{b}_i^T - \boldsymbol{Z}_i\|_F^2 + \alpha\|[\boldsymbol{Z}_1 \cdots \boldsymbol{Z}_k]\boldsymbol{W} + \boldsymbol{1}\boldsymbol{b}^T - \boldsymbol{Y}\|_F^2 + \beta\mathcal{R}, \tag{2.1}$$

where $\boldsymbol{Z}_i$ is learned features of the ith view. $\mathcal{R}$ is a regularization function. With the first least-square loss function, we use the transformation matrix $\boldsymbol{P}_i$ to map the ith view into a subspace $\boldsymbol{Z}_i$. Then we concatenate all the projected subspaces learned in the first loss function from each view and obtain the learned features $[\boldsymbol{Z}_1, \cdots, \boldsymbol{Z}_k]$. We use the second least-square loss function to measure the loss incurred by $\boldsymbol{W}$ on the learned features. Note that there are a number of loss functions. The discussions between loss functions are beyond our focus in this letter in which we choose the least-square loss function for its good performance and simplicity.

We assume that there exists a subspace shared by all the views. We denote the shared subspace as $\boldsymbol{Z}$. Then, for each learned feature from each view, it becomes $[\boldsymbol{Z}\,\boldsymbol{Z}_i]$, and the concatenated learned features should be denoted as $[\boldsymbol{Z}\,\boldsymbol{Z}_1, \cdots, \boldsymbol{Z}_k]$.

Hence, our objective function becomes

$$\min_{\boldsymbol{W},\boldsymbol{P}_i,\boldsymbol{b},\boldsymbol{b}_i,\boldsymbol{Z},\boldsymbol{Z}_i} \sum_{i=1}^{k} \|\boldsymbol{X}_i^T\boldsymbol{P}_i + \boldsymbol{1}\boldsymbol{b}_i^T - [\boldsymbol{Z}\,\boldsymbol{Z}_i]\|_F^2 + \alpha\|[\boldsymbol{Z}\,\boldsymbol{Z}_1 \cdots \boldsymbol{Z}_k]\boldsymbol{W} + \boldsymbol{1}\boldsymbol{b}^T - \boldsymbol{Y}\|_F^2 + \beta\mathcal{R}. \tag{2.2}$$

Note that if there is no shared component among different views of features, we simply set the shared component $\boldsymbol{Z}$ to an empty matrix.

It has been theoretically and empirically indicated that features of a specific view may be informative or not informative for specific classes. Inspired by this fact, group ℓ_1-norm (G_1-norm) is used as a regularization and is defined as follows:

$$\|\boldsymbol{W}\|_{G_1} = \sum_{i=1}^{c} \sum_{j=1}^{k} \|\boldsymbol{w}_i^j\|_2. \tag{2.3}$$

The objective function of the proposed multiview feature learning algorithm is given by

$$\begin{aligned} \min_{\boldsymbol{W},\boldsymbol{P}_i,\boldsymbol{b},\boldsymbol{b}_i,\boldsymbol{Z}_i} & \sum_{i=1}^{k} \|\boldsymbol{X}_i^T\boldsymbol{P}_i + \mathbf{1}\boldsymbol{b}_i^T \\ & - [\boldsymbol{Z}\,\boldsymbol{Z}_i]\|_F^2 + \alpha\|[\boldsymbol{Z}\,\boldsymbol{Z}_1 \cdots \boldsymbol{Z}_k]\boldsymbol{W} + \mathbf{1}\boldsymbol{b}^T - \boldsymbol{Y}\|_F^2 + \beta\|\boldsymbol{W}\|_{G_1}. \end{aligned} \tag{2.4}$$

The sparsity between different views is enhanced by the G_1-norm since it adopts the ℓ_2-norm within each view and the ℓ_1-norm between different views. For example, the objective function will assign very small values to a specific view of features when they are not discriminative for certain tasks. Otherwise, their weights will be assigned large values. In this way, the correlations between different views can be captured.

3 Optimization Algorithm

The difficulties of solving the objective function in equation 2.4 are because of the concatenation of learned intermediate representations and the G_1-norm. We propose to efficiently tackle this problem in the following steps.

By setting the derivative of equation 2.4 with regard to b, b_i to zero, we have

$$\boldsymbol{b} = \frac{1}{n}\boldsymbol{Y}^T\mathbf{1} - \frac{1}{n}\boldsymbol{W}^T[\boldsymbol{Z}\,\boldsymbol{Z}_1, \cdots, \boldsymbol{Z}_k]^T\mathbf{1} \tag{3.1}$$

$$\boldsymbol{b}_i = \frac{1}{n}\boldsymbol{Z}_i^T\mathbf{1} - \frac{1}{n}\boldsymbol{P}_i^T\boldsymbol{X}_i\mathbf{1} \tag{3.2}$$

Substituting $\boldsymbol{b}$ and $\boldsymbol{b}_i$ in equation 2.4 with equations 3.1 and 3.2, the problem becomes

$$\begin{aligned} \min_{\boldsymbol{W},\boldsymbol{P}_i,\boldsymbol{b},\boldsymbol{b}_i} & \sum_{i=1}^{k} \|(\boldsymbol{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^T)\boldsymbol{X}_i^T\boldsymbol{P}_i - \left(\boldsymbol{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^T\right)\boldsymbol{Z}_i\|_F^2 \\ & + \alpha\|\left(\boldsymbol{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^T\right)[\boldsymbol{Z}\,\boldsymbol{Z}_1, \cdots, \boldsymbol{Z}_k]\boldsymbol{W} - \left(\boldsymbol{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^T\right)\boldsymbol{Y}\|_F^2 \\ & + \beta\|\boldsymbol{W}\|_{G_1}, \end{aligned}$$

where $\boldsymbol{I}$ is an identity matrix. Denoting $\boldsymbol{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^T = \boldsymbol{H}$, the problem arrives at

$$\min_{\boldsymbol{W},\boldsymbol{P}_i,\boldsymbol{b},\boldsymbol{b}_i} \sum_{i=1}^{k} \|\boldsymbol{H}\boldsymbol{X}_i^T\boldsymbol{P}_i - \boldsymbol{H}\boldsymbol{Z}_i\|_F^2 + \alpha\|\boldsymbol{H}[\boldsymbol{Z}\,\boldsymbol{Z}_1,\ldots,\boldsymbol{Z}_k]\boldsymbol{W} - \boldsymbol{H}\boldsymbol{Y}\|_F^2 + \beta\|\boldsymbol{W}\|_{G_1}. \tag{3.3}$$

We replace the variables $\boldsymbol{H}\boldsymbol{Z}_i$ by $\boldsymbol{Z}_i$. Then the problem becomes

$$\min_{\boldsymbol{W},\boldsymbol{W}_i,\boldsymbol{Z},\boldsymbol{Z}_i} \sum_{i=1}^{k} \|\boldsymbol{H}\boldsymbol{X}_i^T\boldsymbol{P}_i - \boldsymbol{Z}_i\|_F^2 + \alpha\|[\boldsymbol{Z}\,\boldsymbol{Z}_1,\cdots,\boldsymbol{Z}_k]\boldsymbol{W} - \boldsymbol{H}\boldsymbol{Y}\|_F^2 + \beta\|\boldsymbol{W}\|_{G_1}. \tag{3.4}$$

By setting the derivatives of equation 3.4 with regard to $\boldsymbol{P}_i$ to zero, we have

$$\boldsymbol{P}_i = (\boldsymbol{X}_i\boldsymbol{H}\boldsymbol{X}_i^T)^{-1}\boldsymbol{X}_i\boldsymbol{H}\boldsymbol{Z}_i. \tag{3.5}$$

In multiview classification, $\boldsymbol{W} = [\boldsymbol{w}_1,\cdots,\boldsymbol{w}_c] \in \mathbb{R}^{d\times c}$, where d is the dimension of the learned representation. By taking the derivatives of equation 3.4 with regard to $\boldsymbol{w}_i$ and setting it to zero, we have

$$\boldsymbol{w}_i = \alpha(\alpha[\boldsymbol{Z},\boldsymbol{Z}_1,\cdots,\boldsymbol{Z}_k]^T\boldsymbol{H}[\boldsymbol{Z},\boldsymbol{Z}_1,\cdots,\boldsymbol{Z}_k] + 2\beta\boldsymbol{D}^i)^{-1}([\boldsymbol{Z},\boldsymbol{Z}_1,\cdots,\boldsymbol{Z}_k]^T(\boldsymbol{y}_i - \boldsymbol{b}_i)), \tag{3.6}$$

where $\boldsymbol{D}^i$ is a block diagonal matrix with the jth diagonal block as $\frac{1}{2\|\boldsymbol{w}_i^j\|_2}\boldsymbol{I}_j$, and $\boldsymbol{I}_j$ is an identity matrix with size d_j, $\boldsymbol{w}_i^j$ is used to measure the weights of features in the jth view, and $\boldsymbol{b}_i \in \mathbb{R}^{n\times 1}$ is the bias of the second loss function.

Substituting $\boldsymbol{W}$ and $\boldsymbol{P}_i$ in equation 3.4 with equations 3.6 and 3.5 and setting the derivatives with regard to $[\boldsymbol{Z}\,\boldsymbol{Z}_1,\cdots,\boldsymbol{Z}_k]$, we obtain

$$[\boldsymbol{Z}\,\boldsymbol{Z}_1,\cdots,\boldsymbol{Z}_k]\begin{pmatrix} k\boldsymbol{I} & & \\ & \boldsymbol{I} & \\ & & \cdots \\ & & & \boldsymbol{I} \end{pmatrix} - \left[\sum_{i=1}^{k}\boldsymbol{H}\boldsymbol{X}_i^T\boldsymbol{W}_{i1}, \boldsymbol{H}\boldsymbol{X}_1^T\boldsymbol{W}_{12},\ldots,\boldsymbol{H}\boldsymbol{X}_k^T\boldsymbol{W}_{k2}\right] + \alpha[\boldsymbol{Z}\,\boldsymbol{Z}_1,\cdots,\boldsymbol{Z}_k]\boldsymbol{W}\boldsymbol{W}^T - \alpha\boldsymbol{H}\boldsymbol{Y}\boldsymbol{W}^T = 0 \tag{3.7}$$

$$\Rightarrow [\boldsymbol{Z}\,\boldsymbol{Z}_1\,\cdots\,\boldsymbol{Z}_k] = \left(\left[\sum_{i=1}^{k}\boldsymbol{H}\boldsymbol{X}_i^T\boldsymbol{W}_{i1}\,\boldsymbol{H}\boldsymbol{X}_1^T\boldsymbol{W}_{12}\,\ldots\,\boldsymbol{H}\boldsymbol{X}_k^T\boldsymbol{W}_{k2}\right]\right.$$
$$\left.+\alpha\boldsymbol{H}\boldsymbol{Y}\boldsymbol{W}^T\right)\left(\begin{pmatrix} k\boldsymbol{I} & & \\ & \boldsymbol{I} & \\ & & \cdots \\ & & & \boldsymbol{I}\end{pmatrix}+\alpha\boldsymbol{W}\boldsymbol{W}^T\right)^{-1}, \tag{3.8}$$

where $\boldsymbol{W} = [\boldsymbol{W}_{i1}, \boldsymbol{W}_{i2}]$.

Based on the above mathematical deduction, an iterative algorithm is proposed to optimize the objective function, equation 2.4, which is summarized in algorithm 1. The most expensive step is the inverse operation in equation 3.6. Hence, the computational complexity is $O(d^3)$.

4 Convergence Analysis

In this section, we prove that algorithm 1 converges by the following theorem. We begin with a lemma;

Lemma 1. *By fixing* $\mathbf{P}_i|_{i=1}^k$ *and* $\mathbf{W}$, *we can get the global solutions for* $\mathbf{Z}_i|_{i=1}^k$. *Similarly, we can get the global solutions for* $\mathbf{P}_i|_{i=1}^k$ *and* $\mathbf{W}$ *with fixed* $\mathbf{Z}_i|_{i=1}^k$.

Proof. By fixing $\mathbf{Z}_i|_{i=1}^k$, we can convert the objective function to a convex optimization problem with regard to $\mathbf{P}_i|_{i=1}^k$ and $\mathbf{W}$. Hence, the global solutions for $\mathbf{P}_i|_{i=1}^k$ and $\mathbf{W}$ can be obtained by setting the derivative of equation 3.4 to zero, respectively. In the same manner, we can also prove that by fixing $\mathbf{P}_i|_{i=1}^k$ and $\mathbf{W}$, we can get the global solutions for $\mathbf{Z}_i|_{i=1}^k$. □

Theorem 1. *The objective function value shown in equation 3.4 monotonically decreases until convergence by applying the proposed algorithm.*

Proof. Suppose after the rth iteration, we get $(\mathbf{P}_{i=1}^k)^r$, $(\mathbf{b}_{i=1}^k)^r$, $\mathbf{W}^r$, $\mathbf{b}^r$, $\mathbf{Z}^r$ and $\mathbf{Z}_i^r$. In the next iteration, we fix $\mathbf{Z}$ as $\mathbf{Z}^r$ and $\mathbf{Z}_i$ as $\mathbf{Z}_i^r$, and we solve for $\mathbf{P}_i$ and $\mathbf{W}$. We can get the following inequality according to lemma 1:

$$\begin{aligned}
&\sum_{i=1}^{k}\|\mathbf{X}_i^T\mathbf{P}_i^{r+1}+\mathbf{1}\left(\mathbf{b}_i^{r+1}\right)^T-\mathbf{Z}_i^r\|_{\mathbf{F}}^2+\alpha\|\left[\mathbf{Z}^r\,\mathbf{Z}_1^r,\cdots,\mathbf{Z}_k^r\right]\mathbf{W}^{r+1}\\
&+\mathbf{1}\left(\mathbf{b}^{r+1}\right)^T-\mathbf{Y}\|_F^2+\beta\sum_{i=1}^{c}\sum_{j=1}^{k}\frac{\|\mathbf{w}_i^j\|_2^2}{2\|\mathbf{w}_i^j\|_2}\\
\leq\ &\sum_{i=1}^{k}\|\mathbf{X}_i^T\mathbf{P}_i^{r}+\mathbf{1}\left(\mathbf{b}_i^{r}\right)^T-\mathbf{Z}_i^r\|_F^2+\alpha\|\left[\mathbf{Z}^r\,\mathbf{Z}_1^r,\cdots,\mathbf{Z}_k^r\right]\mathbf{W}^{r}\\
&+\mathbf{1}\left(\mathbf{b}^{r}\right)^T-\mathbf{Y}\|_F^2+\beta\sum_{i=1}^{c}\sum_{j=1}^{k}\frac{\|\mathbf{w}_i^j\|_2^2}{2\|\mathbf{w}_i^j\|_2}.
\end{aligned}\tag{4.1}$$

Algorithm 1: Multiview Correlated Feature Learning.

Data: Training data $\boldsymbol{X}_i|_{i=1}^k \in \mathbb{R}^{d_i \times n}$

Training data labels $\mathbf{Y} \in \mathbb{R}^{n \times c}$

Parameters α and α

Result:

$\mathbf{W} \in \mathbb{R}^{d \times c}$, $\mathbf{P}_i|_{i=1}^k \in \mathbb{R}^{d_i \times d}$, $\mathbf{b} \in \mathbb{R}^{c \times 1}$, $\mathbf{b}_i \in \mathbb{R}^{d_i \times 1}$

Initialize $\mathbf{Z}_i$ with PCA on $\mathbf{X}_i\mathbf{X}_i^T$;

Initialize $\mathbf{Z}$ with PCA on $\sum_{i=1}^k \mathbf{X}_i\mathbf{X}_i^T$;

Initialize $\mathbf{W} \in \mathbb{R}^{d \times c}$;

repeat

  Calculate the block diagonal matrices $\mathbf{D}^i (1 \le i \le c)$, where the jth diagonal block of $\mathbf{D}^i$ is $\frac{1}{2\|\mathbf{w}_i^j\|_2}\mathbf{I}_j$;

  For each $\mathbf{w}_i (1 \le i \le c)$, update it according to equation 3.6;

  Update $\mathbf{P}_i$ according to equation 3.5;

  Update $[\mathbf{Z}\ \mathbf{Z}_1, \cdots, \mathbf{Z}_k]$ according to equation 3.8;

  Update $\mathbf{b}$ according to equation 3.1;

  Update $\mathbf{b}_i$ according to equation 3.2;

until *Convergence*;

Return $\mathbf{W}, \mathbf{P}_i, \mathbf{b}, \mathbf{b}_i$ for $1 \le i \le k$.

In the same manner, when we fix $\mathbf{W}$ and $\mathbf{P}_i$, the following inequality holds:

$$\sum_{i=1}^{k} \|\mathbf{X}_i^T\mathbf{P}_i^r + \mathbf{1}\left(\mathbf{b}_i^r\right)^T - \left[\mathbf{Z}^{r+1}\mathbf{Z}_i^{r+1}\right]\|_F^2 + \alpha\|\left[\mathbf{Z}^{r+1}\ \mathbf{Z}_1^{r+1}, \cdots, \mathbf{Z}_k^{r+1}\right]\mathbf{W}^r + \mathbf{1}\left(\mathbf{b}^r\right)^T - \mathbf{Y}\|_F^2 + \beta \sum_{i=1}^{c}\sum_{j=1}^{k} \frac{\|\mathbf{w}_i^j\|_2^2}{2\|\mathbf{w}_i^j\|_2}$$

$$\leq \sum_{i=1}^{k} \|\mathbf{X}_i^T \mathbf{P}_i^r + \mathbf{1}\left(\mathbf{b}_i^r\right)^T - \left[\mathbf{Z}^r \mathbf{Z}_i^r\right]\|_F^2 + \alpha \|\left[\mathbf{Z}^r \mathbf{Z}_1^r, \cdots, \mathbf{Z}_k^r\right]\mathbf{W}^r + \mathbf{1}\left(\mathbf{b}^r\right)^T - \mathbf{Y}\|_F^2 + \beta \sum_{i=1}^{c}\sum_{j=1}^{k} \frac{\|\mathbf{w}_i^j\|_2^2}{2\|\mathbf{w}_i^j\|_2}. \tag{4.2}$$

By integrating equations 4.1 and 4.2, we obtain

$$\begin{aligned} &\sum_{i=1}^{k} \|\mathbf{X}_i^T \mathbf{P}_i^{r+1} + \mathbf{1}\left(\mathbf{b}_i^{r+1}\right)^T - \left[\mathbf{Z}^{r+1} \mathbf{Z}_i^{r+1}\right]\|_F^2 + \alpha \|\left[\mathbf{Z}_1^{r+1}, \cdots, \mathbf{Z}_k^{r+1}\right]\mathbf{W}^{r+1} \\ &+ \mathbf{1}\left(\mathbf{b}^{r+1}\right)^T - \mathbf{Y}\|_F^2 + \beta \sum_{i=1}^{c}\sum_{j=1}^{k} \frac{\|\mathbf{w}_i^j\|_2^2}{2\|\mathbf{w}_i^j\|_2} \\ \leq &\sum_{i=1}^{k} \|\mathbf{X}_i^T \mathbf{P}_i^r + \mathbf{1}\left(\mathbf{b}_i^r\right)^T - \left[\mathbf{Z}^r \mathbf{Z}_i^r\right]\|_{\mathbf{F}}^2 + \alpha \|\left[\mathbf{Z}_1^r, \cdots, \mathbf{Z}_k^r\right]\mathbf{W}^r \\ &+ \mathbf{1}\left(\mathbf{b}^r\right)^T - \mathbf{Y}\|_F^2 + \beta \sum_{i=1}^{c}\sum_{j=1}^{k} \frac{\|\mathbf{w}_i^j\|_2^2}{2\|\mathbf{w}_i^j\|_2}. \end{aligned} \tag{4.3}$$

From equation 4.3, we can clearly see that the objective function value decrease after each iteration. Thus, theorem 1 has been proved. □

5 Experiment

In this section, we experimentally evaluate the performance of the proposed method MVCL for classification tasks. We first compare our method with the related state-of-the-art multiview methods on four real-world data sets in which the evaluation of shared component is also studied. Then we study the performance variance with regard to parameter sensitivity and the convergence of algorithm 1.

5.1 Data Set Description. We evaluate our new multiview learning framework on four broadly used benchmark data sets. Each data set has a certain number of types of features (views), summarized in Table 1.

We used four data sets:

- *NUS-WIDE-OBJECT data set* (Chua et al., 2009). This is a light version of the NUS-WIDE data set consisting of 30,000 real-world object images and 31 object categories. It is widely used to compared different multiview learning methods in terms of image annotation. In our experiment, the official split is adopted: 17,927 training images and

Table 1: Details of Data Sets and Multiview Features Used in the Experiments.

Feature ID	NUS-WIDE-Object	Outdoor Scene	MSRC-v1	Handwritten Digit
1	Color histogram (64-D)	GIST (512-D)	Color Moment (48-D)	FOU(76-D)
2	Color correlogram (144-D)	Color Moment (432-D)	LBP (256-D)	FAC(216-D)
3	Edge direction histogram (73-D)	HOG (256-D)	HOG (100-D)	KAR (64-D)
4	Wavelet texture (128-D)	LBP (48-D)	SIFT (1230-D)	PIX (240-D)
5	Block-wise color moments (225-D)	-	GIST (512-D)	ZER (47-D)
6	BoW SIFT (500-D)	-	CENTRIST (1320-D)	-
Number of classes	31	8	8	10
Size	30,000	2688	210	2000

12,073 testing images. Each image contained 1134 features within six views.

- *OUTDOOR SCENE* (Monadjemi, Thomas, & Mirmehdi, 2002). This data set contains 2688 color images that belong to eight outdoor scene categories: street, coast, forest, mountain, open country, inside city, highways, and tall buildings. Each image has 1248 features within four views.
- *MSRC* (Grauman & Darrell, 2006). MSRC is a scene recognition data set and consists of 240 images with eight classes. Following Grauman and Darrell (2006), we select seven classes, and each class has 30 images. The selected classes are as follows: building, tree, airplane, cow, face, car, and bicycle. Each images has 3466 features within six views.
- *Handwritten Digits*. This data set consists of 2000 data instances for 0 to 9 10-digit classes. In our experiment, we use five publicly available features to represent multiple views, and each instance has 643 features within five views.

5.2 Experiment Setup. To evaluate the performance of our method for classification tasks, we compare MVCL with several single-view and multiview learning methods. For single-view learning methods, we select the popular standard support vector machine (SVM) to compare the multiview classification performance of the proposed algorithm with their corresponding single-view counterparts and the concatenation of all types of features. In all our experiments, SVM is implemented by LIBMSVM-software package (Chang & Lin, 2011).

For multiview learning methods, several different kinds of state-of-the-art multiview methods are adopted for the comparison. First, we compare the results of our method with several kinds of multiple kernel learning methods (MKL): SVM ℓ_p MKL, LSSVM ℓ_p MKL, LPboost, and GP method. SVM ℓ_p MKL and LSSVM ℓ_p MKL are the SVM-based MKL methods and least square SVM-based (LSSVM) MKL methods, respectively, which all extend the MKL with different norms. According to Sonnenburg et al. (2006), different kernel normalizations can have a significant impact on the performance of MKL. In our experiments, we adopt the popular ℓ_1, ℓ_2, ℓ_∞ norm to normalize MKL methods: SVM ℓ_1 MKL method (Lanckriet et al., 2004), SVM ℓ_2 MKL method (Kloft et al., 2008), SVM ℓ_∞ MKL method (Sonnenburg et al., 2006), LSSVM ℓ_1 MKL method (Suykens & Vandewalle, 1999), LSSVM ℓ_2 MKL method (Yu et al., 2010), and LSSVM ℓ_∞ MKL method (Ye et al., 2008). Different from SVM-based MKL methods, LPboost combines boosting approaches with MKL to mix different kernels. Here we adopt two versions of LPboost: LPboost- -β (Gehler & Nowozin, 2009) and LPboost-B (Gehler & Nowozin, 2009). In LPboost-β, a single-vector β is used to define a combination that works well for all classes, while in LPboost-B, each class has its own weight vector. The GP method (Kapoor, Grauman, Urtasun, &

Darrell, 2010) adopts the gaussian process method and pyramid match Kernel to combine multiple kernels to boost classification performance.

Besides the MKL methods, we also compare our method with multiview correlated algorithms: multiview CCA (Foster, Kakade, & Zhang, 2008), multirelational classification (Guo & Viktor, 2008), and intra-view and inter-view supervised correlation analysis for multiview feature learning (Jing et al., 2014). All of these methods exploit and use the correlations among all the views to improve task performance. Moreover, we compare our method with multiview manifold learning methods, namely manifold alignment preserving global Geomery (MAPGG) (Wang & Mahadevan, 2013).

In all the experiments, five-fold cross-validation is applied, and the average results with standard deviation are reported. The parameters in our proposed algorithm, denoted as α and β in equation 2.4, are tuned from $\{10^{-6}, 10^{-4}, 10^{-2}, 10^{0}, 10^{2}, 10^{4}, 10^{6}\}$. For SVM and MKL methods, we construct one gaussian kernel for each view of features,

$$\mathcal{K}\left(\mathbf{x}_i, \mathbf{x}_j\right) = \exp(-\gamma \|\mathbf{x}_i - \mathbf{x}_j\|^2), \tag{5.1}$$

where γ is also tuned in the same range as the proposed algorithm. The compared MKL methods are implemented using publicly available code (Yu et al., 2010). For LSSVM ℓ_p methods, following the work in Yu et al. (2010), we set the regularization parameter $\lambda = 1$ in the LSSVM ℓ_1 method, while in the LSSVM ℓ_∞ and ℓ_2 MKL methods, λ is estimated jointly as the kernel coefficient of an identity matrix. For LBPboost-β and http://qixianbiao.github.io. LBPboost-B methods, we use the publicly available code on the website of Yu et al. In all other SVM methods, we tune the regularization parameter, C, in the same range used for the proposed algorithm.

As we evaluate the effectiveness of our method for multiclass classification problems in our experiment, we employ mean average precision (MAP) to evaluate classification performance, which is the ranking performance computed under each label.

5.3 Experimental Results. We present the performance comparison of different algorithms in Table 2. It can be seen that the proposed algorithm MVCL consistently outperforms the other compared algorithms for different applications.

Our observations from the experimental results are as follows:

1. MVCL consistently outperforms single-view SVM classification and multiview feature learning algorithms, which indicates the effectiveness of the proposed feature learning algorithm in supervised multiview classification.
2. The approaches that use multiple views generally get better performance than SVM with each single type of features. From this

Table 2: Classification Results of Different Applications Measured by Mean Average Precision.

Method	NUS-WIDE-Object	Scene	MSRC-v1	Handwritten Digit
SVM (type 1)	0.161 ± 0.016	0.830 ± 0.018	0.786 ± 0.026	0.964 ± 0.023
SVM (type 2)	0.152 ± 0.018	0.743 ± 0.015	0.774 ± 0.022	0.764 ± 0.021
SVM (type 3)	0.144 ± 0.020	0.665 ± 0.017	0.794 ± 0.021	0.923 ± 0.018
SVM (type 4)	0.153 ± 0.019	0.581 ± 0.019	0.798 ± 0.019	0.958 ± 0.023
SVM (type 5)	0.142 ± 0.021	-	0.781 ± 0.018	0.798 ± 0.026
SVM (type 6)	0.147 ± 0.017	-	0.799 ± 0.025	-
SVM (all by concatenation)	0.187 ± 0.021	0.846 ± 0.014	0.802 ± 0.018	0.969 ± 0.022
SVM ℓ_∞ MKL method (Sonnenburg et al., 2006)	0.223 ± 0.019	0.852 ± 0.021	0.829 ± 0.021	0.975 ± 0.018
SVM ℓ_1 MKL method (Lanckriet et al., 2004)	0.215 ± 0.026	0.848 ± 0.017	0.824 ± 0.018	0.968 ± 0.023
SVM ℓ_2 MKL method (Kloft et al., 2008)	0.212 ± 0.024	0.847 ± 0.018	0.801 ± 0.022	0.966 ± 0.022
LSSVM ℓ_∞ MKL method (Ye et al., 2008)	0.211 ± 0.020	0.835 ± 0.021	0.795 ± 0.024	0.969 ± 0.020
LSSVM ℓ_1 MKL method (Suykens & Vandewalle, 1999)	0.198 ± 0.021	0.837 ± 0.019	0.812 ± 0.026	0.971 ± 0.019
LSSVM ℓ_2 MKL method (Yu et al., 2010)	0.192 ± 0.022	0.840 ± 0.014	0.819 ± 0.019	0.967 ± 0.022
GP method (Kapoor et al., 2010)	0.190 ± 0.019	0.835 ± 0.018	0.826 ± 0.017	0.969 ± 0.025
LPboost-β (Gehler & Nowozin, 2009)	0.229 ± 0.017	0.859 ± 0.021	0.818 ± 0.022	0.972 ± 0.017
LPboost-B (Gehler & Nowozin, 2009)	0.227 ± 0.014	0.861 ± 0.023	0.810 ± 0.022	0.970 ± 0.014
Multiview CCA (Foster et al., 2008)	0.236 ± 0.025	0.874 ± 0.028	0.832 ± 0.021	0.975 ± 0.023
Multirelational Classification (Guo & Viktor, 2008)	0.268 ± 0.022	0.894 ± 0.024	0.865 ± 0.013	0.987 ± 0.012
MAPGG (Wang & Mahadevan, 2013)	0.258 ± 0.024	0.917 ± 0.022	0.908 ± 0.019	0.993 ± 0.007
KI^{2}SCA (Jing et al., 2014)	0.274 ± 0.023	0.923 ± 0.022	0.902 ± 0.025	0.992 ± 0.011
MVCL (no shared subspace)	0.272 ± 0.021	0.919 ± 0.018	0.891 ± 0.017	0.990 ± 0.009
MVCL (shared subspace)	$\mathbf{0.299 \pm 0.021}$	$\mathbf{0.941 \pm 0.014}$	$\mathbf{0.932 \pm 0.019}$	$\mathbf{0.995 \pm 0.005}$

observation, we can confirm that multiview feature learning contributes to the classification performance improvement.

3. Both multiview CCA and multirelational classification generally perform better than the other multiview algorithms. This demonstrates that considering correlations among different views facilitates further classification.
4. The proposed algorithm outperforms the other compared algorithms. Moreover, when the shared component is considered, the proposed algorithm gets better performance. This indicates that it is beneficial to mine the shared component of different views.

5.4 Parameter Sensitivity and Convergence. In this section, experiments are conducted to study the performance variance with regard to the regularization parameters α and β and validate how fast the proposed algorithm monotonically decreases the objective function value until convergence.

We use the NUS-WIDE-Object data set for the experiments. The parameter sensitivity is shown in Figure 2. From this figure, we can observe that better performance is normally obtained when α and β are comparable. Figure 1 shows how fast algorithm 1 converges by fixing the regularization parameters α and β at 10^0, which is the median value of the tuned range of parameters. We learn that algorithm 1 converges within only 10 iterations, which is very efficient.

6 Conclusion

In this letter, we propose a novel multiview feature learning algorithm has been proposed to efficiently learn an intermediate representation of individual features and combine them for subsequent tasks (e.g., classification). Shared components among different views are taken into consideration. Our algorithm can mine the correlations among different views by incorporating G-norm into the proposed framework. To solve the objective function, we propose an effective and efficient algorithm to reach convergence in an iterative manner. We have tested and compared our algorithm with all other approaches over a number of real-world data sets. Our experimental results show that our algorithm is superior to the others compared in this letter.

We improve our framework based on the assumption that multiple views of features have some shared component. How to automatically determine the dimension of shared component is still an open problem. Normally, we manually estimate the shared component dimension. However, human estimations are sometimes unreliable. Even if humans think that two views of features have some shared components, there may be no common subspace at all. From the experiments conducted in this letter, we can see that performance will drop if we share subspace between nonshared

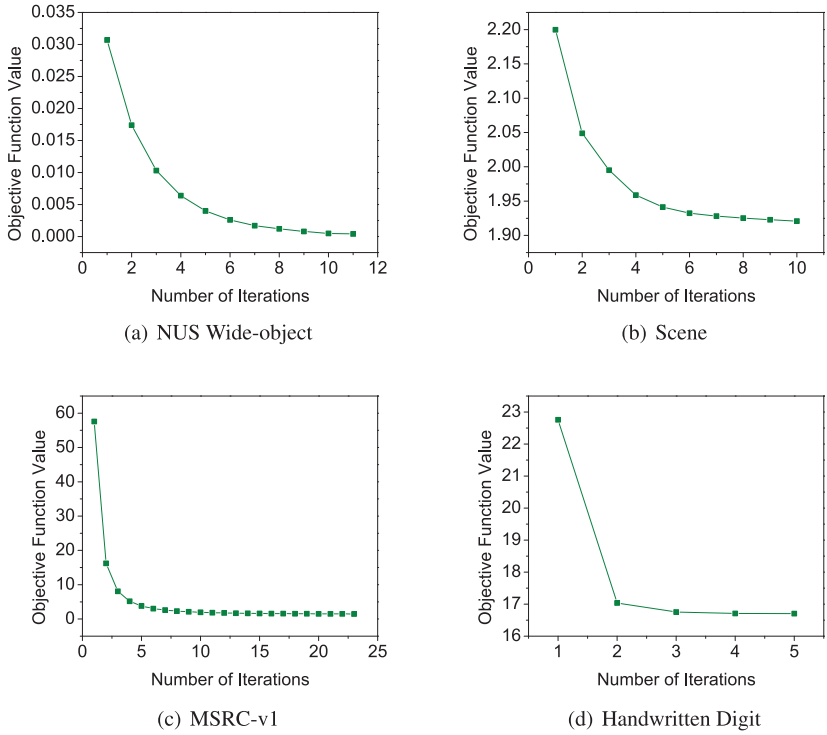

Figure 1: Convergence curves of the objective function value in equation 2.4 using algorithm 1. From this figure, we can observe that the proposed algorithm monotonically decreases the objective function value until convergence.

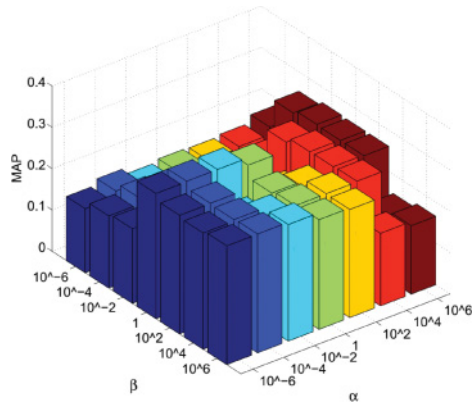

Figure 2: Parameter variance with regard to α and β. The figure shows different results when using different α and β.

views. Based on our discussion, we will learn how to automatically estimate the dimension of shared components among different views in our future work.

Acknowledgments

The research is supported by the Science Foundation of the China (Xi'an) Institute for Silk Road Research (2016SY10, 2016SY18) and the Research Foundation of XAUFE (15XCK14).

References

Bay, H., Tuytelaars, T., & Gool, L. V. (2006). Surf: Speeded up robust features. In *Proceedings of the 9th European Conference on Computer Vision*. Berlin: Springer.

Cai, X., Nie, F., & Huang, H. (2013). Multi-view k-means clustering on big data. In *Proceedings of the Conference on Artificial Intelligence*. Cambridge, MA: AAAI Press.

Chang, C.-C., & Lin, C.-J. (2011). LIBSVM: A library for support vector machines. *ACM TIST*, *2*(3), 27:1–27:27. http://www.csie.ntu.edu.tw/cjlin/libsvm.

Chang, X., Nie, F., Wang, S., Yang, Y., Zhou, X., & Zhang, C. (2016). Compound rank-k projections for bilinear analysis. *IEEE Trans. Neural Netw. Learning Syst.*, *27*(7), 1502–1513.

Chang, X., Nie, F., Yang, Y., & Huang, H. (2014). A convex formulation for semisupervised multi-label feature selection. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence* (pp. 1171–1177). Cambridge, MA: AAAI Press.

Chang, X., Nie, F., Yang, Y., Zhang, C., & Huang, H. (2016). Convex sparse PCA for unsupervised feature learning. *TKDD*, *11*(1), 3:1–3:16.

Chang, X., Yu, Y., Yang, Y., & Xing, E. P. (2016a). They are not equally reliable: Semantic event search using differentiated concept classifiers. In *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1884–1893). Piscataway, NJ: IEEE.

Chang, X., Yu, Y.-L., Yang, Y., & Xing, E. P. (2016b). Semantic pooling for complex event analysis in untrimmed videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. doi:10.1109/TPAMI.2016.2608901

Chen, H., Cai, X., Zhu, D., Nie, F., Liu, T., & Huang, H. (2012). Group-wise consistent parcellation of gyri via adaptive multi-view spectral clustering of fiber shapes. In *Proceedings of the Conference on Medical Computing and Computer-Assisted Intervention*. Berlin: Springer.

Chua, T.-S., Tang, J., Hong, R., Li, H., Luo, Z., & Zheng, Y.-T. (2009). Nus-wide: A real-world web image database from national university of singapore. In *Proceedings of the 8th International Conference on Image and Video Retrieval*. New York: ACM.

Conrad, C., & Mester, R. (2013). Learning multi-view correspondences via subspace-based temporal coincidences. In *Proceedings of the 18th Scandinavian Conference on Image Analysis*. Berlin: Springer.

Dalal, N., & Triggs, B. (2005). Histograms of oriented gradients for human detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* Piscataway, NJ: IEEE.

Fan, M., Chang, X., & Tao, D. (2017). Structure regularized unsupervised discriminant feature analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*. Cambridge, MA: AAAI Press.

Foster, D. P., Kakade, S. M., & Zhang, T. (2008). *Multi-view dimensionality reduction via canonical correlation analysis* (Technical Report). Chicago: Toyota Technological Institute.

Gehler, P., & Nowozin, S. (2009). On feature combination for multiclass object classification. In *Proceedings of the 12th International Conference on Computer Vision*. Piscataway, NJ: IEEE.

Grauman, K., & Darrell, T. (2006). Unsupervised learning of categories from sets of partially matching image features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* Piscataway, NJ: IEEE.

Guo, H., & Viktor, H. L. (2008). Multirelational classification: A multiple view approach. *Knowledge and Information Systems*, *17*(3), 287–312.

Ham, J., Lee, D., & Saul., L. (2005). Semisupervised alignment of manifolds. In *Proceedings of the 10th International Workshop on Artificial Intelligence and Statistics*. N.p.: Society for Artificial Intelligence and Statistics.

Jing, X., Hu, R., Zhu, Y., Wu, S., Liang, C., & Yang, J. (2014). Intra-view and inter-view supervised correlation analysis for multi-view feature learning. In *Proceedings of the AAAI Conference on Artificial Analysis*. Cambridge, MA: MIT Press.

Kapoor, A., Grauman, K., Urtasun, R., & Darrell, T. (2010). Gaussian processes for object categorization. *International Journal of Computer Vision*, *88*(2), 169–188.

Kloft, M., Brefeld, U., Laskov, P., & Sonnenburg, S. (2008). Non-sparse multiple kernel learning. In *Proceedings of the NIPS Workshop Kernel Learning: Automatic Selection of Kernels*.

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

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, *60*(2), 91–110.

Luo, M., Nie, F., Chang, X., Yang, Y., Hauptmann, A. G., & Zhang, Q. (2017). Avoiding optimal mean 2,1-norm maximization-based robust PCA for reconstruction. *Neural Computation*, *29*, 1124–1150.

Monadjemi, A., Thomas, B. T., & Mirmehdi, M. (2002). *Experiments on high resolution images towards outdoor scene classification* (Technical Report). Bristol: University of Bristol, Department of Computer Science.

Sonnenburg, S., Rätsch, G., Schäfer, C., & Schölkopf, B. (2006). Large scale multiple kernel learning. *Journal of Machine Learning Research*, *7*, 1531–1565.

Suykens, J. A., & Vandewalle, J. (1999). Least squares support vector machine classifiers. In *Neural Processing Letters*, *9*, 293–300.

Wang, C., & Mahadevan, S. (2013). Manifold alignment preserving global geometry. In *Proceedings of the 23rd International Joint Conference on Artificial Intelligence*. Cambridge, MA.

Wang, H., Nie, F., & Huang, H. (2013). Multi-view clustering and feature learning via structured sparsity. In *Proceedings of the International Conference on Machine Learning*.

Wang, S., Chang, X., Li, X., Long, G., Yao, L., & Sheng, Q. Z. (2016). Diagnosis code assignment using sparsity-based disease correlation embedding. *IEEE Trans. Knowl. Data Eng.*, *28*(12), 3191–3202.

Wang, S., Chang, X., Li, X., Sheng, Q. Z., & Chen, W. (2016). Multi-task support vector machines for feature selection with shared knowledge discovery. *Signal Processing*, *120*, 746–753.

Wang, S., Ma, Z., Yang, Y., Li, X., Pang, C., & Hauptmann, A. G. (2014). Semisupervised multiple feature analysis for action recognition. *IEEE Transactions on Multimedia*, *16*(2), 289–298.

Xue, X., Nie, F., Wang, S., Chang, X., Stantic, B., & Yao, M. (2017). Multi-view correlated feature learning by uncovering shared component. In *Proceedings of the Conference on Artificial Analysis*. Cambridge, MA: AAAI Press.

Yang, Y., Song, J., Huang, Z., Ma, Z., Sebe, N., & Hauptmann, A. (2012). Multi-feature fusion via hierarchical regression for multimedia analysis. *IEEE Transactions on Multimedia*, *15*, 572–581.

Ye, J., Ji, S., & Chen, J. (2008). Multi-class discriminant kernel learning via convex programming. *Journal of Machine Learning Research*, *9*, 719–758.

Yu, S., Falck, T., Daemen, A., Tranchevent, L.-C., Suykens, J. A., Moor, B. D., & Moreau, Y. (2010). L2-norm multiple kernel learning and its application to biomedical data fusion. *BMC Bioinformatics*, *11*, 309.

Received February 13, 2017; accepted February 26, 2017.