Information Sciences 648 (2023) 119570

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

Generalized multiview regression for feature extraction

Zhihui Lai [a,b,c], Yiling Lin [a], Jiacan Zheng [a], Jie Zhou [a,b,c], Heng Kong [d,*]

[a] College of Computer and Science Software Engineering, Shenzhen University, Shenzhen, 518060, China
[b] Guangdong Key Laboratory of Intelligent Information Processing, Shenzhen University, Shenzhen, 518060, China
[c] Guangdong Laboratory of Artificial - Intelligence and Cyber-Economic (SZ), Shenzhen University, Shenzhen, 518060, China
[d] Department of Thyroid and Breast Surgery, Baoan Central Hospital of Shenzhen, Shenzhen, 518102, China

ARTICLE INFO

Keywords:
Multiview learning
Feature extraction
Local preserving projection
Multiview classification

ABSTRACT

Multiview learning (MVL) has attracted considerable attention since an object can be observed from various views. To use the consistency of multiple views, canonical correlation analysis (CCA) is used as a basic technique for analyzing correlated subspaces. However, CCA-based methods are designed to extract the correlation information between each pair of views. In addition, these methods ignore view-specific geometric structures, which may provide effective complementary information. To address the limitations of CCA-based methods and improve the performance of multiview subspace learning methods, a novel MVL framework called generalized multiview regression (GMR) is proposed. It aims at finding a common subspace to preserve the complementary information of each view and maintain consistency among all the views. Specifically, to preserve the view-specific structures, GMR first considers data reconstruction and local geometrical structures. Subsequently, by introducing an orthogonal dictionary, GMR captures the discriminative consistency suitable for classification tasks. Finally, it uses $L_{2,1}$ as the basic norm to measure errors and regularization, which facilitates robustness and sparsity for feature extraction and selection. An iterative algorithm is designed to solve the proposed GMR. In addition, the convergence and complexity are analyzed theoretically. Extensive experiments on benchmark datasets are conducted to compare the GMR with state-of-the-art or some available multiview methods. The competitive performance implies that GMR is an effective multiview method for learning sparse projections and extracting discriminative and robust features.

1. Introduction

Machine learning has been applied to feature extraction [1,2], image forgery detection [3], face recognition [4], etc. However, most methods assume that the samples are from the same perspective. Actually, an object can be presented in different ways or collected from various channels (i.e., multiview data [5]) in real-world applications. For example, news events are reported by different media, and the news reports from various media are the multiple views on the events. Compared to the data from a single view, the multiview data of an object can provide a more informative and comprehensive description. Thus, multiview learning (MVL) has become a new research field in machine learning [6–10]. MVL has been applied to emotion recognition [11,12], clustering [13], action recognition [14], image classification [15,16], etc.

* Corresponding author.
E-mail addresses: lai_zhi_hui@163.com (Z. Lai), linyilinglyl@163.com (Y. Lin), jiacan_zheng@163.com (J. Zheng), jie_jpu@163.com (J. Zhou), generaldoc@126.com (H. Kong).

https://doi.org/10.1016/j.ins.2023.119570
Received 14 April 2023; Received in revised form 13 August 2023; Accepted 15 August 2023

Available online 22 August 2023

It is important to classify the samples from different perspectives [17]. However, in most cases, the samples (irrespective of whether they are from multiple views) are captured with high dimensionality and redundant information. And reducing dimensionality is necessary. Researchers have proposed quite a lot of dimensionality reduction methods (e.g., principal component analysis (PCA) [1], linear discriminant analysis (LDA) [18], and locality preserving projection (LPP) [19,20]) to obtain a linear transform for extracting features with low dimensionality. However, data from different views lie in different spaces with specific statistical properties and large discrepancies might exist among the various views. The features of the multiview data extracted using single-view methods are physically insignificant. Therefore, an increasing number of methods have been proposed to learn view-specific linear transforms. Each view contains specific statistical information and is complementary. In some MVL methods, the view consistency and complementary properties of each view are considered simultaneously [14–16,21].

To explore complementary information, several researchers have proposed MVL methods that extend manifold learning. Manifold learning assumes that data lies on a low-dimensional manifold embedded in a high-dimensional space. Learning the intrinsic local geometric structure is effective in improving the representativeness of the extracted features. LPP [19] is a classical manifold-learning-based method. Nevertheless, it can not be applied directly to multiview data because of view discrepancies. To obtain complementary information, multiview uncorrelated locality preserving projection (MULPP) [21], Grassmannian regularized structured multiview embedding [15], and group sparse multiview patch alignment framework (GSM-PAF) [16] preserve the local geometric structure by minimizing the sum of the local distances of all views. These methods maintain the local geometric structure of each view to explore complementary information. However, the latent reconstruction information of the original features is not considered.

For consistency among the various views, canonical correlation analysis (CCA) [2] is the most common method for two-view scenarios. This attempts to maximize the correlation between the projected samples of the two views. To address multiple (more than two) views, a few researchers have extended CCA to multiview CCA (MCCA) [22], which aims to maximize the sum of the correlations between each two views [22]. As indicated in [23], this strategy considers the statistics between each pair of views and ignores the high-order correlation. Therefore, tensor CCA (TCCA) [23] generalizes CCA to analyze the covariance tensor of multiple views by directly maximizing the canonical correlation of all views. However, TCCA incurs high space costs when the number of views is high.

Moreover, most of MVL methods use the Euclidean distance as the basic metric and have high sensitivity to outliers. To address this problem, a few researchers have adopted sparse norm to enhance robustness [24–27]. In addition, the original features contain redundant information and should be eliminated after feature extraction. To this end, GSM-PAF [16] introduced regularization term based on $L_{2,1}$ norm to ensure the sparsity of feature extraction [28–31].

In summary, both the consistency among multiple views and complementary information from different views should be considered. Moreover, most of MVL methods are sensitive to outliers and the redundant information should be eliminated. To address these problems, we propose a MVL method called generalized multiview regression (GMR) for feature extraction. It considers both the view consistency and complementary information of all views and attempts to obtain a linear transform for each view to project the samples into a common low-dimensional feature subspace. We design novel models to express the reconstruction and classification error, so as to improve the performance of multiview classification tasks. Moreover, the $L_{2,1}$ norm is adopted to simultaneously ensure robustness and sparsity. The main contributions of this work are summarized below:

1) A multiview method, GMR, is proposed for feature extraction, which considers view consistency and complementary information simultaneously. GMR designs a reconstruction term utilizing view-specific geometric structures to explore the complementary properties among all the views. Moreover, GMR integrates the label information to model the classification error. Therefore, view consistency can be achieved.
2) To enhance robustness, the GMR designs the reconstruction and classification error terms based on the definition of $L_{2,1}$ norm. In addition, $L_{2,1}$ norm is imposed on the regularization term to obtain sparse projections. Thus, the GMR can eliminate redundant information and extract discriminative and robust features.
3) An iterative algorithm to obtain the optimal solution is presented in subsection 4.1. In addition, the analyses of convergence and computational complexity are given. Extensive experiments are conducted on seven well-known datasets. The experimental results illustrate the competitive performance of the GMR.

The rest of this paper is organized as follows. Section 2 briefly introduces some related studies. Section 3 presents the objective function of the GMR. The iterative algorithm for the optimal solution and its theoretical analyses are given in Section 4. Extensive experiments are conducted to evaluate the performance of GMR and the results are shown in Section 5. Finally, the conclusions are summarized in Section 6.

2. Related work

This section introduces the notations and definitions first. Then we briefly introduce LPP and three multiview methods, MCCA [22], MULPP [21], and multiview discriminant analysis (MvDA) [10].

2.1. Notations and definitions

The italic letters denote scalars, i.e., *i, j, T*, etc. The lowercase letters in boldface represent vectors, i.e., $\mathbf{x}$, $\mathbf{y}$, $\mathbf{z}$ etc. And the bold uppercase letters indicate matrices, i.e., $\mathbf{X}$, $\mathbf{Y}$, $\mathbf{P}$, etc.

Given a matrix $\mathbf{H} \in \mathbb{R}^{a\times b}$, the L_m-norm of $\mathbf{H}$ is denoted as $\|\mathbf{H}\|_m$ and is defined as follows:

$$\|\mathbf{H}\|_m = \left(\sum_{i}^{a} \sum_{j}^{b} \left| h_{ij} \right|^m \right)^{\frac{1}{m}} \tag{1}$$

where h_{ij} is the element in the i-th row and j-th column.

Let $\mathbf{H}^i$ define the i-th row of $\mathbf{H}$. $\|\mathbf{H}\|_{2,1}$ represents $L_{2,1}$-norm of $\mathbf{H}$ and is defined as follows:

$$\|\mathbf{H}\|_{2,1} = \sum_{i}^{a} \left(\sum_{j}^{b} \left| h_{ij} \right|^2 \right)^{\frac{1}{2}} = \sum_{i}^{a} ||\mathbf{H}^i||_2 \tag{2}$$

In the multiview scenario, it is assumed that the samples are observed from c classes in v views. Let $\mathbf{X}^{(j)} = [\mathbf{x}_1^{(j)}, \mathbf{x}_2^{(j)}, ..., \mathbf{x}_{n_j}^{(j)}] \in \mathbb{R}^{d_j \times n_j}$ denote the samples from view j ($j = 1, 2, ..., v$). Here, d_j denotes the dimension of view j, and n_j is the number of samples from view j.

2.2. LPP

LPP [19] aims to obtain the projection matrix to preserve the locality of data. The objective function of LPP is formulated as follows:

$$\min_{\mathbf{P}} \frac{1}{2} \sum_{i,j}^{n} ||\mathbf{P}^T \mathbf{x}_i - \mathbf{P}^T \mathbf{x}_j||_2^2 \mathbf{W}_{ij} \tag{3}$$

where $\mathbf{P}$ is the projection matrix, $\mathbf{x}_i$ is the i-th samples, and $\mathbf{W}_{ij}$ indicates whether $\mathbf{x}_i$ is close to $\mathbf{x}_j$. The LPP solves the optimization problem with the constraint $\mathbf{P}^T\mathbf{X}\mathbf{Q}\mathbf{X}^T\mathbf{P} = \mathbf{1}$. Here, $\mathbf{Q}$ is a diagonal matrix and $\mathbf{Q}_{ii} = \sum_j \mathbf{W}_{ij}$. The solution of $\mathbf{P}$ in LPP is obtained by solving the following problem:

$$\begin{aligned} &\min_{\mathbf{P}} tr(\mathbf{P}^T\mathbf{X}\mathbf{L}\mathbf{X}^T\mathbf{P}) \\ &\text{s.t. } \mathbf{P}^T\mathbf{X}\mathbf{Q}\mathbf{X}^T\mathbf{P} = \mathbf{1} \end{aligned} \tag{4}$$

where the matrix $\mathbf{L} = \mathbf{Q} - \mathbf{W}$.

2.3. MCCA

MCCA finds a projection matrix for each view. Then the samples from different views are projected into a common space using the corresponding projections. In the common space, the total correlations of projected samples are maximized as follows:

$$\begin{aligned} &\max_{\mathbf{P}^{(1)},...,\mathbf{P}^{(v)}} \sum_{i<j} \mathbf{P}^{(i)T} \mathbf{C}_{ij} \mathbf{P}^{(j)} \\ &\text{s.t. } \mathbf{P}^{(i)T} \mathbf{C}_{ii} \mathbf{P}^{(i)} = 1, i = 1, 2, ..., v \end{aligned} \tag{5}$$

where $\mathbf{P}^{(j)}$ represents the projection matrix corresponding to j-th view, the covariance matrices $\mathbf{C}_{ij} = \mathbf{X}^{(i)}\mathbf{X}^{(j)T}$ and $\mathbf{C}_{ii} = \mathbf{X}^{(i)}\mathbf{X}^{(i)T}$. MCCA is an unsupervised method considering only the correlation of each pair of views.

2.4. MULPP

MULPP is a multiview method exploring complementary information and view consistency simultaneously. For the r-th projection of the i-th view $\mathbf{p}_r^{(i)}$, the objective function is designed as follows:

$$\begin{aligned} \max_{\mathbf{p}_r^{(1)},...,\mathbf{p}_r^{(v)}} \quad & \frac{1}{2} \sum_{i}^{v} \sum_{j}^{n} \sum_{k}^{n} ||\mathbf{p}_r^{(i)T} \mathbf{x}_{ij} - \mathbf{p}_r^{(i)T} \mathbf{x}_{ik}||_2^2 \mathbf{W}_{j,k}^{(i)} \\ & - \alpha(\mathbf{C}_{12...v} \times_1 \mathbf{p}_r^{(1)T} \times_2 \mathbf{p}_r^{(2)T} ... \times_v \mathbf{p}_r^{(v)T}) - \beta \sum_{i<j} (\mathbf{p}_r^{(i)T} \mathbf{C}_{ij} \mathbf{p}_r^{(j)})^2 \\ \text{s.t.} \quad & \mathbf{p}_r^{(i)T} \mathbf{C}_{ii} \mathbf{p}_r^{(i)} = 1, \mathbf{p}_l^{(i)T} \mathbf{C}_{ii} \mathbf{p}_r^{(i)} = 1, i = 1, 2, ..., v, l = 1, 2, ..., r-1. \end{aligned} \tag{6}$$

MULPP introduces two types of correlations for more flexible view consistency: the correlation of each two views and a high-order correlation. Moreover, it preserves the geometric structure of each view by applying LPP to explore complementary information.

2.5. MvDA

As a multiview extension of LDA, MvDA aims to maximize between-class variations and minimize within-class variations. The objective function is expressed as follows:

$$\min_{\mathbf{P}^{(1)},\ldots,\mathbf{P}^{(v)}} \frac{\sum_j^v \sum_i^c \sum_k^{n_{ij}} \left\| \mathbf{P}^{(j)T}\mathbf{x}_{i,k}^{(j)} - \mathbf{m}_i \right\|_2^2}{\sum_i^c \left\| \mathbf{m}_i - \mathbf{m} \right\|_2^2} \tag{7}$$

where $\mathbf{x}_{i,k}^{(j)}$ is the k-th sample of the i-th class from j-th view, n_{ij} is the number of samples from the i-th class in the j-th view, $\mathbf{m}_i$ denotes the mean of the projected samples of the i-th class from all views, and $\mathbf{m}$ is the mean of all the samples from all the views after projection. MvDA is a supervised method that uses Euclidean distance to characterize the relationship in the objective function.

3. Generalized multiview regression

This section provides the motivation of the proposed GMR method and then proposes the objective function.

3.1. Motivation

MVL has received increasing attention. Moreover, a few researchers have extended CCA to the multiview scenarios [22,23] to achieve view consistency. However, CCA-based methods have some limitations. First, these methods aim to maximize the correlation between each pair of views, but the correlation of all views is ignored. Second, the complementary information should be considered, and the local geometric structure is ignored in the CCA-based methods for MVL. These degrade the performance. Third, the projections learned by the CCA-based methods are not sparse because the projection matrix has no sparsity constraints nor sparse regularization. Fourth, most of the multiview methods have high sensitivity to outliers. Therefore, the development of a robust sparse MVL method considering view consistency and complementary information remains an open problem.

For the first problem, we integrate the label information to make the learned representations more discriminative and achieve view consistency. Maximizing the correlation among various views can achieve consistency, but there are some limitations. MCCA only considers the correlation between each pair of views, and TCCA incurs high space cost. Moreover, view consistency can be achieved by ensuring that the samples of different views from the same class are predicted to have similar results [32]. Thus, we introduce label information to express the classification error and explore the consistency among the different views.

To address the second problem, this study considers view-specific geometric structures. As shown in [33–35], because the high-dimensional data are embedded in a latent low-dimensional manifold, preserving the local geometric structure or local reconstruction relationship could improve feature extraction capability. Moreover, when the local geometric structure is used to characterize the multiview data with view information, it essentially contains the view-specific geometric structures and explores complementary information (as indicated in [21]). Therefore, the local geometric structure of each view is designed in the MVL method GMR and the second problem is solved.

Moreover, Nie et al. [28] demonstrated that adopting the $L_{2,1}$ norm on both the loss function and regularization term in single-view methods could select features and equip them with strong robustness. As indicated in [36], utilizing the $L_{2,1}$ norm as basic measurement of the loss function could improve the robustness of the MVL algorithms. Motivated by these methods, we introduce the $L_{2,1}$ norm as the basic measurement and the regularization term in the loss function to improve robustness and achieve joint sparsity [28] for multiview feature extraction.

To address the above four problems, a tractable strategy is to integrate view-specific geometric structures, label information, a robust measurement (i.e., $L_{2,1}$ norm), and the $L_{2,1}$ norm-based regularization term in a model to extract robust, sparse and representative features to improve the discriminative capability. The details are presented in the following subsections and the overview of GMR are show in Fig. 1.

3.2. Complementary information

For an object, the descriptions from multiple views are capable of characterizing it sufficiently because the information from multiple views is complementary. Thus, it is necessary to consider complementary information among different views. To this end, the geometric structure of the original data would be preserved. Suppose we have $\mathbf{W}^{(j)}$ that captures the geometrical structure of view j. Here, $\mathbf{W}_{k,r}^{(j)}$ indicates that $\mathbf{x}_k^{(j)}$ is close to $\mathbf{x}_r^{(j)}$. We attempt to determine a feature mapping $\phi^{(j)}$, so that we can map the original data $\mathbf{x}_r^{(j)}$ (the r-th sample from j-th view) in a low-dimensional feature space as $\mathbf{z}_r^{(j)}$ using $\mathbf{z}_r^{(j)} = \phi^{(j)}(\mathbf{x}_r^{(j)})$. To preserve the data structure of each view, the following optimization problem is solved:

$$\min \sum_j^v \sum_{k,r}^{n_j} \|\mathbf{z}_k^{(j)} - \mathbf{z}_r^{(j)}\|_2 \mathbf{W}_{k,r}^{(j)} \tag{8}$$

Minimizing Eq. (8) preserves the structure of each view in the feature space. Compared with the objective function in (3), this term circumvents the square operation so that the influence of outliers is weakened.

However, $\mathbf{z}_r^{(j)}$ learned by (8) only preserves the locality but loses the latent reconstruction information of the original features, which is important for MVL. That is to say, although the locality is preserved in the low-dimensional subspace, the extracted features are not suitable to reconstruct the original data, which degrades the representation ability. Thus, we design a reconstruction term to reconstruct the original data $\mathbf{x}_r^{(j)}$ using the projected samples. This renders the features more informative. Then, we minimize the following loss:

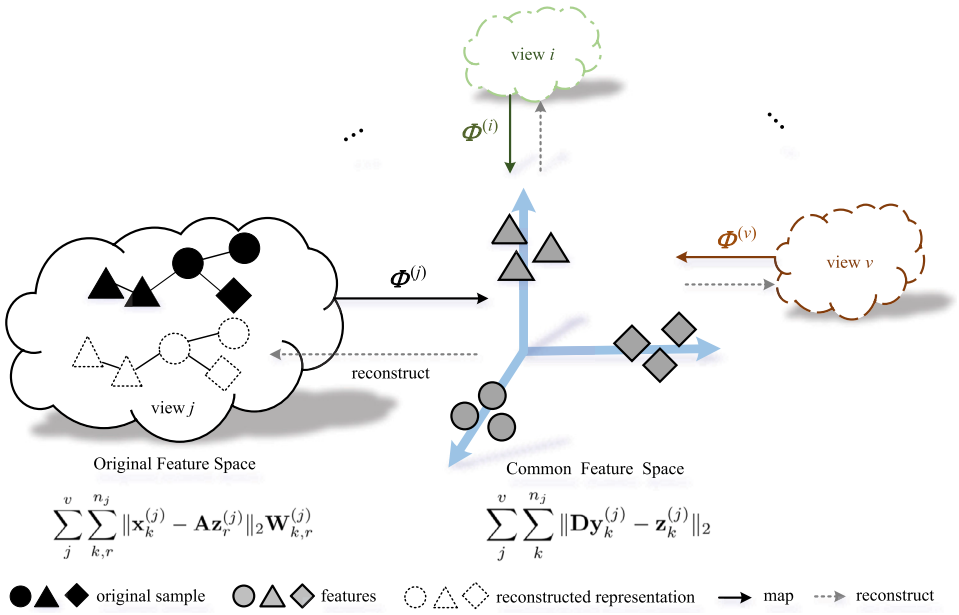


Fig. 1. Overview of GMR. The different shapes denote the samples from different classes. The original samples and extracted features are denoted as solid shapes. The original samples are outlined with solid line in the same color, and the extracted features are outlined in different colors. The hollow shape outlined by the dotted line indicates the reconstructed representations. From the illustration, GMR transforms the original samples into the common feature space, in which the representations are discriminative. Moreover, the learned representations are required to reconstruct the original data $\mathbf{X}^{(j)}$ effectively. In the reconstruction process, the view-specific local structures are preserved.

$$\begin{aligned}&\min\ \widetilde{\mathcal{L}}_1=\sum_{j}^{v}\sum_{k,r}^{n_j}\|\mathbf{x}_k^{(j)}-\mathbf{A}^{(j)}\mathbf{z}_r^{(j)}\|_2\mathbf{W}_{k,r}^{(j)}\\&\text{s.t. }\mathbf{A}^T\mathbf{A}=\mathbf{I},\end{aligned}\tag{9}$$

where $\mathbf{A}^{(j)}$ is an orthogonal matrix. (See Fig. 1.)

However, reconstructing the original data $\mathbf{X}^{(j)}$ using a specific $\mathbf{A}^{(j)}$ would cause a severe unaligned problem. Suppose $\{\mathbf{A}^{(j)*},\mathbf{Z}^{(j)*}\}$ is the optimal solution of (9), we can conveniently verify that $\{\mathbf{A}^{(j)*}\mathbf{R}^{(j)^T},\mathbf{R}^{(j)}\mathbf{Z}^{(j)*}\}$ remains the optimal solution, where $\mathbf{R}^{(j)}$ is an orthogonal matrix. This implies the learned representation $\mathbf{Z}^{(j)}$ of the j-th view is invariant under an arbitrary rotation $\mathbf{R}^{(j)}$. Thus, it is necessary to consider a common basis $\mathbf{A}$ for all views. Finally, we get the following loss:

$$\begin{aligned}&\min\ \mathcal{L}_1=\sum_{j}^{v}\sum_{k,r}^{n_j}\|\mathbf{x}_k^{(j)}-\mathbf{A}\mathbf{z}_r^{(j)}\|_2\mathbf{W}_{k,r}^{(j)}\\&\text{s.t. }\mathbf{A}^T\mathbf{A}=\mathbf{I}.\end{aligned}\tag{10}$$

Eq. (10) differs from Eq. (9) by forcing the basis of each view to be the same and using the common basis $\mathbf{A}$ for all views.

3.3. Consistency information

View consistency has been considered in several multiview methods. CCA-based techniques are developed to achieve the goal. However, CCA-based methods have some limitations. View consistency can be achieved by ensuring that the samples from the same class are predicted to be identical. Thus, label indicators are introduced to explicitly model classification errors and ensure view consistency. The optimal scoring criterion is a method for regressing the label indicator onto the features in a low-dimensional space [37,38].

Inspired by this, we aim to map the learned representations of different classes into different orthogonal subspaces. To this end, we first design an orthogonal dictionary, denoted by $\mathbf{D}$. Suppose that $\mathbf{y}_k^{(j)}$ is the one-hot vector for the k-th samples in j-th view. We minimize the following regression term:

$$\min\ \mathcal{L}_2=\sum_{j}^{v}\sum_{k}^{n_j}\|\mathbf{D}\mathbf{y}_k^{(j)}-\mathbf{z}_k^{(j)}\|_2,\quad\text{s.t. }\mathbf{D}^T\mathbf{D}=\mathbf{I}.\tag{11}$$

Therefore, the extracted feature $\mathbf{z}_k^{(j)}$ is discriminative in the common space. Moreover, GMR explores the view consistency among different views as well.

Algorithm 1 GMR Algorithm.

Input:
training data $\mathbf{X}^{(j)} = [\mathbf{x}_1^{(j)}, ..., \mathbf{x}_{n_j}^{(j)}]$, label matrix $\mathbf{Y}^{(j)} = [\mathbf{y}_1^{(j)}, ..., \mathbf{y}_{n_j}^{(j)}]$, affinity matrix $\mathbf{W}^{(j)}$, $(j = 1,...,v)$, balance parameters β, γ, maximal number of iterations ite_{max}.
Output:
All projection matrices $\mathbf{P} = [\mathbf{P}^{(1)T}, ..., \mathbf{P}^{(v)T}]^T$.

Initialize $\mathbf{P}^{(j)}, \mathbf{A}, \mathbf{D}$ randomly, $j = 1, ..., v$.
Construct matrices $\boldsymbol{Z}$, $\mathbf{D}_g$, $\mathbf{D}_{rr}$, $\mathbf{X}_{diag}$, $\mathbf{X}$.
for $i = 1 : ite_{max}$
  Update $\mathbf{A}$ using $\mathbf{A} = \mathbf{U}\mathbf{V}^T$ in Eq. (21).
  Update $\mathbf{D}$ using $\mathbf{D} = \check{\mathbf{U}}\check{\mathbf{V}}^T$ in Eq. (24).
  Update $\mathbf{P}$ by Eq. (25).
  Update $\mathbf{D}_{rr}$ using $(\mathbf{D}_{rr})_{ii} = 1/\|(\mathbf{Y}^T\mathbf{D}^T - \mathbf{X}_{diag}^T\mathbf{P})^i\|_2$.
  Update $\mathbf{G}^{(j)}$ using $\mathbf{G}_{k,r}^{(j)} = \mathbf{W}_{k,r}^{(j)}/\|\mathbf{x}_k^{(j)T} - \mathbf{x}_r^{(j)T}\mathbf{P}^{(j)}\mathbf{A}^T\|_2$.
  Update $\mathbf{D}_g^{(j)} = diag(\sum_k^{n_j} \mathbf{G}_{1,k}^{(j)}, ..., \sum_k^{n_j} \mathbf{G}_{n_j,k}^{(j)})$.
  Construct $\mathbf{G} = diag(\mathbf{G}^{(1)}, ..., \mathbf{G}^{(v)})$, $\mathbf{D}_g = diag(\mathbf{D}_g^{(1)}, ..., \mathbf{D}_g^{(v)})$
  if the value of objective function converges
    break
  end
Return all view-specific projections $\mathbf{P}^{(j)} (j = 1, ..., v)$.

3.4. Generalized multiview regression

By integrating $\mathcal{L}_1$ and $\mathcal{L}_2$ into one framework, we derive the following formulation:

$$\begin{aligned} &\min\ \mathcal{L}_1 + \beta\mathcal{L}_2 + \gamma\mathcal{R} \\ &\text{s.t. } \mathbf{A}^T\mathbf{A} = \mathbf{I},\ \mathbf{D}^T\mathbf{D} = \mathbf{I} \end{aligned} \tag{12}$$

where $\mathcal{R}$ indicates the regularization term.

To demonstrate the effectiveness of the framework above by constraining the feature maps as linear projections, we design a regression method called GMR. The feature maps $\phi^{(j)}(j = 1, ..., v)$ can be expressed explicitly in matrix form as $\mathbf{z}_i^{(j)} = \phi^{(j)}(\mathbf{x}_i^{(j)}) = \mathbf{P}^{(j)T}\mathbf{x}_i^{(j)}$. Moreover, the third part $\mathcal{R}$ is assigned as the $L_{2,1}$ norm-based regularization term. Then the proposed GMR is formulated as follows:

$$\begin{aligned} \min_{\mathbf{A},\mathbf{D},\mathbf{P}^{(j)}} &\sum_j^v \sum_{k,r}^{n_j} \|\mathbf{x}_k^{(j)} - \mathbf{A}\mathbf{P}^{(j)T}\mathbf{x}_r^{(j)}\|_2 \mathbf{W}_{k,r}^{(j)} + \beta \sum_j^v \sum_k^{n_j} \|\mathbf{D}\mathbf{y}_k^{(j)} - \mathbf{P}^{(j)T}\mathbf{x}_k^{(j)}\|_2 \\ &+ \gamma \sum_j^v \|\mathbf{P}^{(j)}\|_{2,1} \\ \text{s.t. } &\mathbf{A}^T\mathbf{A} = \mathbf{I},\ \mathbf{D}^T\mathbf{D} = \mathbf{I}. \end{aligned} \tag{13}$$

In (13), GMR adopts $L_{2,1}$ norm to measure the loss function. From Eq. (2), the square operation is circumvented and the influence of outliers is weakened so that the extracted features are more robust. Moreover, GMR introduces the $L_{2,1}$ norm-based regularization on the projection matrices to prevent the over-fitting problem and improve the explanations of the projected features. Compared with the widely used $\|\mathbf{P}^{(j)}\|_F^2$, minimizing $\|\mathbf{P}^{(j)}\|_{2,1}$ tends to force the row of $\mathbf{P}^{(j)}$ to be zero. Thus, the proposed method derives row-sparse projection matrices that they can filter out some uncorrelated features and select the most important features for the recognition task. In addition, the $L_{2,1}$ norm introduces the diagonal weight matrix, which is computed by the corresponding variables. To obtain the optimal solution, we design an iterative algorithm and the details can be seen in the following.

4. Optimization and theoretical analysis

In this section, an iterative algorithm for obtaining the optimal solution of the proposed model is presented. Then the theoretical analyses of the convergence and computational complexity are presented.

4.1. Optimal solution

It can be seen from Eq. (13) that the square operation is circumvented. Using the technique proposed in [28], we can rewrite the first term as:

$$\sum_j^v \sum_{k,r}^{n_j} \left\|\mathbf{x}_k^{(j)} - \mathbf{A}\mathbf{P}^{(j)T}\mathbf{x}_r^{(j)}\right\|_2 \mathbf{W}_{k,r}^{(j)} = \sum_j^v \sum_{k,r}^{n_j} \left\|\mathbf{x}_k^{(j)T} - \mathbf{x}_r^{(j)T}\mathbf{P}^{(j)}\mathbf{A}^T\right\|_2^2 \mathbf{G}_{k,r}^{(j)} \tag{14}$$

where $\mathbf{G}_{k,r}^{(j)} = \mathbf{W}_{k,r}^{(j)}/\|\mathbf{x}_k^{(j)T} - \mathbf{x}_r^{(j)T}\mathbf{P}^{(j)}\mathbf{A}^T\|_2$.

Furthermore, for Eq. (14), we can have:

$$\begin{aligned}&\sum_{j}^{v}\sum_{k,r}^{n_j}\left\|\mathbf{x}_k^{(j)T}-\mathbf{x}_r^{(j)T}\mathbf{P}^{(j)}\mathbf{A}^T\right\|_2^2\mathbf{G}_{k,r}^{(j)}\\&=\sum_{j}^{v}tr(\mathbf{X}^{(j)}\mathbf{D}_g^{(j)}\mathbf{X}^{(j)^T}-2\mathbf{X}^{(j)}\mathbf{G}^{(j)}\mathbf{X}^{(j)^T}\mathbf{P}^{(j)}\mathbf{A}^T+\mathbf{A}\mathbf{P}^{(j)^T}\mathbf{X}^{(j)}\mathbf{D}_g^{(j)}\mathbf{X}^{(j)^T}\mathbf{P}^{(j)}\mathbf{A}^T)\\&=tr(\mathbf{X}\mathbf{D}_g\mathbf{X}^T-2\mathbf{X}\mathbf{G}\mathbf{X}_{diag}^T\mathbf{P}\mathbf{A}^T+\mathbf{P}^T\mathbf{X}_{diag}\mathbf{D}_g\mathbf{X}_{diag}^T\mathbf{P})\end{aligned}\tag{15}$$

where $\mathbf{G}^{(j)}$ is a matrix whose element in the k-th row and the r-th column is $\mathbf{G}_{k,r}^{(j)}$, $\mathbf{D}_g^{(j)}=diag(\sum_r^{n_j}\mathbf{G}_{1,r}^{(j)},...,\sum_r^{n_j}\mathbf{G}_{n_j,r}^{(j)})$, $\mathbf{X}=[\mathbf{X}^{(1)},...,\mathbf{X}^{(v)}]$, $\mathbf{X}_{diag}=diag(\mathbf{X}^{(1)},...,\mathbf{X}^{(v)})$, $\mathbf{G}=diag(\mathbf{G}^{(1)},...,\mathbf{G}^{(v)})$, $\mathbf{D}_g=diag(\mathbf{D}_g^{(1)},...,\mathbf{D}_g^{(v)})$ and $\mathbf{P}=[\mathbf{P}^{(1)T},...,\mathbf{P}^{(v)T}]^T$.

In Eq. (13), the second part can be reformulated as:

$$\begin{aligned}&\sum_{j}^{v}\sum_{k}^{n_j}\left\|\mathbf{D}\mathbf{y}_k^{(j)}-\mathbf{P}^{(j)T}\mathbf{x}_k^{(j)}\right\|_2\\&=\sum_{j}^{v}\sum_{k}^{n_j}\left\|\mathbf{y}_k^{(j)T}\mathbf{D}^T-\mathbf{x}_k^{(j)T}\mathbf{P}^{(j)}\right\|_2\\&=\left\|\begin{array}{c}\mathbf{Y}^{(1)T}\mathbf{D}^T-\mathbf{X}^{(1)T}\mathbf{P}^{(1)}\\\mathbf{Y}^{(2)T}\mathbf{D}^T-\mathbf{X}^{(2)T}\mathbf{P}^{(2)}\\\cdots\\\mathbf{Y}^{(v)T}\mathbf{D}^T-\mathbf{X}^{(v)T}\mathbf{P}^{(v)}\end{array}\right\|_{2,1}\\&=\left\|\mathbf{Y}^T\mathbf{D}^T-\mathbf{X}_{diag}^T\mathbf{P}\right\|_{2,1}\end{aligned}\tag{16}$$

where $\mathbf{Y}=[\mathbf{Y}^{(1)},...,\mathbf{Y}^{(v)}]$, and $\mathbf{Y}^{(j)}=[\mathbf{y}_1^{(j)},...,\mathbf{y}_{n_j}^{(j)}]$. According to the definition of $L_{2,1}$-norm, the above formulation becomes:

$$\begin{aligned}&\left\|\mathbf{Y}^T\mathbf{D}^T-\mathbf{X}_{diag}^T\mathbf{P}\right\|_{2,1}\\&=tr(\mathbf{Y}\mathbf{D}_{rr}\mathbf{Y}^T-2\mathbf{P}^T\mathbf{X}_{diag}\mathbf{D}_{rr}\mathbf{Y}^T\mathbf{D}^T+\mathbf{P}^T\mathbf{X}_{diag}\mathbf{D}_{rr}\mathbf{X}_{diag}^T\mathbf{P})\end{aligned}\tag{17}$$

where $\mathbf{D}_{rr}$ is a diagonal matrix whose i-th diagonal element equal to $(\mathbf{D}_{rr})_{ii}=1/\left\|\mathbf{Y}^T\mathbf{D}^T-\mathbf{X}_{diag}^T\mathbf{P}\right\|_2$ and $(\mathbf{Y}^T\mathbf{D}^T-\mathbf{X}_{diag}^T\mathbf{P})^i$ indicates the i-th row of $(\mathbf{Y}^T\mathbf{D}^T-\mathbf{X}_{diag}^T\mathbf{P})$.

With Eqs. (15) and (17), the Eq. (13) can be reformulated as follows:

$$\begin{aligned}\min_{\mathbf{P},\mathbf{A},\mathbf{D}}\quad&tr(-2\mathbf{X}\mathbf{G}\mathbf{X}_{diag}^T\mathbf{P}\mathbf{A}^T+\mathbf{P}^T\mathbf{X}_{diag}\mathbf{D}_g\mathbf{X}_{diag}^T\mathbf{P}+\mathbf{X}\mathbf{D}_g\mathbf{X}^T)\\&+\beta tr(-2\mathbf{P}^T\mathbf{X}_{diag}\mathbf{D}_{rr}\mathbf{Y}^T\mathbf{D}^T+\mathbf{P}^T\mathbf{X}_{diag}\mathbf{D}_{rr}\mathbf{X}_{diag}^T\mathbf{P}+\mathbf{D}\mathbf{Y}\mathbf{D}_{rr}\mathbf{Y}^T\mathbf{D}^T)\\&+\gamma tr(\mathbf{P}^T\mathbf{D}_p\mathbf{P})\\&\text{s.t. }\mathbf{A}^T\mathbf{A}=\mathbf{I},\mathbf{D}^T\mathbf{D}=\mathbf{I}\end{aligned}\tag{18}$$

where $\mathbf{D}_p$ is a diagonal matrix with i-th diagonal element $\mathbf{D}_{pii}=1/\left\|\mathbf{P}^i\right\|_2$, and $\mathbf{P}^i$ is the i-th row of the matrix $\mathbf{P}$.

We propose an iterative algorithm to obtain the optimal solutions of $\mathbf{A}$, $\mathbf{P}$ and $\mathbf{D}$. The details of each iteration are presented below:

For the A-subproblem: The optimal solution of $\mathbf{A}$ can be obtained by solving the following optimization problem:

$$\begin{aligned}&\max_{\mathbf{A}}tr(\mathbf{X}\mathbf{G}\mathbf{X}_{diag}^T\mathbf{P}\mathbf{A}^T)\\&\text{s.t. }\mathbf{A}^T\mathbf{A}=\mathbf{I}\end{aligned}\tag{19}$$

From the theory in [39], we compute the optimal $\mathbf{A}$ by SVD of $\mathbf{X}\mathbf{G}\mathbf{X}_{diag}^T\mathbf{P}$:

$$\mathbf{X}\mathbf{G}\mathbf{X}_{diag}^T\mathbf{P}=\mathbf{U}\Sigma\mathbf{V}^T\tag{20}$$

And we have:

$$\mathbf{A}=\mathbf{U}\mathbf{V}^T\tag{21}$$

For the D-subproblem: The optimal $\mathbf{D}$ can be computed by solving the following optimization problem:

$$\begin{aligned}&\max_{\mathbf{D}}tr(\mathbf{P}^T\mathbf{X}_{diag}\mathbf{D}_{rr}\mathbf{Y}^T\mathbf{D}^T)\\&\text{s.t. }\mathbf{D}^T\mathbf{D}=\mathbf{I}\end{aligned}\tag{22}$$

Similarly, the optimal solution of $\mathbf{D}$ is given by SVD of $\mathbf{P}^T\mathbf{X}_{diag}\mathbf{D}_{rr}\mathbf{Y}^T$:

$$\mathbf{P}^T\mathbf{X}_{diag}\mathbf{D}_{rr}\mathbf{Y}^T=\breve{\mathbf{U}}\Psi\breve{\mathbf{V}}^T\tag{23}$$

Then:

$$\mathbf{D} = \check{\mathbf{U}}\check{\mathbf{V}}^T \tag{24}$$

For the P-subproblem: The derivative of $\mathbf{P}$ is set to zero. Subsequently, we obtain the following equation:

$$\begin{aligned}\mathbf{P} = &(\mathbf{X}_{diag}\mathbf{D}_g\mathbf{X}_{diag}^T + \beta\mathbf{X}_{diag}\mathbf{D}_{rr}\mathbf{X}_{diag}^T + \gamma\mathbf{D}_p)^{-1} \\ &(\mathbf{X}_{diag}\mathbf{G}\mathbf{X}^T\mathbf{A} + \beta\mathbf{X}_{diag}\mathbf{D}_{rr}\mathbf{Y}^T\mathbf{D}^T)\end{aligned} \tag{25}$$

The details of GMR are presented in the Algorithm 1.

4.2. Convergence

We denote $(\cdot)_{(i)}$ as the variable obtained in the i-th iteration. To prove the convergence of GMR, we utilize the following lemmas:

Lemma 1 *([28]). Given two random non-zero constants ρ and σ, the following inequality holds:*

$$\sqrt{\rho} - \frac{\rho}{2\sqrt{\sigma}} \leq \sqrt{\sigma} - \frac{\sigma}{2\sqrt{\sigma}} \tag{26}$$

Lemma 2 *([28]). Given a random non-zero matrix $\mathbf{U}$, the following inequality holds:*

$$\sum_i \left\|\mathbf{U}_{(\tau)}^i\right\|_2 - \sum_i \frac{\|\mathbf{U}_{(\tau)}^i\|_2^2}{2\|\mathbf{U}_{(\tau-1)}^i\|_2} \leq \sum_i \|\mathbf{U}_{(\tau-1)}^i\|_2 - \sum_i \frac{\left\|\mathbf{U}_{(\tau-1)}^i\right\|_2^2}{2\left\|\mathbf{U}_{(\tau-1)}^i\right\|_2} \tag{27}$$

Theorem 1. *The objective function of GMR will monotonously decrease in each iteration until it converges and a local optimal solution can be obtained.*

Proof. The proof is shown in Appendix.

4.3. Computational complexity

We define the sum of the dimensions of all the views as $d = \sum_j^v d_j$. The highest computational complexity in each iteration of the SVD in Eqs. (20) and (23) are $O(d^3)$ in the worst-case scenario. Assuming that the value of the objective function converges after T iterations, the computational cost is $O(2Td^3)$. From the above analysis, the dimensions determine the computational complexity. Therefore, dimensionality reduction should be performed during data pre-processing. The experiments in Section 5 demonstrate that the GMR converges in a few iterations.

5. Experiments

This section describes the datasets used in the experiments first. Subsequently, the performance of GMR is evaluated and compared with other multiview methods. Finally, we list the experimental analyses of GMR from different perspectives.

A series of experiments were performed on the CMU PIE [40], 100Leaves [41], AR face [42], NUS-WIDE [43], COIL-100, 3Sources dataset, and HUMBI [44].

The **CMU PIE [40]** face dataset contains 41,368 face images of 68 objects at 7 poses with changing illumination and expression. Based on the varying poses, the samples are regarded as the data from 7 views. The subset of the CMU PIE dataset used in our experiments contains samples from 60 objects.

The **AR [42]** face dataset is comprised of 126 objects, 4,000 images in total. The samples vary in lighting, facial occlusion, and expression. In experiments on this dataset, the number of views is 3 based on the facial occlusion (i.e., with glasses, with a scarf, and no occlusion). Images of 100 individuals were selected for the experiments.

The **100Leaves [41]** dataset collects 100 plant species. Each one has 16 specimens represented by 3 features containing various types of information (i.e., shape, texture, and margin). That is, there are 1600 samples in one feature set. Each type of feature provides a 64-element vector per sample.

The **NUS-WIDE [43]** dataset extracts 6 types of low-level features from each real-world image. The images of 10 concepts (i.e., airport, bear, cat, dancing, elk, fire, glacier, harbor, leaf, and plane) were selected and employed in our experiments. Moreover, the subset consisted of 4 types of features (500-D bag of words, 225-D block-wise color moments, 144-D color correlogram, and 128-D wavelet texture) for each image. Based on the type of feature, the samples can be viewed from 4 perspectives.

The **COIL-100** dataset contains 7,200 images of 100 objects. The samples are divided into 3 groups based on the camera angle (i.e., three groups containing the photographs taken from $0° - 115°$, $120° - 235°$, $240° - 355°$, respectively).

The **3Sources** dataset is a text dataset containing 418 news stories. Furthermore, 169 stories are reported in three news sources (i.e., BBC, *The Guardian*, and Reuters), which are from 6 categories.

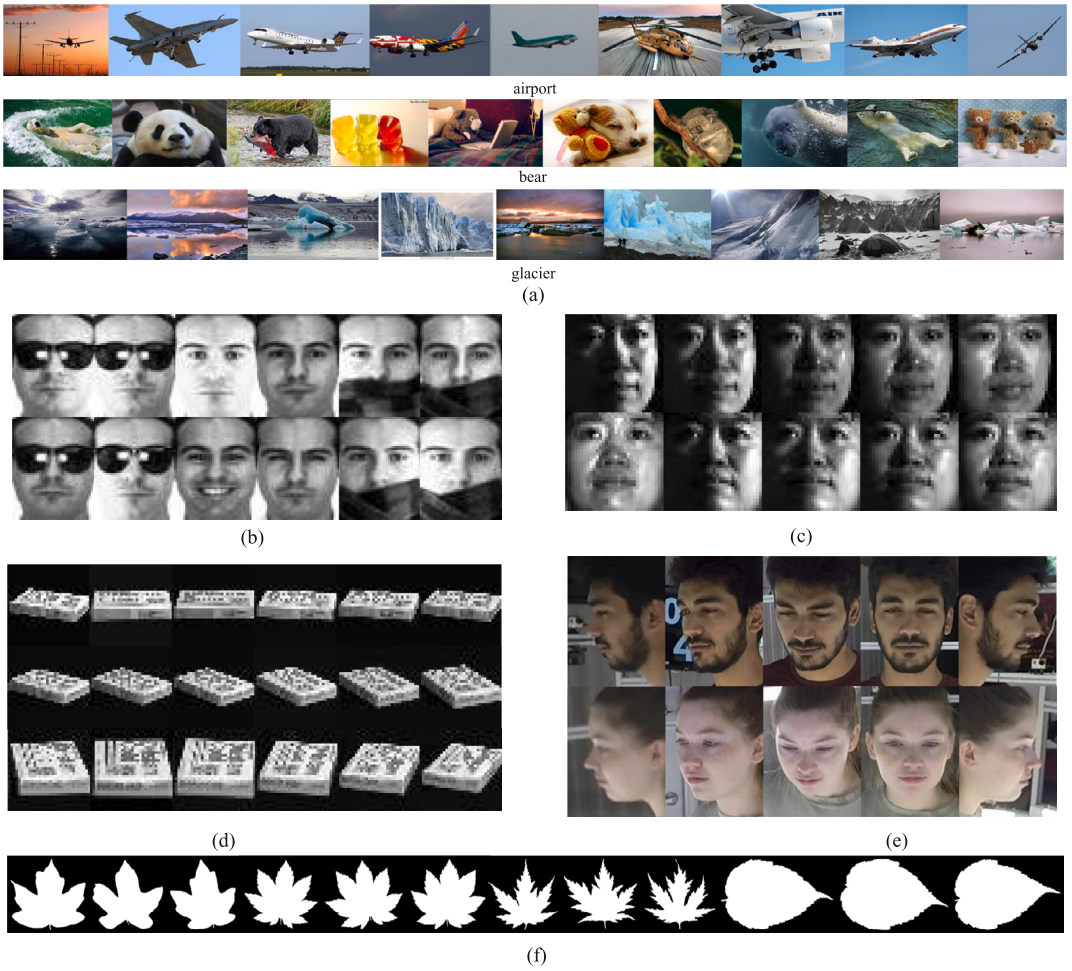


Fig. 2. The samples from (a) NUS-WIDE, (b) AR face, (c) CMU PIE, (d) COIL-100, (e) HUMBI, (f) 100Leaves.

Table 1
The statistical information of the datasets in experiments.

Dataset	CMU PIE [1]	AR[2]	100Leaves[3]	NUS-WIDE[4]	COIL100 [5]	3Sources[6]	HUMBI[7]
instances	8820	1800	4800	8000	7200	507	32000
instances per view	1260	600	1600	2000	2400	169	6400
views	7	3	3	4	3	3	5
classes	60	100	100	10	100	6	64

[1] http://www.cs.cmu.edu/afs/cs/project/PIE/MultiPie/Multi-Pie/Home.html.
[2] http://www2.ece.ohio-state.edu/~aleix/ARdatabase.
[3] https://archive.ics.uci.edu/ml/datasets/One-hundred+plant+species+leaves+data+set.
[4] https://lms.comp.nus.edu.sg/wp-content/uploads/2019/research/nuswide/NUS-WIDE.html.
[5] https://www.cs.columbia.edu/CAVE/software/softlib/coil-100.php.
[6] http://mlg.ucd.ie/datasets/3sources.html.
[7] https://humbi-data.net/.

The **HUMBI [44]** dataset is a multiview dataset that uses 107 synchronized cameras to capture human body expressions from 772 subjects. A subset containing the facial expressions of the first 64 subjects captured with 5 randomly selected cameras was used in experiments. Each subject contained 100 images under one camera with changing facial expressions.

The statistical information of the datasets is presented in Table 1 and the samples of the image datasets are shown in Fig. 2 (a)~(f).

5.1. Experiment setting

In our experiments, the proposed GMR is compared with existing multiview feature extraction methods including MvDA [10], GMA [45], MCCA [22], cross-regression for multi-view feature extraction (CRMvFE) [36], robust cross-regression for multi-view fea-

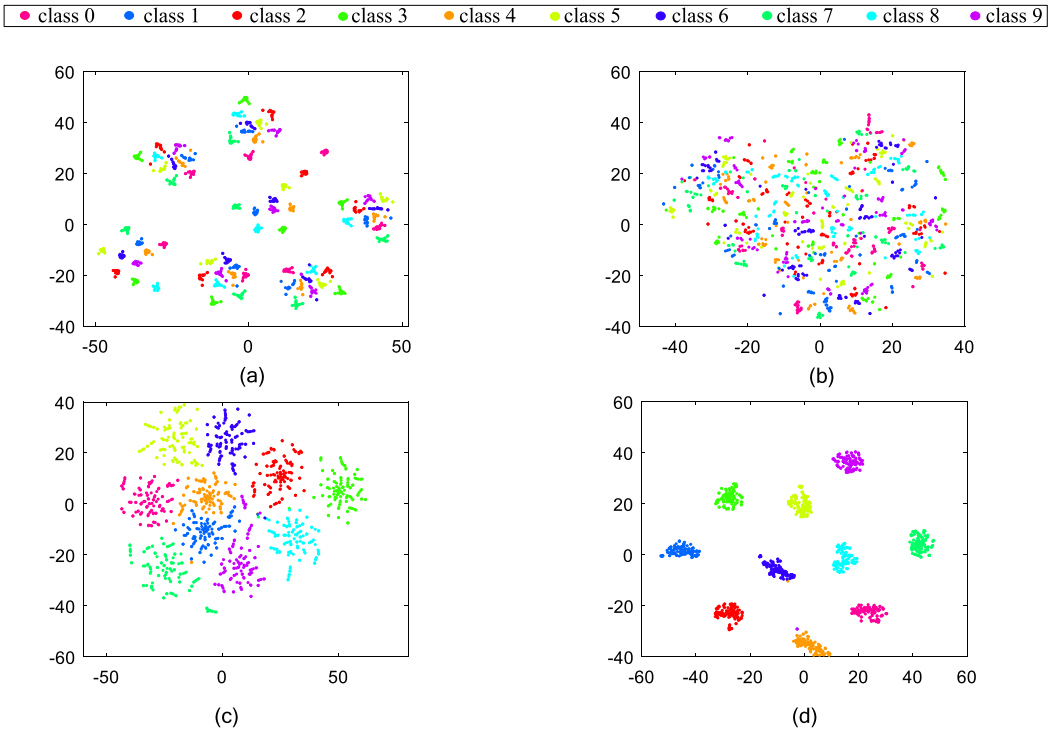


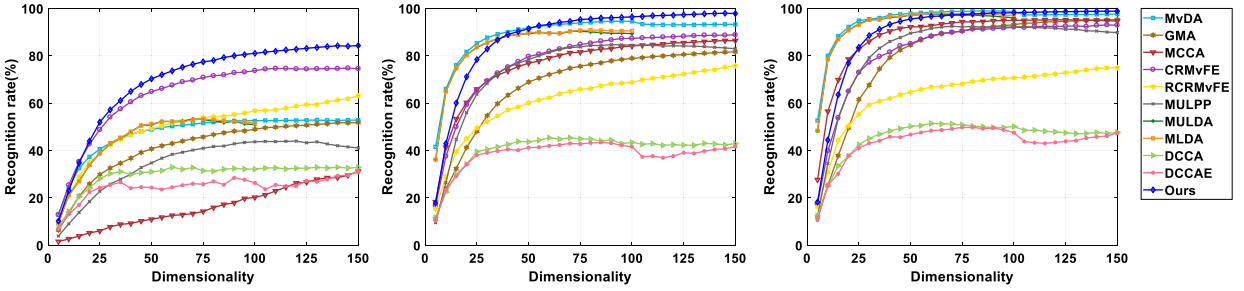
Fig. 3. t-SNE visualizations on CMU PIE dataset learned by (a) GMA, (b) MCCA, (c) MvDA, (d) GMR.

Table 2
Recognition rate (%), standard deviation, dimension and training time of different models on AR, CMU PIE, and 100Leaves datasets.

Dataset (c)	AR (100)			CMU PIE (60)			100Leaves (100)		
N	2	3	4	3	4	5	3	5	7
MvDA	52.87 ± 7.10 (130) (0.35 s)	94.60 ± 5.19 (90)(0.35 s)	**98.93 ± 0.70 (95) (0.37 s)**	72.06 ± 6.07 (65) (0.71 s)	94.57 ± 10.40 (60) (0.71 s)	99.03 ± 0.02 (65) (0.71 s)	43.85 ± 0.90 (48) (0.07 s)	53.55 ± 0.28 (36) (0.08 s)	59.30 ± 0.22 (40) (0.08 s)
GMA	51.82 ± 15.04 (150) (0.10 s)	81.69 ± 12.50 (150) (0.11 s)	95.35 ± 3.38 (150) (0.14 s)	73.90 ± 5.63 (150) (0.49 s)	85.29 ± 10.30 (150) (0.63 s)	96.50 ± 0.36 (150) (0.59 s)	46.98 ± 0.76 (60) (0.03 s)	54.24 ± 0.08 (60) (0.04 s)	54.57 ± 0.05 (56) (0.07 s)
MCCA	31.22 ± 9.20 (150) (0.24 s)	86.40 ± 6.09 (145) (0.24 s)	95.25 ± 1.54 (120) (0.25 s)	78.83 ± 5.34 (150)(1.08 s)	91.25 ± 9.34 (150) (1.39 s)	98.69 ± 0.08 (150) (1.36 s)	35.46 ± 0.38 (64) (0.06 s)	43.87 ± 0.07 (52) (0.05 s)	50.18 ± 0.22 (64) (0.06 s)
CRMvFE	74.80 ± 8.50 (145)(0.08 s)	88.85 ± 4.68 (150) (0.08 s)	93.02 ± 2.54 (145) (0.09 s)	75.06 ± 6.44 (145) (0.67 s)	96.17 ± 7.89 (150)(0.65 s)	**99.43 ± 0.16 (135) (0.68 s)**	51.79 ± 0.52 (40)(0.03 s)	59.70 ± 0.27 (36)(0.04 s)	62.80 ± 0.05 (48)(0.05 s)
RCRMvFE	63.12 ± 4.72 (150) (2.74 s)	75.80 ± 4.83 (150) (2.95 s)	74.97 ± 6.05 (150) (3.23 s)	38.08 ± 8.36 (150) (17.95 s)	85.56 ± 11.03 (150) (22.71 s)	96.45 ± 0.34 (150) (22.79 s)	47.82 ± 0.15 (64) (0.83 s)	51.31 ± 0.55 (64) (1.03 s)	55.89 ± 0.33 (60) (1.70 s)
MULPP	43.95 ± 9.46 (120) (59.92 s)	84.70 ± 8.42 (90) (65.93 s)	92.47 ± 3.76 (70) (67.92 s)	-	-	-	40.63 ± 0.68 (24) (4.86 s)	49.12 ± 0.48 (16) (6.70 s)	56.14 ± 0.11 (16) (8.12 s)
MULDA	53.29 ± 10.33 (70) (13.29 s)	90.45 ± 7.64 (75) (6.86 s)	97.68 ± 1.43 (60) (6.54 s)	71.93 ± 6.64 (60) (46.49 s)	95.73 ± 9.07 (60) (26.21 s)	98.82 ± 0.13 (60) (24.82 s)	51.66 ± 0.45 (40) (0.68 s)	54.12 ± 0.64 (48) (0.78 s)	56.68 ± 0.50 (36) (0.92 s)
MLDA	53.32 ± 9.70 (75) (0.26 s)	90.92 ± 7.10 (80) (0.22 s)	97.93 ± 1.32 (65) (0.22 s)	72.08 ± 6.61 (60) (0.90 s)	95.74 ± 9.04 (60) (0.69 s)	98.88 ± 0.12 (60) (0.64 s)	51.38 ± 0.15 (60) (0.06 s)	55.86 ± 0.32 (48) (0.05 s)	60.77 ± 0.59 (64) (0.07 s)
DCCA	32.99 ± 5.35 (60) (20.57)	45.39 ± 1.81 (70) (27.71)	51.47 ± 4.36 (60) (33.91)	46.69 ± 2.89 (60) (37.44 s)	64.78 ± 6.17 (60) (43.09 s)	72.97 ± 0.41 (60) (48.27 s)	32.50 ± 0.10 (48) (28.41 s)	35.77 ± 0.23 (44) (25.74 s)	41.00 ± 0.43 (52) (32.23 s)
DCCAE	31.03 ± 2.22 (150) (20.41)	43.43 ± 5.27 (85) (35.91 s)	49.78 ± 3.16 (80) (42.25 s)	41.92 ± 3.04 (35) (43.09 s)	60.75 ± 6.71 (40) (51.39 s)	71.59 ± 0.04 (65) (55.81 s)	33.87 ± 0.14 (36) (20.93 s)	38.14 ± 0.42 (36) (27.40 s)	43.22 ± 0.11 (36) (38.61 s)
Ours	**84.32 ± 8.93 (150) (30.82 s)**	**97.97 ± 1.54 (145) (74.23 s)**	98.83 ± 0.85 (145)(116.37 s)	**85.10 ± 4.03 (150) (52.50 s)**	**97.42 ± 6.89 (150) (90.83 s)**	99.40 ± 0.04 (140)(133.95 s)	**54.15 ± 0.20 (64) (44.60 s)**	**60.38 ± 0.49 (64) (116.65 s)**	**65.76 ± 0.36 (64) (253.13 s)**

ture extraction (RCRMvFE) [36], MULPP [21], multiview uncorrelated linear discriminant analysis (MULDA) [46], multiview linear discriminant analysis (MLDA) [46], Deep Canonical Correlation Analysis (DCCA) [47], and deep canonically correlated autoencoders (DCCAE) [48]. From the formulation of the objective function of GMR in Eq. (13), there are 2 balance parameters (i.e., β and γ). We explore the optimal value of them in $[10^{-5}, 10^{5}]$.

For data pre-processing, PCA was performed to reduce dimensions. N images of an object from each view were randomly selected for training. The remaining samples were used for testing. The value of N varied with different datasets. After obtaining the pro-

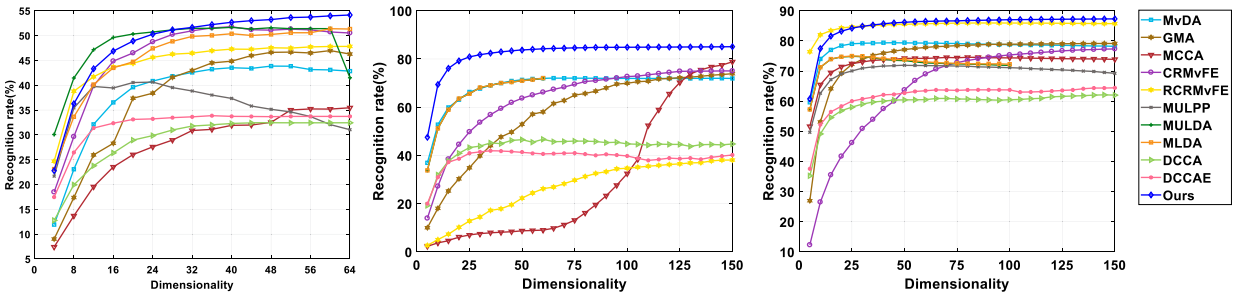


(a) (b) (c)

Fig. 4. Recognition rate versus dimension on AR face dataset when (a) $N = 2$, (b) $N = 3$, (c) $N = 4$.

Table 3
Recognition rate (%), standard deviation, dimension and training time of different models on COIL-100 and NUS-WIDE datasets.

Dataset (c)	COIL-100 (100)			NUS-WIDE (10)		
N	4	5	6	20	25	30
MvDA	79.41 ± 3.25 (50) (0.17 s)	84.17 ± 3.33 (40) (0.17 s)	84.74 ± 5.48 (40) (0.17 s)	23.85 ± 0.55 (10) (0.14 s)	26.07 ± 1.00 (10) (0.13 s)	26.84 ± 0.50 (10) (0.13 s)
GMA	79.15 ± 2.42 (145) (0.13 s)	84.55 ± 3.91 (145) (0.12 s)	85.06 ± 6.15 (125) (0.16 s)	23.80 ± 0.76 (40) (0.12 s)	25.87 ± 0.97 (45) (0.11 s)	26.71 ± 0.53 (45) (0.12 s)
MCCA	74.56 ± 2.99 (80) (0.24 s)	80.43 ± 3.76 (55) (0.23 s)	81.38 ± 5.53 (45) (0.30 s)	16.45 ± 0.69 (100) (0.21 s)	16.95 ± 0.50 (100) (0.20 s)	17.52 ± 0.46 (80) (0.21 s)
CRMvFE	77.33 ± 2.92 (150) (0.09 s)	82.33 ± 3.75 (150) (0.08 s)	82.31 ± 5.76 (150) (0.09 s)	24.37 ± 0.49 (20) (0.12 s)	25.31 ± 0.99 (25) (0.13 s)	25.87 ± 0.54 (25) (0.16 s)
RCRMvFE	85.95 ± 3.73 (85)(3.80 s)	89.54 ± 2.78 (85)(4.81 s)	88.81 ± 5.34 (95)(5.43 s)	26.37 ± 0.81 (40)(2.05 s)	27.49 ± 1.00 (55)(2.43 s)	27.58 ± 0.50 (45)(2.97 s)
MULPP	71.92 ± 3.24 (50) (117.18 s)	77.96 ± 4.03 (55) (119.59 s)	79.03 ± 6.25 (40) (121.80 s)	17.42 ± 0.89 (10) (505.83 s)	19.10 ± 1.00 (10) (557.40 s)	19.74 ± 0.76 (15) (607.30 s)
MULDA	74.83 ± 2.97 (25) (8.18 s)	80.02 ± 3.59 (25) (7.24 s)	81.03 ± 6.08 (30) (6.89 s)	20.66 ± 0.69 (10) (0.40 s)	21.78 ± 0.67 (10) (0.38 s)	22.32 ± 0.51 (10) (0.39 s)
MLDA	74.97 ± 3.01 (25) (0.36 s)	80.11 ± 3.57 (25) (0.34 s)	81.14 ± 5.99 (30) (0.33 s)	20.73 ± 0.73 (10) (0.08 s)	22.18 ± 0.78 (10) (0.07 s)	22.65 ± 0.47 (10) (0.07 s)
DCCA	62.10 ± 2.27 (145) (41.70 s)	68.07 ± 2.27 (150) (49.42 s)	70.09 ± 3.57 (145) (57.26 s)	25.78 ± 0.75 (100) (16.18 s)	27.46 ± 0.78 (100) (18.31 s)	27.74 ± 1.15 (100) (18.03 s)
DCCAE	64.43 ± 1.63 (150) (48.95 s)	70.39 ± 1.97 (150) (62.35 s)	71.71 ± 3.41 (145) (81.59 s)	24.51 ± 0.85 (65) (14.48 s)	25.53 ± 1.35 (55) (19.16 s)	26.41 ± 0.72 (100) (20.93 s)
Ours	**87.30 ± 4.24 (150) (148.14 s)**	**90.40 ± 2.77 (150) (191.29 s)**	**89.21 ± 5.41 (135) (266.32 s)**	**28.22 ± 0.84 (85) (27.13 s)**	**28.77 ± 0.80 (100) (43.91 s)**	**29.04 ± 0.74 (95) (62.82 s)**



(a) (b) (c)

Fig. 5. Recognition rate versus dimension on (a) 100Leaves, (b) CMU PIE, (c) COIL100 dataset.

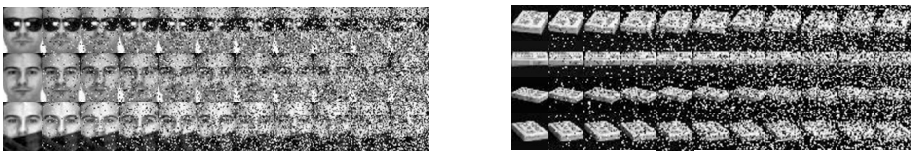
jections of all views using the iterative algorithm shown in Algorithm 1, the 1 NN classifier was utilized to predict which class the projected samples belong to. Moreover, the final recognition rate was the average of 10 independent experiments.

5.2. Discussions

The experimental results for the AR, CMU PIE, 100Leaves, COIL-100, NUS-WIDE, and HUMBI datasets are listed in Tables 2–4. In these tables, the highest recognition rates are shown in bold, and the second-highest rates are underlined. Fig. 3 presents the t-SNE visualizations learned by GMA, MCCA, MvDA and GMR on the CMU PIE dataset. Figs. 4 and 5 show the recognition rate versus the

Table 4
Recognition rate (%), standard deviation, dimension and training time of different models on 3Sources and HUMBI datasets.

Dataset (c)	3Sources (6)			HUMBI (64)		
N	3	5	7	2	3	4
MvDA	45.05 ± 2.58 (6) (0.09 s)	48.90 ± 5.11 (6) (0.09 s)	55.96 ± 4.09 (6) (0.10 s)	94.19 ± 1.23 (72) (0.16 s)	98.26 ± 0.52 (100)(0.16 s)	**98.82 ± 0.28 (100) (0.16 s)**
GMA	42.96 ± 2.25 (18) (0.07 s)	42.61 ± 5.25 (18) (0.08 s)	42.62 ± 5.83 (18) (0.18 s)	97.65 ± 0.55 (100)(0.06 s)	97.26 ± 0.70 (100) (0.07 s)	98.74 ± 0.28 (100) (0.07 s)
MCCA	44.26 ± 6.95 (24) (0.24 s)	41.15 ± 0.43 (39) (0.26 s)	44.41 ± 2.52 (42) (0.28 s)	81.33 ± 3.36 (100) (0.11 s)	97.28 ± 1.02 (100) (0.12 s)	98.65 ± 0.26 (100) (0.12 s)
CRMvFE	58.83 ± 0.99 (9) (0.17 s)	65.73 ± 4.96 (15) (0.18 s)	62.26 ± 0.94 (12) (0.17 s)	95.87 ± 0.64 (64) (0.09 s)	97.37 ± 0.57 (64) (0.10 s)	97.84 ± 0.38 (64) (0.12 s)
RCRMvFE	59.49 ± 4.97 (12)(1.08 s)	68.66 ± 4.82 (15)(1.07 s)	63.44 ± 0.55 (21) (1.07 s)	66.19 ± 4.59 (100) (1.55 s)	95.18 ± 1.73 (100) (1.73 s)	97.47 ± 0.88 (100) (1.90 s)
MULPP	23.35 ± 4.07 (21) (570.94 s)	28.78 ± 8.39 (48) (630.55 s)	25.28 ± 11.04 (81) (651.82 s)	-	-	-
MULDA	51.19 ± 1.39 (6) (0.19 s)	58.49 ± 1.51 (6) (0.34 s)	66.22 ± 3.39 (6)(0.47 s)	92.91 ± 1.63 (60) (4.87 s)	97.21 ± 0.74 (48) (2.88 s)	97.91 ± 0.49 (60) (2.59 s)
MLDA	51.19 ± 1.39 (6) (0.07 s)	58.70 ± 1.58 (6) (0.20 s)	66.22 ± 3.39 (6)(0.25 s)	93.23 ± 1.45 (60) (0.09 s)	97.31 ± 0.70 (60) (0.09 s)	98.06 ± 0.49 (60) (0.08 s)
DCCA	37.00 ± 3.58 (135) (17.39 s)	31.29 ± 1.80 (147) (18.06 s)	23.49 ± 0.39 (81) (18.28 s)	70.83 ± 3.19 (40) (56.73 s)	82.01 ± 1.17 (48) (70.48 s)	87.60 ± 0.76 (52) (86.10 s)
DCCAE	35.98 ± 5.96 (132) (16.50 s)	42.11 ± 4.60 (96) (18.21 s)	44.15 ± 1.42 (84) (21.16 s)	67.56 ± 2.54 (52) (39.23 s)	83.12 ± 0.92 (52) (50.07 s)	90.83 ± 1.46 (100) (77.01 s)
Ours	**66.07 ± 2.54 (144) (0.60 s)**	**73.38 ± 2.42 (150) (1.03 s)**	**75.46 ± 0.39 (150) (1.27 s)**	**98.44 ± 0.36 (100) (9.04 s)**	**98.69 ± 0.29 (100) (14.41 s)**	98.78 ± 0.32 (100)(27.39 s)



(a) (b)

Fig. 6. The samples from (a) AR face dataset, (b) COIL-100 dataset with Salt & Pepper noise. For both of these two datasets, the noise densities are 0, 0.05, 0.1, 0.15, 0.2, 0.25, 0.3, 0.35, 0.4, 0.45, 0.5 from left to right.

dimensions of the subspaces obtained by different methods. The samples used to evaluate the robustness of these methods are shown in Fig. 6, and the performances of the various methods versus noise density are visualized in Fig. 7. From the results, we can arrive at the following conclusions:

(1) The proposed MVL method, GMR, outperformed the other multiview methods in most cases, particularly on the 100Leaves, COIL-100 and NUS-WIDE datasets. Moreover, the features learned by GMR are more discriminative and compact as shown in Fig. 3. This is because GMR takes the local geometric structure and label indicators into consideration, and enforces $L_{2,1}$ norm on the loss function and regularization term. The local geometric structure of each view enables the features to learn more information. The label indicators enable the features to be more discriminative. Moreover, the $L_{2,1}$ norm-based regularization term can make the projections sparser and eliminate redundant information. With the aid of these factors, GMR can achieve remarkable performance in multi-view classification tasks.

(2) MULPP failed to work on the CMU PIE with 7 views and HUMBI with 5 views. Because the number of views in these two datasets were large and the required space cost of the tensor object in MULPP was too large to be created. GMR utilizes label information to explore the consistency of all views, while MCCA optimizes the sum of the correlations of each pair views. Moreover, the reconstruction term in GMR preserves complementary information compared with MvDA, MCCA, DCCA and DCCAE. In addition, the extracted features in GMR could reconstruct the original data compared with MULPP. The experimental results indicated that GMR achieved the best performance among all these methods in most cases. That implied GMR could extract informative features.

(3) From Tables 2–4 and Fig. 4, GMR could rank the best two among all the MVL methods as the number of training samples increases. In addition, the GMR performed better than the other methods with fewer training samples. For example, when $N = 3$ in the CMU PIE dataset, the recognition rate of GMR was the highest (85.10%) among all methods and the second was 78.83%. Moreover, GMR was the second highest (99.40%) with N equal to 5. This phenomenon can be seen in the experiments on the AR and HUMBI datasets as well. The gap between GMR and the other methods narrowed with the increasing N according to Fig. 4. This was because more information was contained as the number of training samples increased. And most of these methods

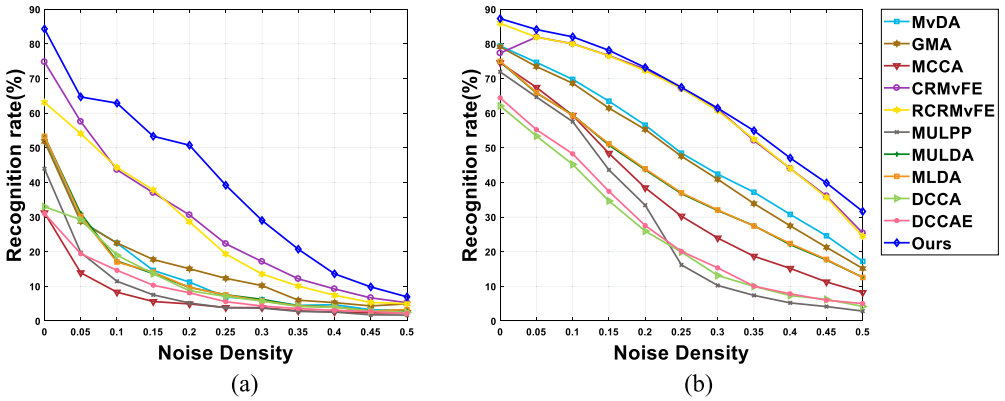


Fig. 7. The recognition rate versus noise density on (a) AR face dataset, (b) COIL-100 dataset with Salt & Pepper noise.

achieved good performance. When the size of the training set was small, the features extracted by GMR remained informative and discriminative even if the number of training samples decreased to 2 in each class. This is because GMR integrates view-specific local geometric structures, label information, and $L_{2,1}$ regularization term.

(4) As Figs. 4 and 5 show, the performances of some methods declined with the increasing dimensionality, but the performance of GMR remained the best or continued to improve. This is because redundant information exists when the dimensionality of the feature is high. The regularization term based $L_{2,1}$ norm helps GMR eliminate redundant information so that it can outperform the other methods when the learned representations have high dimensionality. However, the deep learning methods (i.e., DCCA and DCCAE) perform poorly in most cases. There are two reasons. First, these two methods maximize only the pairwise correlation in each class, which may lead to overlapping between different views' samples. Second, these two deep learning methods cannot learn enough information when the number of training samples is small. These two deep learning methods require more training samples, which cannot be satisfied in this case. This will cause the networks to be overfitting or underfitting. Moreover, the results indicated that the conventional multiview methods performed better than deep learning methods on the classification tasks when the number of training samples was small. Thus, it is still necessary to develop conventional multiview techniques.

(5) GMR maintained the best performance among these MVL methods when noise existed in samples (the images with noise are shown in Fig. 6). For example, the average recognition rate of GMR remained the best with increasing noise density (see Fig. 7). That is because GMR adopts $L_{2,1}$ norm as the basic metric and the impact of outliers could be weakened. Therefore, GMR could extract discriminative features with stronger robustness than the other methods.

5.3. Ablation study

In this subsection, we examine the contributions of $L_{2,1}$ norm-based regularization and consistency information. GMR_1 refers to the GMR after removing the regularization term, and its objective function is as follows:

$$\min_{\mathbf{A},\mathbf{D},\mathbf{P}^{(j)}} \sum_{j}^{v} \sum_{k,r} ||\mathbf{x}_k^{(j)} - \mathbf{A}\mathbf{P}^{(j)^T}\mathbf{x}_r^{(j)}||_2 \mathbf{W}_{k,r}^{(j)} + \beta \sum_{j}^{n} \sum_{k}^{n_j} ||\mathbf{D}\mathbf{y}_k^{(j)} - \mathbf{P}^{(j)^T}\mathbf{x}_k^{(j)}||_2 \quad (28)$$
$$\text{s.t. } \mathbf{A}^T\mathbf{A} = \mathbf{I}, \mathbf{D}^T\mathbf{D} = \mathbf{I}$$

$L_{2,1}$ norm-based regularization term could enhance the sparsity of the learned projections. Without the $L_{2,1}$ norm-based regularization term, the projections learned by GMR_1 lack sparsity, which leads to the lack of the function of feature selection. GMR_2 refers to the GMR without the classification error term and its objective function can be formulated as follows:

$$\min_{\mathbf{A},\mathbf{P}^{(j)}} \sum_{j}^{v} \sum_{k,r} ||\mathbf{x}_k^{(j)} - \mathbf{A}\mathbf{P}^{(j)^T}\mathbf{x}_r^{(j)}||_2 \mathbf{W}_{k,r}^{(j)} + \gamma \sum_{j}^{v} ||\mathbf{P}^{(j)}||_{2,1} \quad (29)$$
$$\text{s.t. } \mathbf{A}^T\mathbf{A} = \mathbf{I}$$

The optimal solution of $\mathbf{P}$ in GMR_2 is $\mathbf{P} = (\mathbf{X}_{diag}\mathbf{D}_g\mathbf{X}_{diag}^T + \gamma\mathbf{D}_p)^{-1}(\mathbf{X}_{diag}\mathbf{G}\mathbf{X}^T\mathbf{A})$. The features extracted from GMR_2 could not integrate discriminative information. The performances of these models were evaluated in the AR face, COIL-100, NUS-WIDE, and 100Leaves datasets and listed in Table 5. As presented in the table, the GMR outperformed GMR_1 even when the dimensionality of features in GMR_1 was higher than that in GMR. Because the projection of GMR_1 is not sparse enough, GMR can eliminate more redundant information than GMR_1. For GMR_2, it does not utilize the label information so that the learned features are less discriminative. Thus, each component of GMR contributes to its performance.

Table 5
Recognition rate (%) and dimension of different models on different datasets.

Dataset	AR			coil100			NUS-WIDE			100Leaves		
N	2	3	4	4	5	6	20	25	30	3	5	7
GMR_1	58.14 (150)	92.87 (150)	97.90 (145)	78.23 (145)	85.08 (150)	86.59 (150)	22.58 (95)	24.73 (100)	25.79 (95)	49.56 (64)	57.61 (64)	63.53 (64)
GMR_2	12.67 (120)	25.43 (135)	30.57 (150)	83.52 (150)	87.89 (150)	88.26 (150)	12.51 (15)	13.41 (100)	14.61 (100)	50.07 (64)	58.00 (64)	63.55 (64)
GMR	**84.32 (150)**	**97.97 (145)**	**98.83 (145)**	**87.30 (150)**	**90.40 (150)**	**89.21 (135)**	**28.22 (85)**	**28.77 (100)**	**29.05 (100)**	**54.15 (64)**	**60.38 (64)**	**65.76 (64)**

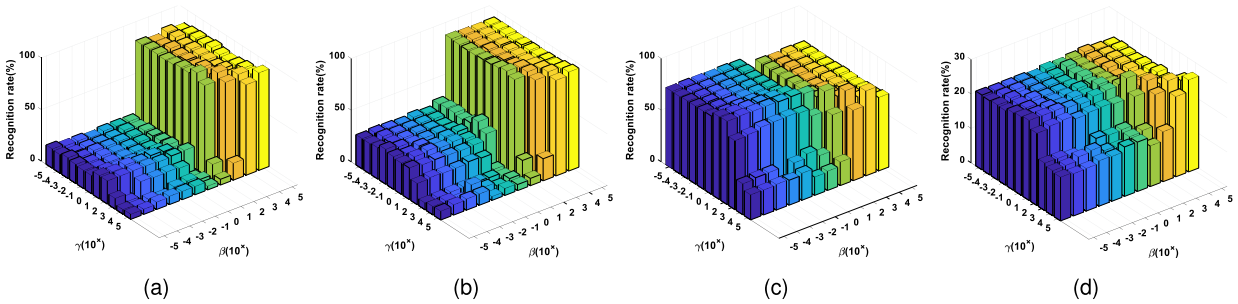


Fig. 8. Recognition rate versus β and γ on (a) AR face, (b) CMU PIE, (c) COIL100, (d) NUS-WIDE dataset.

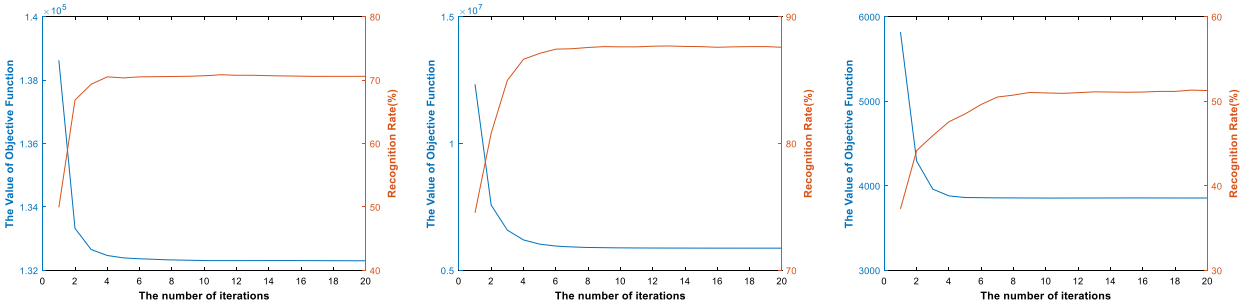


Fig. 9. The curves of recognition rate and convergence on (a) AR face, (b) COIL-100, (c) 100Leaves datasets.

5.4. Parameter sensitivity study

Fig. 8 shows the recognition rate of GMR on different datasets when β and γ varied from 10^{-5} to 10^{5}. Parameter γ implies the importance of the regularization term for the performance of the GMR, and β indicates the contribution of the consistency information term. All the illustrations in Fig. 8 indicate that both parameters have impacts on experimental results. It can be observed that the accuracy on the face data sets (i.e., the AR and CMU PIE datasets) decreased abruptly when $\beta \leq 1$. This indicates label information makes the extracted features more discriminative in face recognition. In addition, it can be observed that the accuracy decreased in all the datasets when $\gamma \geq 10^{3}$ and $\beta \leq 10^{4}$. This is because when γ is large and β is small, discriminative information is not emphasized sufficiently and the projection $\mathbf{P}^{(j)}$ is so sparse that it eliminates an excessive number of features. In general, GMR performs effectively when the parameter $\beta \geq 100$.

5.5. Convergence analysis

To obtain the optimal solution of $\mathbf{P}^{(j)}$, an iterative algorithm is designed and section 4.2 presents the theoretical analysis of convergence. The curves of recognition rate and convergence on the AR, COIL-100 and 100Leaves datasets are shown in Fig. 9. As can be seen in Fig. 9, the value of the objective function decreased gradually and attained stability in 10 times. This implies that GMR converges highly rapidly. As the objective function value decreased, the recognition rate increased and tended to be stable within several iterations. In addition, the training time of all the methods can be seen in Tables 2–4. From the experimental results, GMR was more time-consuming than the methods obtaining optimal solution by eigenvalue decomposition only once, (i.e., MvDA, GMA, MCCA, and CRMvFE). However, the training process of GMR is conducted offline and the time cost is affordable.

6. Conclusion

The MVL methods based on CCA aim to maximize the correlation between each two views so that the correlation of all views and the complementary information is not considered. In this paper, we propose a method for multiview feature extraction called GMR. It explores both view consistency and complementary information. GMR reconstructs the original data and preserves the view-specific local geometrical structures. As such, GMR could explore complementary information among all views. In addition, we introduce label information to make the extracted features more discriminant and achieve view consistency. GMR uses the $L_{2,1}$ norm as basic metric to measure the loss function and regularization term. An iterative algorithm for computing the optimal solution and its theoretical analyses are presented.

In terms of robustness and recognition capability, the experimental results demonstrate that GMR outperforms the other multiview methods in most cases. On the AR, COIL-100, NUS-WIDE, and 100Leaves datasets, the $L_{2,1}$ regularization term and consistency information are demonstrated to be effective. Hence, GMR is a multiview method for extracting discriminative, sparse and robust features. In addition, the deep-learning methods perform worse than GMR because of the limited number of training samples and the importance of the consistency of all views. Thus, developing conventional multiview methods is still necessary.

However, the computational complexity of GMR is marginally high since it is an iterative algorithm, and SVD is used in each iteration. For real-time scenarios (e.g. fast similarity search), GMR can learn the projections off-line even if it is time-consuming. Then, the discriminative features of the test samples could be extracted online and used in similarity searching. Moreover, some real-world applications (e.g., emotion recognition) require the multiview methods to be equipped with the capability of multiview fusion. In fact, GMR is based on complete multiview data and it is not designed for fast similarity searching or multiview fusion. Therefore, GMR could be improved or extended in four respects, including the reduction of the computational complexity, multiview fusion, incomplete MVL and binary MVL, which will be explored in the future.

CRediT authorship contribution statement

Zhihui Lai: Conceptualization, Investigation, Methodology. **Yiling Lin:** Data curation, Writing – original draft. **Jiacan Zheng:** Validation, Visualization. **Jie Zhou:** Writing – review & editing. **Heng Kong:** Writing – review & editing.

Declaration of competing interest

We declare that we have no financial and personal relationships with other people or organizations that can inappropriately influence our work, there is no professional or other personal interest of any nature or kind in any product, service and/or company that could be construed as influencing the position presented in, or the review of, the manuscript entitled.

Data availability

Data will be made available on request.

Acknowledgement

This work was supported in part by the Natural Science Foundation of China under Grant 61976145, Grant 62272319 and Grant 62076164, and in part by the Natural Science Foundation of Guangdong Province (Grant 2314050002242, 2023A1515010677, 2021A1515011861), and in part by the Shenzhen Municipal Science and Technology Innovation Council under Grant JCYJ20210324094413037, JCYJ20220818095803007, and JCYJ20210324094601005.

Appendix A

Proof of the Theorem 1. Let $\mathbf{A}_{(\tau)}$, $\mathbf{P}_{(\tau)}$, $\mathbf{D}_{(\tau)}$, $\mathbf{G}_{(\tau)}$, $\mathbf{D}_{p(\tau)}$ $\mathbf{D}_{g(\tau)}$, and $\mathbf{D}_{rr(\tau)}$ be the matrices $\mathbf{A}$, $\mathbf{P}$, $\mathbf{D}$, $\mathbf{G}$, $\mathbf{D}_g$, $\mathbf{D}_p$, and $\mathbf{D}_{rr}$ obtained in τ-th iteration, respectively. We define the objective function in Eq. (18) as $F(\mathbf{A},\mathbf{P},\mathbf{D},\mathbf{G},\mathbf{D}_g,\mathbf{D}_p,\mathbf{D}_{rr})$. The following inequality must be proved:

$$\begin{aligned} &F(\mathbf{A}_{(\tau)},\mathbf{P}_{(\tau)},\mathbf{D}_{(\tau)},\mathbf{G}_{(\tau)},\mathbf{D}_{g(\tau)},\mathbf{D}_{p(\tau)},\mathbf{D}_{rr(\tau)}) \\ &\le F(\mathbf{A}_{(\tau-1)},\mathbf{P}_{(\tau-1)},\mathbf{D}_{(\tau-1)},\mathbf{G}_{(\tau-1)},\mathbf{D}_{g(\tau-1)},\mathbf{D}_{p(\tau-1)},\mathbf{D}_{rr(\tau-1)}) \end{aligned} \tag{30}$$

In the $(\tau-1)$-th iteration, $\mathbf{P}_{(\tau-1)}$, $\mathbf{D}_{g(\tau-1)}$, $\mathbf{D}_{rr(\tau-1)}$, $\mathbf{P}_{(\tau-1)}$ have been obtained. And we can compute $\mathbf{A}_{(\tau)}$ by Eq. (21), which can reduce the objective function. Then we have:

$$\begin{aligned} &F(\mathbf{A}_{(\tau)},\mathbf{P}_{(\tau-1)},\mathbf{D}_{(\tau-1)},\mathbf{G}_{(\tau-1)},\mathbf{D}_{g(\tau-1)},\mathbf{D}_{p(\tau-1)},\mathbf{D}_{rr(\tau-1)}) \\ &\le F(\mathbf{A}_{(\tau-1)},\mathbf{P}_{(\tau-1)},\mathbf{D}_{(\tau-1)},\mathbf{G}_{(\tau-1)},\mathbf{D}_{g(\tau-1)},\mathbf{D}_{p(\tau-1)},\mathbf{D}_{rr(\tau-1)}) \end{aligned} \tag{31}$$

$\mathbf{D}_{(\tau)}$ is obtained by SVD. Thus it is a optimal solution that decreases the value of Eq. (24), and we get:

$$F(\mathbf{A}_{(\tau)},\mathbf{P}_{(\tau-1)},\mathbf{D}_{(\tau)},\mathbf{G}_{(\tau-1)},\mathbf{D}_{g(\tau-1)},\mathbf{D}_{p(\tau-1)},\mathbf{D}_{rr(\tau-1)}) \le F(\mathbf{A}_{(\tau-1)},\mathbf{P}_{(\tau-1)},\mathbf{D}_{(\tau-1)},\mathbf{G}_{(\tau-1)},\mathbf{D}_{g(\tau-1)},\mathbf{D}_{p(\tau-1)},\mathbf{D}_{rr(\tau-1)}) \tag{32}$$

After obtaining $\mathbf{D}_{(\tau)}$ and $\mathbf{A}_{(\tau)}$, $\mathbf{P}_{(\tau)}$ can be computed by Eq. (25), and we further have:

$$F(\mathbf{A}_{(\tau)},\mathbf{P}_{(\tau)},\mathbf{D}_{(\tau)},\mathbf{G}_{(\tau-1)},\mathbf{D}_{g(\tau-1)},\mathbf{D}_{p(\tau-1)},\mathbf{D}_{rr(\tau-1)}) \le F(\mathbf{A}_{(\tau-1)},\mathbf{P}_{(\tau-1)},\mathbf{D}_{(\tau-1)},\mathbf{G}_{(\tau-1)},\mathbf{D}_{g(\tau-1)},\mathbf{D}_{p(\tau-1)},\mathbf{D}_{rr(\tau-1)}) \tag{33}$$

The matrix $\mathbf{G}_{(\tau)}$ comprises $(\mathbf{G}^{(j)})_{(\tau)}(j=1,...,v)$. They are updated using $(\mathbf{G}^{(j)}_{k,r})_{(\tau)} = \mathbf{W}^{(j)}_{k,r}/\|\mathbf{x}^{(j)T}_{k} - \mathbf{x}^{(j)T}_{r}\mathbf{P}^{(j)}\mathbf{A}^{T}\|_2$ and $(\mathbf{D}^{(j)}_{g})_{(\tau)} = diag(\sum_k^{n_j}(\mathbf{G}^{(j)}_{1,k})_{(\tau)},...,\sum_k^{n_j}(\mathbf{G}^{(j)}_{n_j,k})_{(\tau)})$. So the following inequality holds after obtaining $\mathbf{A}_{(\tau)}$ and $\mathbf{P}_{(\tau)}$:

$$F(\mathbf{A}_{(\tau)},\mathbf{P}_{(\tau)},\mathbf{D}_{(\tau)},\mathbf{G}_{(\tau)},\mathbf{D}_{g(\tau)},\mathbf{D}_{p(\tau-1)},\mathbf{D}_{rr(\tau-1)}) \le F(\mathbf{A}_{(\tau-1)},\mathbf{P}_{(\tau-1)},\mathbf{D}_{(\tau-1)},\mathbf{G},\mathbf{D}_{g(\tau-1)},\mathbf{D}_{p(\tau-1)},\mathbf{D}_{rr(\tau-1)}) \tag{34}$$

For simplicity, let $\Phi = \mathbf{Y}^T\mathbf{D}^T - \mathbf{X}^T_{diag}\mathbf{P}$. And Eq. (34) can be written as:

$$\begin{aligned}
&\mathcal{L}_{1(\tau)} + \beta tr(\Phi^T_{(\tau)}\mathbf{D}_{rr(\tau-1)}\Phi_{(\tau)}) + \gamma tr(\mathbf{P}^T_{(\tau)}\mathbf{D}_{p(\tau-1)}\mathbf{P}_{(\tau)}) \\
&\le \mathcal{L}_{1(\tau-1)} + \beta tr(\Phi^T_{(\tau-1)}\mathbf{D}_{rr(\tau-1)}\Phi_{(\tau-1)}) + \gamma tr(\mathbf{P}^T_{(\tau-1)}\mathbf{D}_{p(\tau-1)}\mathbf{P}_{(\tau-1)}) \\
&\Rightarrow \mathcal{L}_{1(\tau)} + \beta\sum_i \frac{\|\Phi^i_{(\tau)}\|_2^2}{\|\Phi^i_{(\tau-1)}\|_2} + \gamma\sum_i \frac{\|\mathbf{P}^i_{(\tau)}\|_2^2}{\|\mathbf{P}^i_{(\tau-1)}\|_2} \\
&\le \mathcal{L}_{1(\tau-1)} + \beta\sum_i \frac{\|\Phi^i_{(\tau-1)}\|_2^2}{\|\Phi^i_{(\tau-1)}\|_2} + \gamma\sum_i \frac{\|\mathbf{P}^i_{(\tau-1)}\|_2^2}{\|\mathbf{P}^i_{(\tau-1)}\|_2} \\
&\Rightarrow \mathcal{L}_{1(\tau)} + \beta\sum_i(\|\Phi^i_{(\tau)}\|_2 - (\|\Phi^i_{(\tau)}\|_2 - \frac{\|\Phi^i_{(\tau)}\|_2^2}{\|\Phi^i_{(\tau-1)}\|_2})) \\
&\quad + \gamma\sum_i(\|\mathbf{P}^i_{(\tau)}\|_2 - (\|\mathbf{P}^i_{(\tau)}\|_2 - \frac{\|\mathbf{P}^i_{(\tau)}\|_2^2}{\|\mathbf{P}^i_{(\tau-1)}\|_2})) \\
&\le \mathcal{L}_{1(\tau-1)} + \beta\sum_i(\|\Phi^i_{(\tau-1)}\|_2 - (\|\Phi^i_{(\tau-1)}\|_2 - \frac{\|\Phi^i_{(\tau-1)}\|_2^2}{\|\Phi^i_{(\tau-1)}\|_2})) \\
&\quad + \gamma\sum_i(\|\mathbf{P}^i_{(\tau-1)}\|_2 - (\|\mathbf{P}^i_{(\tau-1)}\|_2 - \frac{\|\mathbf{P}^i_{(\tau-1)}\|_2^2}{\|\mathbf{P}^i_{(\tau-1)}\|_2}))
\end{aligned} \tag{35}$$

According to Lemma 2, the following inequalities hold on:

$$\sum_i\|\mathbf{P}^i_{(\tau)}\|_2 - \sum_i\frac{\|\mathbf{P}^i_{(\tau)}\|_2^2}{2\|\mathbf{P}^i_{(\tau-1)}\|_2} \le \sum_i\|\mathbf{P}^i_{(\tau-1)}\|_2 - \sum_i\frac{\|\mathbf{P}^i_{(\tau-1)}\|_2^2}{2\|\mathbf{P}^i_{(\tau-1)}\|_2} \tag{36}$$

$$\sum_i\|\Phi^i_{(\tau)}\|_2 - \sum_i\frac{\|\Phi^i_{(\tau)}\|_2^2}{2\|\Phi^i_{(\tau-1)}\|_2} \le \sum_i\|\Phi^i_{(\tau-1)}\|_2 - \sum_i\frac{\|\Phi^i_{(\tau-1)}\|_2^2}{2\|\Phi^i_{(\tau-1)}\|_2} \tag{37}$$

With Eq. (36) and Eq. (37), we further rewrite Eq. (35) as follows:

$$\mathcal{L}_{1(\tau)} + \beta\sum_i\|\Phi^i_{(\tau)}\|_2 + \gamma\sum_i\|\mathbf{P}^i_{(\tau)}\|_2 \le \mathcal{L}_{1(\tau-1)} + \beta\sum_i\|\Phi^i_{(\tau-1)}\|_2 + \gamma\sum_i\|\mathbf{P}^i_{(\tau-1)}\|_2 \tag{38}$$

Eq. (38) indicates we have:

$$F(\mathbf{A}_{(\tau)},\mathbf{P}_{(\tau)},\mathbf{D}_{(\tau)},\mathbf{G}_{(\tau)},\mathbf{D}_{g(\tau)},\mathbf{D}_{p(\tau)},\mathbf{D}_{rr(\tau)}) \le F(\mathbf{A}_{(\tau-1)},\mathbf{P}_{(\tau-1)},\mathbf{D}_{(\tau-1)},\mathbf{G}_{(\tau-1)},\mathbf{D}_{g(\tau-1)},\mathbf{D}_{p(\tau-1)},\mathbf{D}_{rr(\tau-1)}) \tag{39}$$

Eq. (39) implies that the objective function value monotonically decreases in each iteration and finally converges.

References

[1] M. Kirby, L. Sirovich, Application of the Karhunen-Loeve procedure for the characterization of human faces, IEEE Trans. Pattern Anal. Mach. Intell. 12 (1) (Jan. 1990) 103–108.

[2] H. Hotelling, Relations between two sets of variates, Biometrika 28 (3–4) (1936) 321–377.

[3] B. Gurunlu, S. Ozturk, Efficient approach for block-based copy-move forgery detection, in: Smart Trends in Computing and Communications: Proceedings of SmartCom 2021, 2022, pp. 167–174.

[4] A.M. Martinez, A.C. Kak, PCA versus LDA, IEEE Trans. Pattern Anal. Mach. Intell. 23 (2) (2001) 228–233.

[5] Z. Ren, X. Li, M. Mukherjee, Y. Huang, Q. Sun, Z. Huang, Robust multi-view graph clustering in latent energy-preserving embedding space, Inf. Sci. 569 (2021) 582–595.
[6] Z. Gui, J. Yang, Z. Xie, Robust dimensionality reduction method based on relaxed energy and structure preserving embedding for multiview clustering, Inf. Sci. 621 (2023) 506–523.
[7] X. Ma, X. Yan, J. Liu, G. Zhong, Simultaneous multi-graph learning and clustering for multiview data, Inf. Sci. 539 (2022) 472–487.
[8] X. Gao, Y. Xiong, G. Zhang, H. Deng, K. Kou, Exploiting key points supervision and grouped feature fusion for multiview pedestrian detection, Pattern Recognit. 121 (2022) 108866.
[9] X. Sang, J. Lu, H. Lu, Consensus graph learning for auto-weighted multi-view projection clustering, Inf. Sci. 609 (2022) 816–837.
[10] M. Kan, S. Shan, H. Zhang, S. Lao, X. Chen, Multi-view discriminant analysis, IEEE Trans. Pattern Anal. Mach. Intell. 38 (1) (2016) 188–194.
[11] M. Hou, Z. Zhang, Q. Cao, D. Zhang, G. Liu, Multi-view speech emotion recognition via collective relation construction, IEEE/ACM Trans. Audio Speech Lang. Process. 30 (2022) 218–229.
[12] G.N. Dong, C.M. Pun, Z. Zhang, Temporal relation inference network for multimodal speech emotion recognition, IEEE Trans. Circuits Syst. Video Technol. 32 (9) (2022) 6472–6485.
[13] Z. Zhang, L. Liu, F. Shen, H.T. Shen, L. Shao, Binary multi-view clustering, IEEE Trans. Pattern Anal. Mach. Intell. 41 (7) (2019) 1774–1782.
[14] B. Sheng, J. Li, F. Xiao, Q. Li, W. Yang, J. Han, Discriminative multi-view subspace feature learning for action recognition, IEEE Trans. Circuits Syst. Video Technol. 30 (12) (2020) 4591–4600.
[15] X. Wang, W. Bian, D. Tao, Grassmannian regularized structured multi-view embedding for image classification, IEEE Trans. Image Process. 22 (7) (Jul. 2013) 2646–2660.
[16] J. Gui, D. Tao, Z. Sun, Y. Luo, X. You, Y.Y. Tang, Group sparse multiview patch alignment framework with view consistency for image classification, IEEE Trans. Image Process. 23 (7) (2014) 3126–3137.
[17] X. You, J. Xu, W.Yuan.X.Y. Jing, D. Tao, T. Zhang, Multi-view common component discriminant analysis for cross-view classification, Pattern Recognit. 92 (2019) 37–51.
[18] M.A. Turk, A.P. Pentland, Face recognition using eigenfaces, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit, 1991, pp. 586–591.
[19] X. He, S. Yan, Y. Hu, P. Niyogi, H.-J. Zhang, Face recognition using laplacianfaces, IEEE Trans. Pattern Anal. Mach. Intell. 27 (3) (Mar. 2005) 328–340.
[20] M. Wan, Y. Yao, T. Zhan, G. Yang, Supervised low-rank embedded regression (SLRER) for robust subspace learning, IEEE Trans. Circuits Syst. Video Technol. 32 (4) (2022) 1917–1927.
[21] J. Yin, S. Sun, Multiview uncorrelated locality preserving projection, IEEE Trans. Neural Netw. Learn. Syst. 31 (9) (2020) 3442–3455.
[22] J. Rupnik, J. Shawe-Taylor, Multi-view canonical correlation analysis, in: Proc. Slovenian KDD Conf. Data Mining Data Warehouses, 2010, pp. 1–4.
[23] Y. Luo, D. Tao, K. Ramamohanarao, C. Xu, Y. Wen, Tensor canonical correlation analysis for multi-view dimensionality reduction, IEEE Trans. Knowl. Data Eng. 27 (11) (2015) 3111–3124.
[24] Z. Gui, J. Yang, Z. Xie, Robust dimensionality reduction method based on relaxed energy and structure preserving embedding for multiview clustering, Inf. Sci. 621 (2023) 506–523.
[25] X. Yang, C. Li, Y. Shao, Robust multi-view discriminant analysis with view-consistency, Inf. Sci. 596 (2022) 153–168.
[26] M. Najafi, L. He, P.S. Yu, Error-robust multi-view clustering, in: Proc. IEEE Int. Conf. Big Data 2017, 2017, pp. 736–745.
[27] B. Jiang, J. Xiang, X. Wu, Y. Wang, H. Chen, W. Cao, W. Sheng, Robust multi-view learning via adaptive regression, Inf. Sci. 610 (2022) 916–937.
[28] F. Nie, H. Huang, X. Cai, C. Ding, Efficient and robust feature selection via joint $L_{2,1}$ norms minimization, in: Proc. Adv. Neural Inf. Process. Syst, 2010, pp. 1813–1821.
[29] J. Lu, J. Lin, Z. Lai, H. Wang, J. Zhou, Target redirected regression with dynamic neighborhood structure, Inf. Sci. 544 (2021) 564–584.
[30] F. Nie, Z. Wang, R. Wang, Z. Wang, X. Li, Towards robust discriminative projections learning via non-greedy $\ell_{2,1}$-norm MinMax, IEEE Trans. Pattern Anal. Mach. Intell. 43 (6) (2021) 2086–2100.
[31] Y. Chen, X. Xiao, C. Peng, G. Lu, Y. Zhou, Low-rank tensor graph learning for multi-view subspace clustering, IEEE Trans. Circuits Syst. Video Technol. 32 (1) (2022) 92–104.
[32] C. Zhang, J. Cheng, Q. Tian, Multi-view image classification with visual, semantic and view consistency, IEEE Trans. Image Process. 29 (2020) 617–627.
[33] C. Li, H. Che, M.-F. Leung, C. Liu, Z. Yan, Robust multi-view non-negative matrix factorization with adaptive graph and diversity constraints, Inf. Sci. 634 (2023) 587–607.
[34] D. Cai, X. He, J. Han, H.-J. Zhang, Orthogonal laplacianfaces for face recognition, IEEE Trans. Image Process. 15 (11) (2006) 3608–3614.
[35] X. He, D. Cai, S. Yan, H.-J. Zhang, Neighborhood preserving embedding, in: Proc. IEEE Int. Conf. Comput. Vis., ICCV, Beijing, China, 2005, pp. 1208–1213.
[36] J. Zhang, L. Jing, J. Tan, Cross-regression for multi-view feature extraction, Knowl.-Based Syst. 200 (11) (2020) 105997.
[37] L. Clemmensen, T. Hastie, D. Witten, B. Ersbøll, Sparse discriminant analysis, Technometrics 53 (4) (2011) 406–413.
[38] T. Hastie, A. Buja, R. Tibshirani, Penalized discriminant analysis, Ann. Stat. 23 (1) (1995) 73–102.
[39] H. Zou, T. Hastie, R. Tibshirani, Sparse principal component analysis, J. Comput. Graph. Stat. 15 (2) (2004) 265–286.
[40] T. Sim, S. Baker, M. Bsat, The CMU pose, illumination, and expression database, IEEE Trans. Pattern Anal. Mach. Intell. 25 (12) (2003) 1615–1618.
[41] C. Mallah, J. Cope, J. Orwell, Plant leaf classification using probabilistic integration of shape, texture and margin features, in: Signal Process., Pattern Recog. Appl., 2013, pp. 279–286.
[42] A.A. Martinez, R. Benavente, The AR face database, CVC, Barcelona, Spain, Tech. Rep. #24, 1998.
[43] T. Chua, J. Tang, R. Hong, H. Li, Z. Luo, Y. Zheng, NUS-WIDE: a real-world web image database from national university of Singapore, in: Proc. of ACM Conf. on Image and Video Retrieval, 2009, pp. 1–9.
[44] Z. Yu, J.S. Yoon, I.K. Lee, P. Venkatesh, J. Park, J. Yu, H.S. Park, HUMBI: a large multiview dataset of human body expressions and benchmark challenge, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit, 2020, pp. 2990–3000.
[45] A. Sharma, A. Kumar, H. Daume III, D.W. Jacobs, Generalized multiview analysis: a discriminative latent space, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit, 2012, pp. 2160–2167.
[46] S. Sun, X. Xie, M. Yang, Multiview uncorrelated discriminant analysis, IEEE Trans. Cybern. 46 (12) (2016) 3272–3284.
[47] G. Andrew, R. Arora, J. Bilmes, K. Livescu, Deep canonical correlation analysis, in: Proc. Int. Conf. Mach. Learn., 2013, pp. 1247–1255.
[48] W. Wang, R. Arora, K. Bilmes, J. Bilmes, On deep multi-view representation learning, in: Proc. Int. Conf. Mach. Learn., 2015, pp. 1083–1092.