A Novel Robust Low-rank Multi-view Diversity Optimization Model with Adaptive-Weighting Based Manifold Learning

Junpeng Tan [1], Zhijing Yang [1, *], Jinchang Ren[2], Bing Wang [3], Yongqiang Cheng [3], Wing-Kuen Ling [1]

[1] School of Information Engineering, Guangdong University of Technology, Guangzhou, 510006, China

[2] National Subsea Centre, Robert Gordon University, Aberdeen, AB10 7RQ, U.K.

[3] Department of Computer Science and Technology, University of Hull, Hull, HU6 7RX, UK.

[*]Corresponding author. Tel.: +86-20-39322438; Fax: +86-20-39322253. E-mail address: yzhj@gdut.edu.cn

Abstract

Multi-view clustering has become a hot yet challenging topic, due mainly to the independence of and information complementarity between different views. Although good results are achieved to a certain extent from typical methods including multi-view based k-means clustering, sparse cooperative representation clustering and subspace clustering, they still suffer from several drawbacks or limitations: (1) When each view is sparse decomposed, it still contains some hidden information for mining, such as the structure of samples, the intra-class similarity measure, and the inter-class diversity discrimination, etc. (2) Most of the existing multi-view methods only consider the local features within each view, but fail to effectively balance the importance of and combine information among different views in a diversified way. To tackle these issues, we propose a novel multi-view diversity learning model based on robust bilinear error decomposition (BED). The BED term with a low rank sparse constraint is an improved non-negative matrix factorization (NMF), which is used to extract the hidden structure information in sparse decomposition and useful diversity discrimination information in error matrix. The preservation of local features and selection of important views are achieved by adaptive weighted manifold learning. Furthermore, the Hilbert Schmidt independence criterion is used as a diversity learning term for mutual learning and fusion among views. Finally, the proposed robust low-rank multi-view diversity learning spectral clustering method is evaluated and benchmarked with eight state-of-the-art methods.

Experiments in six real datasets have fully validated the significantly improved accuracy and efficiency of the proposed methodology for effective clustering of multi-view images.

Keywords: Low-rank Representation (LRR), Multi-view Subspace Clustering (MVSC), Hilbert Schmidt Independence Criterion (HSIC), Non-negative Matrix Factorization (NMF), Adaptive-Weighting Manifold Learning (AWML).

1 Introduction

With the rapid development of information technology, it is difficult to meet increasing demands with solely single-view data clustering methods based on clustering data from a singular perspective, especially in analyzing datasets with complex relationships [1]. Therefore, it is essential to extract effective information from different views, which can then be combined for data fusion. Here, we define multi-view data as multiple features extracted from the samples using different techniques. By fusing the most diversity of various features, the effectiveness of clustering can be significantly improved.

Multi-view clustering (MVC) has been widely applied for many practical applications, such as natural language processing [2], computer vision [3], big data [4] and biomedical information analysis [5]. Simultaneously, it can handle distinctive types of data including images [6], texts [7] and their combinations [8]. The core of multi-view data is to share features with the same structure among multiple views, so that data points can be uniformly partitioned according to multiple representations of different views. By fusing information from different views, MVC algorithms can achieve high accuracy of clustering rather than simply connecting features from different views, which seems often incomplete as different views describe various perspectives. Therefore, the major issue of multi-view data analysis is how to effectively integrate multiple features and explore the underlying structures [9]-[11], such as, adding related constraints to the conventional k-means approach [12].

In order to explore different information from multiple feature sets while revealing a consistent cluster structure of the dataset, Jiang et al [13] proposed a collaborative fuzzy MVC algorithm with differently weighed views. Zhang et al [15] proposed a k-means-based two-level weighted fusion MVC method, yet the structure of the view and the integration of different views were not taken into consideration. In order to solve the problem of retaining the internal structural features of each view, Wang et al [16] proposed a belief propagation-based MVC method, where clustering consistency between the clustering qualities and similarity between different views were used. This led to the realization of different viewpoints through the passing between single view and cross view, where information between views was directly integrated.

Many existing MVC methods employ graphs in which typically pre-calculated inputs were used to reveal the data distribution independently in each view. Manifold learning can better obtain the inherent geometric information and the natural discriminant information in a single view. Gao et al [43] proposed double graphs-based discriminant projections (DGDP), which designed the graph constructions to preserve the informative globality and locality. Gao et al [44] also proposed the discriminant global and local preservation graph embedding (DGLPGE) to weight the edges of graphs. Cai et al used manifold learning to smooth the data and preserve the local structural features of the graph in MVC [17], i.e. the internal features of each view. As to MVC, most of them seldom consider the correlation between the internal graph structure of the view and the clustering results, with the results depending largely on the quality of the pre-defined affinity graph [18]. In practical applications, manifold learning is often used as an auxiliary part, which is combined with some other techniques to complete the tasks. Wang et al [19] proposed to combine manifold regularization with non-negative matrix factorization (NMF) for similarity measurement of regions within a view or between individual views, where a more compact multi-view data space representation through the regularization of views was obtained. NMF is essential for the sparse representation of the data, which has become a hot topic of research over the years. Based on the manifold regularization, we can also use a part-based NMF representation to maintain the local geometry of the data space.

Due to the widespread use of low-rank in single-view data compression [21], low-rank representation (LRR) has also been applied to MVC as it can improve clustering performance by exploring structural consistency among multi-views. In Ref. [22], a multi-view spectral clustering method was proposed for structural low-rank matrix decomposition. By decomposing potential low-dimensional data for clustering representations, a structured LRR was proposed to provide a high-quality description of the data's clustering structure for each view.

Another thing to note, high-dimensional multi-view datasets need to be projected onto a low-dimensional space to simplify the problem. The importance of different views should be taken into account. For instance, Yu et al [46] used the low-rank matrix representation to capture the global structure from the weighted multiple views. Zhang et al [14] proposed a multi-view cooperative local adaptive weighted information entropy of each view clustering method based on the Minkowski matrix. Zhan et al [20] proposed a novel multi-view document clustering method with the graph-regularized concept factorization, which was suitable for feature extraction and adaptive weighting of each view, enabling the preservation of the local structural features by using the manifold learning. This approach only considers the importance of the subspaces of the different views, but the importance of the internal structure of each view is not taken into account. Recently, Zhang et al [45] proposed kernelized multi-view subspace clustering (MVSC) via auto-weighted graph learning (KMSC-AGL) to evaluate the importance of multiple views. However, the update of weights in this method depends on the derivation of the common similarity matrix, showing a lack of mathematical meaning for weights that will be solved in this paper.

Although the importance of different views was considered to tackle the problems above, this method did not fully consider the diversity between views. Rather than focusing purely on the similarity between different views, it is also critical to use the diversity of different views. Consequently, to make full use of all view information an ideal solution needs to solve the following three key challenges: 1) extraction of hidden information in the noise part of the low-rank decomposition; 2) adaptive trade-off between local feature based structure consistency of different views; and 3) the diversity fusion between different views.

To tackle the aforementioned challenges in MVC, this paper proposes a new unified robust low-rank multi-view diversity optimization model (RLMDOM), which features a novel bilinear error matrix decomposition (BEMD) module, an adaptive-weighting based manifold learning (AWML) module, and the Hilbert Schmidt Independence Criterion (HSIC) for inter-view learning. The structural block diagram of the proposed RLMDOM with the AWML is shown in Fig. 1. In the proposed method, the BEMD module is an improved NMF, where robust low-rank constraints are added to the relevant error matrices. In this way, the robustness and sparsity of matrix decomposition are improved, such that the internal structure information of the sparse matrix and some useful discriminant information in the error matrix can be fully explored. Subsequently, the addition of AWML can ensure the local geometric structure inside each view to a large extent, and the discriminant information of data will not be seriously degraded when the input data is sparsely decomposed. Moreover, the structure of each view is applied with adaptive weighting, so that the importance of different view structures can be weighed and changed during the algorithm updates. To further enhance mutual learning and diversity fusion among different views, we added the HSIC module to facilitate the interconnection between views, the mutual learning and integration of information during the model optimization. Experimental results have shown that the proposed RLMDOM is a promising diversity optimization model with a superior performance than existing approaches.

The major contributions of the proposed RLMDOM algorithm can be summarized as follows:

(1) A novel BED model based on NMF is proposed. One error matrix is the sparse representation error of each view obtained by NMF. This can effectively remove any redundant information from the input data, whilst reducing unnecessary noise interference. Another error matrix is between the basis matrix obtained by NMF and the input data matrix of each view. This can help to extract the effective information by the sparse representation model more sufficiently. The novel bilinear error matrix decomposition can more effectively extract the hidden information in the noisy error part.

(2) The L_{21} norm and the nuclear norm are used to constrain the error matrix of NMF and the error matrix of the basis matrix, respectively. The L_{21} norm is to ensure that the coefficient matrix obtained by NMF has more complete information than the one obtained by linear decomposition. The nuclear norm that represents the sum of the singular values of the minimization matrix, belongs to the low-rank constraint with the effect of sparsity.

(3) In order to enhance the relationship between views and the diversity of views during matrix decomposition, two sub-modules are combined in the proposed model, including AWML and the HSIC, respectively. AWML takes into account contributions of the internal structures of different views during the process of continuous optimization. Furthermore, HSIC is used to make a variety of judgments on each view to enhance the connection between views;

(4) We effectively integrate BED, AWML and HSIC to construct a unified robust low-rank multi-view diversity optimization model. The internal structure information of sparse representation, the useful discriminant information of each view, and the diversity learning are fully considered.

Experimental results have demonstrated the superior performance of the proposed methodology in comparison to a few state-of-the-arts.

The remaining of this paper is organized as follows. Section 2 briefly introduces the background of related techniques, including NMF, manifold learning for data smoothing and HSIC covariance constraints. The proposed algorithm and its optimization process are presented in Section 3. Section 4 details the experimental results and analysis, and finally some concluding remarks are drawn in Section 5.

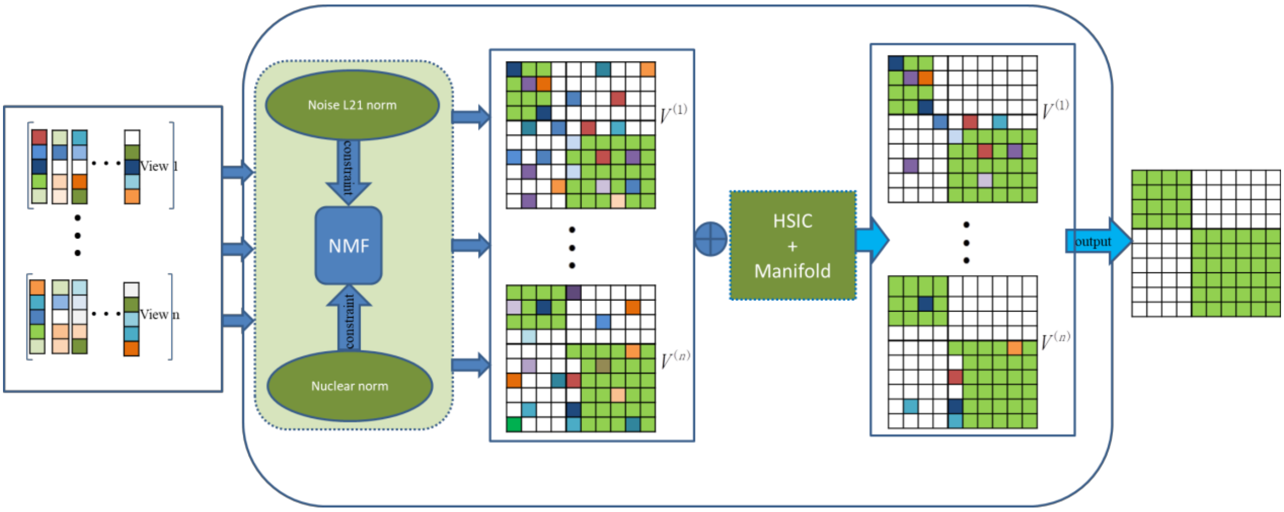

Figure 1: The structural block diagram of the proposed RLMDOM method.

2 Related Works

In this section, the fundamentals and background of the key techniques used in our proposed approach, including NMF, manifold regulation and HSIC are briefly introduced as follows.

2.1 Non-negative Matrix Factorization (NMF)

NMF [23] has been widely applied in LRR, sparse decomposition and cooperative representation [24] in recent years. In NMF, a given matrix X can be decomposed into two matrix forms, i.e. a basis matrix U and a coefficient matrix V, where all the elements in these two matrices are non-negative. That is

$$X \approx UV \tag{1}$$

where $X = [x_1, x_2, \cdots, x_n] \in R^{m \times n}$, m is the number of sample features, n is the number of input samples, and each column of X is a sample. The basis matrix $U = [u_{ij}] \in R^{m \times n}$ and the coefficient matrix $V = [v_{ij}] \in R^{n \times n}$ can be obtained.

The Euclidean distance is adopted to measure the objective function below:

$$O_1 = \|X - UV\|_F^2 = \sum_{i=1,j=1}^{i=m,j=n} (x_{ij} - u_{i,\cdot} * v_{\cdot,j})^2 \tag{2}$$

where * represents dot product; x_{ij} denotes the i-th feature of the j-th sample of the input data X; $u_{i,\cdot}$ and $v_{\cdot,j}$ denote the i-th row and j-th column, respectively. In fact, NMF is used as a sparse decomposition algorithm. With added constraints, the coefficient matrix can be used as an objective function of clustering detailed in the next subsections.

2.2 Manifold Regularization

Although NMF can generate a series of useful sparse matrices for a given data matrix, it does not take into account the integrity of the internal structural features of the data matrix. Recently, manifold learning has been applied in NMF to smooth the sparse subspace matrix [25]. In order to preserve the integrity of local features, the graph theory is used to construct the association of local features. It is found that this local similarity relationship can be constructed using the k-nearest neighbor based multi-label data clustering (ML-KNN) [38].

Three representative approaches of the general graph method based on KNN are given as follows:

1) **Binary representation**: In the constructed graph, each element is treated as a node. If the node j is within or on the edge of the nearest neighbor of the node i, W_{ij}=1; otherwise W_{ij}=0, where W_{ij} denotes the weight value between nodes i and j, that is, the similarity between these two samples.

2) **Heat kernel weighting**: If the node j is within the k-nearest neighbor of node i, we have

$$W_{ij} = e^{-\frac{\left\|x_j - x_i\right\|^2}{\sigma}}. \tag{3}$$

3) **Dot-product weighting**: If the node j is within the k-nearest neighbor of node i, we have

$$W_{ij} = \frac{x_i^T x_j}{max\ (X^T X)}. \tag{4}$$

Let $V_j = [v_{j1}, v_{j2}, \cdots, v_{jn}]^T$ denote the j-th column of the coefficient matrix. The Euclidean distance is used to calculate the deviation between the columns for measuring the smoothness of the low dimensional representation in multi-views. Here, the general manifold regularization O_2 can be defined as follows:

$$O_2 = \frac{1}{2} \sum_{i,j=1}^{n} \left\| V_i^{(v)} - V_j^{(v)} \right\|_F^2 W_{ij}^{(v)} \tag{5a}$$

$$= \sum_{i=1}^{n} (V_i^{(v)})^T V_i^{(v)} D_{ii}^{(v)} - \sum_{i,j=1}^{n} (V_i^{(v)})^T V_j^{(v)} \tag{5b}$$

$$= tr\left(V^{(v)} D^{(v)} (V^{(v)})^T\right) - tr\left(V^{(v)} W^{(v)} (V^{(v)})^T\right) \tag{5c}$$

$$= tr\left(V^{(v)} L^{(v)} (V^{(v)})^T\right) \tag{5d}$$

where tr denotes the matrices trace, $W_{ij}^{(v)}$ denotes the weight value between nodes i and j in the v-th view, $W^{(v)} \in R^{n \times n}$ is the similarity matrix, and $L^{(v)} = D^{(v)} - W^{(v)}$ is the v-th view Laplacian matrix, in which $D^{(v)}$ is a diagonal matrix with $D_{ii}^{(v)} = \sum_{j=1}^{n} W_{ij}^{(v)}$.

2.3 **The Hilbert Schmidt Independence Criterion**

According to Refs. [26][27], we may recall the definition of cross-covariance C_{xy}. A mapping $\emptyset(\mathrm{x})$ from a sample $x \in X$ to the kernel space Γ is defined, after that the inner product vector in the mapping space is represented as $K_1(x_i, x_j) = \langle \emptyset(x_i), \emptyset(x_j) \rangle$, where $\langle \cdot \rangle$ is the inner product function, and x_i, x_j

are the i, j-th column of the input data x, respectively. Then, $\varUpsilon$ is defined as the second kernel space, with a kernel function $K_2(y_i, y_j) = \langle\varphi(y_i), \varphi(y_j)\rangle$, where y_i and y_j are the i, j-th column of the input data y, respectively. As a result, the covariance function for two random variables, x and y, can be defined as:

$$C_{xy} = E_{xy}\left[(\emptyset(x) - \mu_x)\otimes(\varphi(y) - \mu_y)\right] \tag{6}$$

where μ_x and μ_y are the expectations of x and y, respectively, which can be obtained by $\mu_x = E(\emptyset(x))$ and $\mu_y = E(\varphi(y))$, and $\otimes$ denotes the matrix product.

Definition 1 (definition of HSIC): Given the separable reproducing kernel Hilbert space (RKHS) $\varGamma$, and a joint probability distribution ρ_{xy}, we can use the associated operator C_{xy} to identify the HSIC as the squared Hilbert-Schmidt norm:

$$\mathrm{HSIC}(\rho_{xy}, \varGamma, \varUpsilon) := \left\|C_{xy}\right\|_{HS}^2 \tag{7}$$

where the squared Hilbert-Schmidt norm is represented as follows:

$$\|X\|_{HS}^2 = \sum_{i,j} x_{ij}^2. \tag{8}$$

Definition 2: (The general form of HSIC): Given n independent observations from ρ_{xy}, $Z := \{(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)\} \in \varGamma \times \varUpsilon$, an estimator of HSIC, redefine $\mathrm{HSIC}(Z, \varGamma, \varUpsilon)$ as

$$\mathrm{HSIC}(Z, \varGamma, \varUpsilon) = (n-1)^{-2} tr(K_1 \mathrm{H} K_2 \mathrm{H}) \tag{9}$$

where K_1 and K_2 are the two inner product matrix $K_{1,ij} = K_1(x_i, x_j)$,$K_{2,ij} = K_2(x_i, x_j)$ and $h_{ij} = \delta_{ij} - 1/n$ is the center matrix. For more details of HSIC, please refer to Refs. [26][27].

3 The Proposed Method

As for MVC, a straightforward way to make use of all views is to perform clustering-based fusion of multi-view features. We propose the multi-view spectral clustering method based on NMF, called RLMDOM. This method can effectively decompose the input data, thus improving the computational efficiency of the algorithm. To preserve the local structure in MVC, we propose to obtain the latent representation in an adaptive way by using the AWML and the HSIC. The presented method is robust, as the adaptive weight of each view is used and the relevant robust constraints in each part are applied. Lastly, the spectral clustering method is utilized to classify datasets with the decomposing coefficient matrix.

3.1 Sparse Low-rank Subspace Clustering

According to the characteristics of NMF [23], a data matrix X can be decomposed into two non-negative matrices. As NMF is an approximate decomposition, a certain degree of errors exists. In general, these errors partially miss some useful information. Therefore, to ensure the integrity of the information provided by NMF, an error matrix E_x for the non-negative matrix is defined below:

$$X = UV + E_x. \tag{10}$$

The focus of the non-negative matrix is to use the corresponding object function and optimization method to minimize the decomposition error matrix. Actually, Eq. (2) contains two new matrices, i.e. the original data is decomposed into a basis matrix and a coefficient matrix, yet the decomposition in Eq. (2) is to use the original matrix as the basis matrix. There is a deviation in the basis matrix in the decomposition of the Eq. (2), and there is a noise term in the basis matrix in the decomposition of Ref. [27]. Therefore, in order to ensure that the coefficient matrix obtained by NMF has more complete information than the coefficient matrix obtained by linear decomposition, we add the error matrix and use the kernel approximation norm as $\|\cdot\|_* = \sum_i \delta_i$ and the L_{21} norm as $\|E_x\|_{2,1} = \sum_{i=1}^{n} \sqrt{\sum_{j=1}^{m} [E_x]_{i,j}^2} = \sum_{i=1}^{n} \left\| [E_x]_{i,:} \right\|_2$, where δ_i denotes the i-th singular value of the matrix. Here, the novel BEMD model based on NMF, denoted by O_3, can be defined as follows:

$$O_3 = \|E_x\|_{2,1} + \|E_u\|_*$$

$$s.t.\ \ E_x = X - UV, E_u = X - U, U \geq 0, V \geq 0 \tag{11}$$

where E_x represents the error of NMF. It is simultaneously with low-rank and sparsity, and can be used to make the decomposition more robust than other norms. By comparing with Eq. (2) and Ref. [27], more useful information can be extracted by our proposed method. E_u denotes the difference between the basis matrix of NMF and the original data. This will not only preserve the characteristics of the coefficient matrix, but also enhance the robustness of the model.

3.2 Model Construction

Before discussing the model construction, we firstly define some notations. The input multi-view datasets are represented as $\{X^{(1)},\cdots,X^{(v)},\cdots,X^{(n_v)}\}$, where $X^{(v)}$ represents the dataset from the v-th view and n_v denotes the number of total views. For each $X^{(v)}$, it has n instances, and each instance has m features, thus it can be presented by $X^{(v)} = \left\{x_1^{(v)}, x_2^{(v)}, \cdots, x_n^{(v)}\right\} \in \mathcal{R}^{m\times n}$. $U^{(v)} \in \mathcal{R}^{m\times n}$ and $V^{(v)} \in \mathcal{R}^{n\times n}$ represent the v-th view's basis matrix and coefficient matrix of NMF, respectively. $L^{(v)}$ is the Laplacian matrix for the v-th view with $E_x^{(v)}$ and $E_u^{(v)}$ denoting the error of the v-th view matrix $X^{(v)}$ and basis matrix $U^{(v)}$, respectively.

Based on the definitions above, the objective function of the proposed RLMDOM method can be given by:

$$\min_{U^{(v)}, V^{(v)}, E_x^{(v)}, E_u^{(v)}, \alpha^{(v)}} \sum_{v=1}^{n_v} \lambda_1 \left[\left\|E_x^{(v)}\right\|_{2,1} + \left\|E_u^{(v)}\right\|_*\right] + \lambda_v \sum_{w=1, v\neq w}^{n_v} HSIC\big(V^{(v)}, V^{(w)}\big)$$
$$+\sum_{v=1}^{n_v}\big(\alpha^{(v)}\big)^\gamma tr(V^{(v)}L^{(v)}(V^{(v)})^T) \tag{12}$$
$$\mathrm{s.t.}\, E_x^{(v)} = X^{(v)} - U^{(v)}V^{(v)}, E_u^{(v)} = X^{(v)} - U^{(v)};\ \sum_{v=1}^{n_v}(\alpha^{(v)})^\gamma = 1, 0 \leq \alpha^{(v)} \leq 1; U^{(v)} \geq 0, V^{(v)} \geq 0.$$

There are three regularization parameters in Eq. (12), where λ_1 is to measure the importance of sparse representation, $(\alpha^{(v)})^\gamma$ and λ_v are the trade-off correlations of the smooth term and the diverse regression terms, respectively. $\alpha^{(v)}$ and γ are the weight of v-th view and power index, respectively. In the first term, the improved NMF is used to decompose the original data matrix into the basis matrix, the coefficient matrix and two error matrices. A sparse sub-matrix can be obtained, which is sparser than the original data matrix with the redundant information discarded. If the feature dimension of each sample (m) is greater than the number of samples in the input dataset (n), i.e. $m > n$, the decomposition actually has an effect of dimension reduction. The L_{21} norm is used for more robust decomposition in the first term of the objective function. In order to ensure that the coefficient matrix obtained by NMF has more complete information than the one obtained by linear decomposition, the error matrix is added with using nuclear norm, which not only preserves the characteristics of the coefficient matrix, but also enhances the robustness of the model. $\|\cdot\|_*$ represents the sum of the singular values of the minimization matrix, which is different from other norms, and belongs to the low-rank constraint with the effect of sparse and dimensionality reduction [28]. In the second term, HSIC is used to fuse the structural features of various views, and also to determine the covariance between different views to enhance the connection between views. In addition, the differences between them are minimized through the related kernel space mapping. In the third term of the model, manifold learning is used to smooth each view whilst preserving its structural features. Actually, AWML can adaptively update the importance of different views according to the optimized results of the algorithm. The exponential form of AWML helps to enhance the diversity among different views for improving the discriminability in between. In comparison to other similar models, our proposed RLMDOM model mainly considers the deviation of matrix decomposition, whilst taking into account the importance of each view and the fusion of multi-view data for the following-on tasks of object detection and recognition.

3.3 Optimization

To solve the non-convex objective function in Eq. (12), the Augmented Lagrangian Method (ALM) [29] is adopted. The specific algorithm flow is summarized in Algorithm 1 and the details are explained as follows.

Firstly, the objective function can be rewritten as:

$$L\left(U^{(v)},V^{(v)},E_x^{(v)},E_u^{(v)},\alpha^{(v)}\right)=\sum_{v=1}^{n_v}\lambda_1\left[\left\|E_x^{(v)}\right\|_{2,1}+\left\|E_u^{(v)}\right\|_*\right]+\sum_{v=1}^{n_v}\left(\alpha^{(v)}\right)^\gamma tr(V^{(v)}L^{(v)}(V^{(v)})^T)$$

$$+\sum_{v=1}^{n_v}\emptyset\left(Q_1^{(v)},X^{(v)}-U^{(v)}V^{(v)}-E_x^{(v)}\right)$$

$$+\lambda_v\sum_{w=1,v\neq w}^{n_v}HSIC\left(V^{(v)},V^{(w)}\right)\ +\sum_{v=1}^{n_v}\emptyset\left(Q_2^{(v)},X^{(v)}-U^{(v)}-E_u^{(v)}\right)$$

$$\text{s.t.}\sum_{v=1}^{n_v}(\alpha^{(v)})^\gamma=1,0\leq\alpha^{(v)}\leq 1;U^{(v)}\geq 0,V^{(v)}\geq 0 \tag{13}$$

where $\emptyset(J,K)=\langle J,K\rangle+\frac{\mu}{2}\|K\|_F^2$, $\langle.,.\rangle$ is the matrix inner product, and $Q_1^{(v)}$ and $Q_2^{(v)}$ are related to the Lagrangian multipliers. Furthermore, the parameter μ is a regularity coefficient. In this paper, the inner product kernel is used for HSIC, which is defined as $K^{(v)}=(V^{(v)})^TV^{(v)}$. For notation convenience, HSIC is represented by

$$\sum_{w=1,w\neq v}^{n_v}HSIC\left(V^{(v)},V^{(w)}\right)=\sum_{w=1,w\neq v}^{n_v}tr\left(HK^{(v)}HK^{(w)}\right)$$

$$=\sum_{w=1,w\neq v}^{n_v}tr\left(V^{(v)}HK^{(w)}H(V^{(v)})^T\right)=tr(V^{(v)}K(V^{(v)})^T) \tag{14a}$$

$$K=\sum_{w=1,v\neq w}^{n_v}HK^{(w)}H,h_{ij}=\delta_{ij}-\frac{1}{n}. \tag{14b}$$

Based on the analysis above, an alternate optimization method can be derived in order to convert it to a few solvable sub-problems. In every round of parameter updating, the solution of one variable is obtained whilst all others are fixed.

1) The updating of the variable $V^{(v)}$: With fixed variables $U^{(v)}, E_x^{(v)}, E_u^{(v)}, (\alpha^{(v)})^\gamma$ to update $V^{(v)}$, the optimization problem of Eq. (13) turns into

$$\arg min \sum_{v=1}^{n_v} (\alpha^{(v)})^\gamma \, tr\big(V^{(v)} L^{(v)} (V^{(v)})^T\big) + \sum_{v=1}^{n_v} \langle Q_1^{(v)}, X^{(V)} - U^{(V)} V^{(V)} - E_x^{(v)} \rangle + \lambda \, V^{(v)}$$

$$+\lambda_v \sum_{w=1, v \neq w}^{n_v} HSIC\big(V^{(v)}, V^{(w)}\big) \; + \sum_{v=1}^{n_v} \frac{\mu}{2} \left\| X^{(V)} - U^{(V)} V^{(V)} - E_x^{(v)} \right\|_F^2. \qquad (15)$$

where λ is the Lagrangian coefficient. We take the derivative to Eq. (15) about $V^{(v)}$ and let it be 0. According to the Karush-Kuhn-Tucker condition [31], it can be rewritten as:

$$\mathrm{A} V^{(v)} + V^{(v)} B = C \qquad (16a)$$

$$A = \mu (U^{(v)})^T U^{(v)} + \lambda, B = \alpha^{(v)} L^{(v)} + \lambda_v K \qquad (16b)$$

$$C = \mu (U^{(v)})^T X^{(v)} + \mu (U^{(v)})^T E_x^{(v)} - (U^{(v)})^T Q_1^{(v)}. \qquad (16c)$$

Note that Eq. (16a) is the standard Sylvester equation [30][42]. As the matrix A and $-B$ do not have common eigenvalues, according to the following Theorem 1, Eq. (16a) has a standard solution.

Theorem 1: If the equation $AX + XB = C$ is a standard Sylvester equation, it has a standard solution for X, as $X = QX^* R^T$, provided that the matrix A and $-B$ have no common eigenvalues, where X^* is the solution to the following equation,

$$A_{kk}^* X_{kl}^* + X_{kl}^* B_{ll}^* = C_{kl}^* - \sum_{j=1}^{k-1} A_{kj}^* \, X_{jl}^* - \sum_{i=1}^{l-1} X_{ki}^* \, B_{il}^*$$

$$(k = 1,2, \cdots, p; l = 1,2, \cdots, q) \qquad (17)$$

where $A^\star = Q^T A Q, B^\star = R^T B R,\ R \in \mathcal{R}^{n \times q}$ are the orthogonal similarity transformation matrix, and the matrix A and the matrix B are reduced to lower and upper real Schur form; $C^\star = Q^T C R,\ Q \in \mathcal{R}^{n \times p}$ and, where p and q are the reduced target dimension of $A^\star \in \mathcal{R}^{p \times p}$ and $B^\star \in \mathcal{R}^{q \times q}$, respectively. For more details please refer to Ref. [42].

2) The updating of $U^{(v)}$: To update the parameter $U^{(v)}$, the above method is adopted by fixing other variables except $U^{(v)}$, where the loss function of Eq. (13) can be rewritten as:

$$\arg min \sum_{v=1}^{n_v} \frac{\mu}{2} \left\| X^{(v)} - U^{(v)} - E_u^{(v)} \right\|_F^2 + \sum_{v=1}^{n_v} \langle Q_1^{(v)}, X^{(v)} - U^{(v)} - E_u^{(v)} \rangle + \lambda U^{(v)}$$

$$+ \sum_{v=1}^{n_v} \frac{\mu}{2} \left\| X^{(v)} - U^{(v)} V^{(v)} - E_x^{(v)} \right\|_F^2 + \sum_{v=1}^{n_v} \langle Q_2^{(v)}, X^{(v)} - U^{(v)} V^{(v)} - E_x^{(v)} \rangle. \quad (18)$$

Similarly, we take a derivative to Eq. (18) about $U^{(v)}$ and let it be 0. Again, according to the Karush-Kuhn-Tucker condition [31], the optimization solution of $U^{(v)}$ is derived as:

$$U^{(v)} = A * inv(B) \quad (19a)$$

$$A = \mu X^{(v)} (V^{(v)})^T + Q_1^{(v)} V^{(v)} + \mu X^{(v)} + Q_2^{(v)} - \mu E_x^{(v)} (V^{(v)})^T - \frac{\mu}{2} E_x^{(v)} \quad (19b)$$

$$B = \mu V^{(v)} (V^{(v)})^T. \quad (19c)$$

Here, B can be regarded as the positive definite matrix, hence B will have an inverse matrix.

3) The updating of $\alpha^{(v)}$: According to Eq. (13) and the relative constraint, the sub-problem of $\alpha^{(v)}$ can be derived as:

$$\arg\min \sum_{v=1}^{n_v} (\alpha^{(v)})^\gamma tr(V^{(v)} L^{(v)} (V^{(v)})^T) \quad (20)$$

$$\text{s.t.} \sum_{v=1}^{n_v} (\alpha^{(v)})^\gamma = 1.$$

Eq. (20) can be solved by using the Lagrange function method as follows：

$$\mathrm{L}\big(\alpha^{(v)}\big) = \sum_{v=1}^{n_v} (\alpha^{(v)})^\gamma tr(V^{(v)} L^{(v)} \big(V^{(v)}\big)^T\big) - \lambda\big(\sum_{v=1}^{n_v} (\alpha^{(v)})^\gamma - 1\big). \quad (21)$$

With the fixed parameter γ, partial derivation is able to be taken to $\alpha^{(v)}$ and set it to be 0, thus we have:

$$S = tr(V^{(v)} L^{(v)} (V^{(v)})^T) \quad (22a)$$

$$\frac{\partial L}{\partial \alpha^{(v)}} = \gamma (\alpha^{(v)})^{\gamma-1} (S - \lambda) \quad (22b)$$

$$\alpha^{(v)} = (\frac{\lambda}{S})^{\frac{1}{\gamma-1}}. \tag{22c}$$

4) The updating of $E_x^{(v)}$: To update $E_x^{(v)}$, the original objective function can be rewritten as:

$$\arg min \frac{\lambda_1}{\mu} \left\| E_x^{(v)} \right\|_{2,1} + \frac{1}{2} \left\| E_x^{(v)} - (X^{(v)} - U^{(v)} V^{(v)} - \frac{Q_1^{(v)}}{\mu}) \right\|_F^2. \tag{23}$$

One can see that this problem has a closed form solution [21][32]. Let $B = X^{(v)} - U^{(v)}V^{(v)} - Q_1^{(v)}/\mu$, we can then update $E_x^{(v)}$ by:

$$E_{ij}^{(v)} = \begin{cases} (1 - \frac{\lambda_1}{\mu \|B_j\|_2}) B_j & if \|B_j\|_2 \geq \frac{\lambda_1}{\mu} \\ 0 & oterwise \end{cases} \tag{24}$$

where $E_{ij}^{(v)}$ is the (i,j)-th element of $E_x^{(v)}$, and B_j is the j-th column of B. For a large dataset, the following normalization method is applied to $E_x^{(v)}$ for greater robustness and smoothness:

$$E_x^{(v)} = \frac{E_x^{(v)}}{\left\| E_x^{(v)} \right\|_F^2} \ . \tag{25}$$

5) The updating of $E_u^{(v)}$: To update $E_u^{(v)}$, the original objective function can be rewritten as:

$$\arg min \frac{\lambda_1}{\mu} \left\| E_u^{(v)} \right\|_* + \frac{1}{2} \left\| E_u^{(v)} - (X^{(v)} - U^{(v)} + \frac{Q_2^{(v)}}{\mu}) \right\|_F^2. \tag{26}$$

Let $C = X^{(v)} - U^{(v)} + \frac{Q_2^{(v)}}{\mu}$, and this can be solved by the Singular Value Threshold (SVT) method [33], where $U \sum V^T$ may become the standard SVT of the matrix C. Therefore, the solution of the above formula can be derived as:

$$E_u^{(v)} = U S_{\frac{\lambda_1}{\mu}}(\sum) V^T, \tag{27a}$$

$$S_\delta(X) = max(X - \delta, 0) + min\ (X + \delta, 0) \tag{27b}$$

where $S_\delta(X)$ is the shrinkage operator. To ensure that the variable $E_u^{(v)}$ does not overflow during the entire optimization process, we also apply the following normalization method to $E_u^{(v)}$ for greater robustness and smoothness:

$$E_u^{(v)} = \frac{E_u^{(v)}}{\left\|E_u^{(v)}\right\|_F^2} \ . \tag{28}$$

6) The updating of ALM parameters: These parameters can be updated as follows:

$$Q_1^{(v)} = Q_1^{(v)} + \mu\left(X^{(v)} - U^{(v)}V^{(v)} - E_x^{(v)}\right) \tag{29a}$$

$$Q_2^{(v)} = Q_2^{(v)} + \mu\left(X^{(v)} - U^{(v)} - E_u^{(v)}\right) \tag{29b}$$

$$\mu = \rho\mu \tag{29c}$$

where ρ is a constant parameter.

Algorithm 1: The algorithm steps for solving RLMDOM

```
Input: Unlabeled datasets $X = \{X^{(v)}\}_{v=1}^{n_v}, k, k_v, \lambda_1, \lambda_v, \gamma, \{(\alpha^{(v)})^\gamma\}_{v=1}^{n_v}, \mu, \rho$.
Output: The features of data points grouped in $k$ clusters.
Initialize:$k = 0, V^{(v)} = 0, U^{(v)} = 0, E_x^{(v)} = 0, E_u^{(v)} = 0, Z = 0, Q_1^{(v)} = 0, Q_2^{(v)} = 0, t = 0$.
While $t \leq$ max _iter do
    for $v$=1 to $n_v$ do
        $k = 0$
        for w=1 to $n_v$ do
            Update $k_v$ by Eq. (14b).
            if $v \neq w$ do
                $k = k + k_v$.
            end if
        end for
        Fix other parameters and update $V^{(v)}$ by solving Eq. (17).
        Fix other parameters and update $U^{(v)}$ by solving Eq. (19a).
        Fix other parameters and update $\alpha^{(v)}$ by solving Eq. (22c).
        Fix other parameters and update $E_x^{(v)}$ by solving Eq. (24) and Eq. (25).
        Fix other parameters and update $E_u^{(v)}$ by solving Eq. (27a) and Eq. (28).
        Update $Q_1^{(v)}$ by solving Eq. (29a).
        Update $Q_2^{(v)}$ by solving Eq. (29b).
        Update $\mu$ by solving Eq. (29c).
    end for
    Determine whether the convergence condition is met.
    $t = t + 1$.
end while
for $v$=1 to $n_v$ do
    $Z = Z + (|V^{(v)}| + |V^{(v)}|^T)/2$.
end for
Apply spectral clustering to the affinity matrix $Z$.
```

4 Experiments Results and Analysis

4.1 Experimental Setting

In this section, comprehensive experiments on six well-known MVC datasets are used for performance assessment of the proposed approach. These datasets span various applications, such as news, texts, and facial images/videos under various conditions, which are widely used in MVC field. The general information and statistics about these six datasets are summarized in Table 1, and the experimental settings are detailed as follows.

(1) Notting-Hill Video Face dataset [34]: This is a face clustering dataset which contains 4660 faces of 5 main casts in 76 video sequences. Our experiments used three views, including 6750-D Gabor features, 3304-D LBP features and the 2000-D gray features. As suggested in Ref. [34], the first 1206 samples of the dataset are used in our experiment.

(2) 3-Sources dataset [36]: Covering three online news sources, i.e. BBC, Reuters, and the Guardian, this dataset contains 416 cases. As suggested in Ref. [36], we used the three views with 169 distinct pieces of news from each view.

(3) ORL_mtv dataset [28]: The ORL_mtv dataset includes 10 different gray scale face images of 40 distinct subjects. With selected subjects, the images are taken under different conditions such as: varied lighting, facial expressions and details.

(4) COIL20 dataset [10]: It contains 1440 samples in total, each of which is a 32x32 gray scale image from 20 objects captured from different view angles, including three views.

(5) FERET dataset [35]: This dataset contains 1400 samples, each of which is 80x80 gray scale image from 200 objects captured from different view angles.

(6) Reuter dataset [36]: This is a document dataset containing features derived from 5 languages and translations over a common set of six categories, with all files being in the text bag representative. The original document language is English, with 4 other views in French, German, Spanish and Italian translation. For each class, 100 samples are randomly selected, resulting in a dataset of 600 documents.

Table 1: Summary of the six datasets used in our experiments

Dataset	Notting-Hill	3-Sources	ORL_mtv	COIL20	FERET	Reuter
Samples	1206	169	400	1440	1400	600
Views	3	3	3	3	1	5
Clusters	20	6	40	20	200	6

4.2 Compared Methods

In this section, we will compare the proposed RLMDOM method with the existing well-known MVSC algorithms, including the Graph regularized multi-view NMF (GMVNMF) [41], Exclusivity-consistency regularized MVSC (ECRMSC) [10], centroid-based multi-view low-rank sparse subspace clustering (CMLRSSC) [36], pairwise kernel multi-view low-rank sparse subspace clustering (PKMLRSSC) [36], diversity-induced MVSC (DiMSC) [27], latent MVSC (LMSC) [28], MVSC via co-training robust data representation (CoMSC) [47], and consensus one-step MVSC (COMVSC) [48]. Details of these benchmarking approaches are briefly introduced as follows.

1) **GMVNMF** [41]: This is mainly composed of two terms in the loss function, with NMF and weighting graph regularized.

2) **ECRMSC** [10]: This method mainly uses a large number of L_1 norm to constrain the components of sparse representation.

3) **CMLRSSC** [36]: Similar to pairwise MLRSSC, the only difference is that the view-specific representations are enforced towards a common centroid.

4) **PKMLRSSC** [36]: Aiming to recover the non-linear subspace, PKMLRSSC is similar to the first two algorithms yet the original dataset is affine mapped into a high dimensional feature space.

5) **DiMSC** [27]: This method extends the existing subspace clustering by joining the HSIC for preserving the relevant information between different views.

6) **LMSC** [28]: Different from the general latent subspace clustering algorithms, LMSC uses two linear representations to derive the final sparse matrix which has greatly reduced the dimension of the data.

7) **CoMSC** [47]: A method that utilizes multi-kernel spaces to process redundant information, and constructs consensus self-representations by exploring complementary details of learned representations.

8) **COMVSC** [48]: This is a unified MVSC framework, which jointly optimizes similarity learning, clustering partition and final clustering labels.

4.3 Evaluation Criteria

For quantitative performance evaluation, the following seven well-known evaluation criteria are used. They are the clustering accuracy (ACC) [27], normalized information (NMI) [37], precision [36], recall [36], F-score [36], adjusted rand index (ARI) and confusion matrix, respectively.

4.4 Parameter Analysis

There are several parameters in the objective function of RLMDOM, which includes the regularization parameters for NMF, manifold learning, HSIC and the Lagrangian regularization, i.e.

$\lambda_1, (\alpha^{(v)})^{\gamma}$, λ_v, μ and ρ, respectively. In this section, the effect of these parameters on the proposed model will be discussed.

The first important parameter is the sparse parameter λ_1, which plays a crucial role in the sparse decomposition of the original data and directly affects the clustering performance of the entire algorithm. As shown in Fig. 2, the magnitudes of the fluctuations in the six datasets are quite stable under various λ_1, which means the proposed approach is insensitive to λ_1. To this end, we set λ_1= 0.001 by unified verification of different datasets in the experiment.

The second important parameter $(\alpha^{(v)})^{\gamma}$ has a great effect on the weight of different views when preserving their structural integrity. A larger weight indicates that the view contains more information thus it is more important. Although the parameter $\alpha^{(v)}$ can be adaptively determined by Eq. (23), the parameter γ has to be manually set. Actually, the initial value of γ is empirically set to 1. If someone wants to obtain the best performance of clustering, the parameter of power index γ can be adjusted in $[1.1,1.2,\cdots,2]$, and the initial value of $\alpha^{(v)}$ is set to 0.3.

For the parameter λ_v, it ensures the diversity of different subspace representations based on HSIC. The larger the parameter is, the more important the common feature structure between views is. As shown in Fig. 3, λ_v is set to 0.001 for all the datasets.

The last two important parameters are the regulation parameters μ and ρ, which are hard to be determined empirically. Here, the initial value of μ can be obtained in a ranging from [1*e*-2, 1*e*4]. Actually, the initial parameter μ is empirically set to $1e1$, and ρ is set to 1.2. The effects of these two parameters are shown in Fig. 4, where the parameter μ is shown in exponential scale.

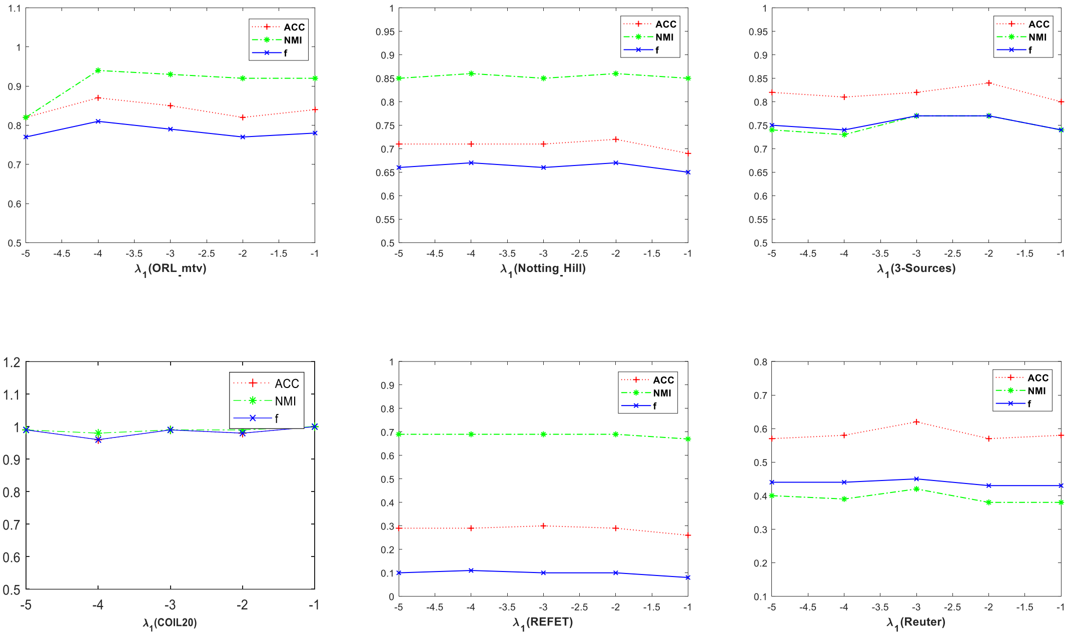

Figure 2: ACC, NMI and *f* are presented respectively according to the data decomposition sparse representation parameter λ_1 in six datasets.

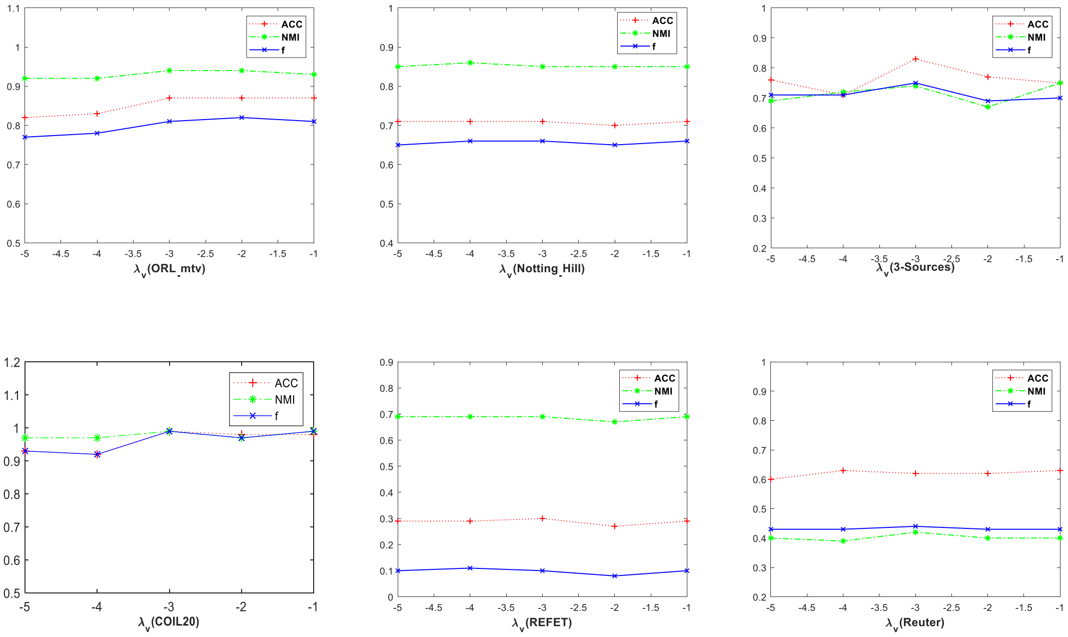

Figure 3: ACC, NMI and *f* are presented respectively according to the importance of HSIC parameter λ_v in six datasets.

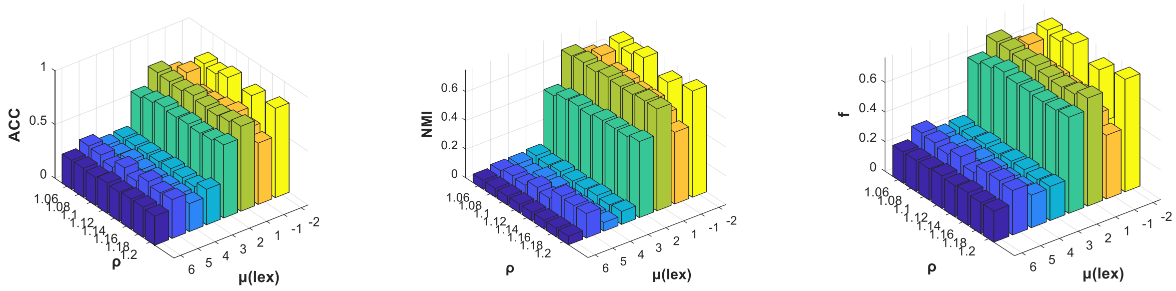

Figure 4: ACC, NMI and *f* respectively of the regulation parameter μ and parameter ρ in the 3-Sources dataset. The results are similar in other five datasets.

4.5 Clustering Results

In this section, the clustering results on the six datasets are presented and shown in Tables 2-7, including those from our method and eight benchmarking algorithms. As seen, the proposed method has consistently outperformed all its peers in terms of all the six evaluation criteria. Detailed analysis of these results is presented as follows.

(1) For all six datasets excluding FERET, our approach has yielded significantly improved results. Firstly, we can analyze the two clustering indicators ACC and NMI of all experiment results. More specifically, the average of ACC of our method is higher than the second best algorithm by 2.1%, 3.7%, 11%, 0.2%, 5% and 0.6% on the ORL_mtv, 3-Sources, COIL20, Notting-Hill, Reuters and FERET, respectively. Besides, the average of NMI of our method is higher than the second best algorithm by 0.8%, 6%, 3%, 2%, 4% and 0.6% on the ORL_mtv, 3-Sources, COIL20, Notting-Hill, Reuters and FERET, respectively. The clustering results for dataset on ORL_mtv, 3-Source, COIL20, Notting-Hill, Reuters and FERET are shown in Tables 2-7, which show the great improvement of our algorithm compared to other algorithms.

(2) We analyze the experimental results from the image dataset. Tables 2, 5 and 7 show the face clustering results on ORL_mtv, Notting-Hill and FERET. Table 4 is the results on the multi-target dataset

COIL20. Here, we divide the datasets into two groups, i.e. multi-view datasets (ORL_mtv, COIL20 and Notting-Hill) and single-view dataset (FERET), respectively. As can be seen from these tables, the experimental results of our proposed algorithm for multi-view face datasets are very good. In particular, from the Notting-Hill dataset, it can be clearly seen that the lowest improvement of the evaluation index is approximately 1% compared to the other method (such as ECRMSC and COMVSC for ACC). As to the dataset FERET, the improvement of experimental results is not so obvious compared to CoMSC and COMVSC, which also take into account the mapping of the kernel space, the fusion of diversity, and the preservation of geometric structures. Especially, the comparison method COMVSC also utilizes the optimal similarity learning, clustering partition and final clustering labels. Therefore, our proposed method experimental result is only 0.6%, 0.7%, 0.1%, 0.2%, 0.5%, 0.1% better than the second best algorithm COMVSC in ACC, NMI, F-score, recall, Precision and ARI. From the experimental results in Table 7, we can analyze from two aspects on the reasons of low experimental results, including other comparison algorithms (GMVNMF, CMLRSSC, PKMLRSSC, ECRMSC, DiMSC, LMSC, CoMSC and COMVSC). First, the dataset FERET is a single-view dataset. In order to ensure that the proposed algorithm is not only applicable to multi-view datasets, but also to single-view datasets (FERET), for this dataset we do not use multi-view features like other datasets. Second, to illustrate the generality of our proposed algorithm, we did not use the different feature extraction methods to extract features from the original image, like other datasets ORL_mtv and Notting-Hill. The dataset without feature extraction has a lot of noise information when it is integrated into the matrix, which will affect the cluster performance to some extent. Therefore, the experiment result on dataset FERET is not so good compared to other datasets. As to multi-view dataset COIL20, it shows that the experimental results are very good with great improvement compared to other algorithms. According to the experimental results of the second best algorithm CMLRSSC, there are 11%, 3%, 13%, 3%, 21% and 14% improvement on the six evaluation metrics ACC, NMI, F-score, Recall, Precision and ARI, respectively. Therefore, the proposed algorithm has good clustering performance on face datasets, no matter it is multi-view feature space or single-view feature space.

(3) We analyze the experimental results from the text dataset. Table 3 and Table 6 show the clustering results on the text datasets 3-Sources and Reuters, respectively. As can be seen, our proposed algorithm has obvious improvement compared to other comparison algorithms. ECRMSC is the second best algorithm on the text dataset 3-Sources, and our propose algorithm is much better than that. The biggest improvement of the evaluation metrics is ARI, which is 13% higher than ECRMSC, while the lowest is 3% for ACC. As can be seen from Table 3, the proposed method has greatly improved in various indicators compared with CMLRSSC and PKMLRSSC. That is because different norm constraints are used. Another difference between our method and the CMLRSSC and PKMLRSSC algorithms are that the manifold learning method is used to preserve the local structural features of the view, and ensure the effectiveness of each view. Table 6 shows the clustering results on another text dataset Reuters. We can see that the proposed algorithm still has obvious improvement from the comparison algorithms with different evaluation indicators. For example, the lowest improvement evaluation index is 3.8% in NMI. However, the clustering performance evaluation index recall is lower than the comparison algorithm DiMSC, CoMSC and COMVSC. The reason is that the text data classification context is closely related, and we filter some negative information when performing NMF on the text dataset. However, the clustering results of GMVNMF, CMLRSSC and PKMLRSSC in this dataset are much worse than our proposed method. That is because our method pays attention to the deviation of matrix decomposition and adds the AWML module. Besides, compared with the DiMSC and COMVSC methods, when the original data matrix is decomposed, the basis matrix may contain noise which in turn affects the clustering performance to some extents.

Table 2: Results (mean ± standard deviation) on the ORL_mtv dataset (best results in bold)

Method	ACC	NMI	F-score	Recall	Precision	ARI
GMVNMF [41]	72.70±0.40	89.10±0.10	66.00±0.44	78.37±1.53	57.90±0.65	65.00±0.46
ECRMSC [10]	85.41±1.10	94.70±0.90	82.10±1.50	84.72±1.21	78.30±0.82	81.01±1.12
CMLRSSC [36]	68.80±3.04	83.67±1.33	58.67±3.45	63.41±2.81	5466±4.21	57.64±3.55
PKMLRSSC [36]	76.70±2.83	89.46±1.46	70.30±3.41	75.03±3.15	66.17±3.93	69.57±3.49
DiMSC [27]	83.80±0.10	94.00±0.30	80.20±0.70	85.60±0.40	76.40±1.20	80.70±0.20
LMSC [28]	81.94±1.71	93.10±1.16	74.83±0.90	78.94±0.43	71.12±0.43	74.22±0.46
CoMSC [47]	76.75±1.25	87.22±0.75	68.31±1.54	71.11±1.47	64.75±1.85	67.31±1.89
COMVSC [48]	79.25±0.45	90.59±0.14	73.34±0.87	79.57±0.23	68.03±1.10	72.68±0.48
Ours	**87.55±1.45**	**95.56±2.78**	**83.17±0.82**	**85.75±1.34**	**82.92±1.45**	**82.34±1.34**

Table 3: Results (mean ± standard deviation) on 3-Sources dataset (best results in bold)

Method	ACC	NMI	F-score	Recall	Precision	ARI
GMVNMF [41]	55.42±3.05	49.60±2.29	48.46±2.52	51.42±2.61	45.35±2.37	32.23±1.51
ECRMSC [10]	80.47±0.00	70.27±0.00	68.94±0.00	62.75±0.00	76.50±0.00	60.71±0.00
CMLRSSC [36]	65.80±6.01	58.31±2.95	62.09±5.85	58.31±7.63	66.81±4.87	51.72±6.92
PKMLRSSC [36]	60.47±3.60	52.44±2.18	54.60±3.74	49.35±4.19	61.23±3.82	42.74±4.51
DiMSC [27]	68.27±0.45	60.52±0.61	59.95±0.83	53.51±0.53	66.47±0.82	48.49±0.83
LMSC [28]	71.60±0.10	68.37±1.32	65.58±0.23	58.62±0.52	74.41±0.30	56.72±0.42
CoMSC [47]	71.60±4.85	60.57±3.54	66.71±5.89	61.08±3.57	73.49±2.89	57.81±4.69
COMVSC [48]	71.60±3.58	60.43±4.19	66.86±4.56	67.66±3.96	66.08±2.95	56.67±5.45
Ours	**84.20±7.60**	**76.49±4.33**	**79.77±5.66**	**75.10±6.69**	**85.17±4.67**	**74.18±7.03**

Table 4: Results (mean ± standard deviation) on COIL20 dataset (best results in bold)

Method	ACC	NMI	F-score	Recall	Precision	ARI
GMVNMF [41]	81.01±2.20	94.05±1.42	81.00±2.30	92.57±0.56	72.35±3.72	79.90±2.50
ECRMSC [10]	77.71±2.22	93.26±1.02	78.47±2.45	91.25±1.57	68.88±1.09	77.19±1.18
CMLRSSC [36]	87.99±0.32	96.42±0.10	85.69±0.25	96.42±0.13	77.18±0.33	84.86±0.26
PKMLRSSC [36]	79.72±2.52	85.00±1.97	76.09±2.23	77.27±2.31	74.95±2.16	74.83±2.49
DiMSC [27]	70.81±2.36	80.95±2.04	65.43±1.85	67.10±1.42	63.99±1.42	64.55±1.25
LMSC [28]	80.10±0.70	87.90±0.11	75.80±0.70	83.20±0.40	78.30±0.80	82.10±1.70
CoMSC [47]	67.91±1.56	75.99±0.84	62.20±2.58	64.04±1.49	60.46±2.68	60.18±2.89
COMVSC [48]	65.76±1.56	79.59±0.53	61.39±1.24	68.57±0.12	55.57±1.12	59.17±1.58
Ours	**99.44±0.52**	**99.46±0.36**	**99.06±0.42**	**99.11±0.33**	**99.02±0.48**	**99.01±0.61**

Table 5: Results (mean ± standard deviation) on Notting-Hill dataset (best results in bold)

Method	ACC	NMI	F-score	Recall	Precision	ARI
GMVNMF [41]	67.64±4.65	82.10±1.02	54.78±1.36	67.09±2.24	46.45±3.50	49.82±2.83
ECRMSC [10]	70.07±2.45	82.52±1.96	58.21±2.58	57.84±1.29	58.57±1.07	54.97±1.28
CMLRSSC [36]	62.69±3.32	76.30±1.63	54.39±3.98	50.61±4.35	58.88±4.03	51.13±4.24
PKMLRSSC [36]	63.84±2.98	78.49±1.58	56.04±2.70	50.37±3.37	63.26±2.18	53.04±2.82
DiMSC [27]	63.56±2.03	77.14±1.71	54.20±1.11	47.35±1.69	60.70±1.54	50.04±1.17

LMSC [28]	68.57±1.28	80.87±1.25	58.42±1.91	53.71±0.33	64.20±1.93	54.18±2.12
CoMSC [47]	67.00±1.26	77.08±1.21	54.14±0.35	54.69±0.56	53.61±0.48	50.54±0.89
COMVSC [48]	71.47±0.89	83.80±1.12	65.79±1.56	60.27±2.51	72.41±2.87	63.39±2.45
Ours	**71.74±2.86**	**86.10±1.45**	**67.06±3.43**	**61.31±4.17**	**74.15±3.84**	**64.76±3.66**

Table 6: Results (mean ± standard deviation) on the Reuter dataset (best results in bold)

Method	ACC	NMI	F-score	Recall	Precision	ARI
GMVNMF [41]	16.80±0.00	4.50±0.00	28.30±0.00	28.00±0.00	16.50±0.00	11.84±0.00
ECRMSC [10]	30.33±0.20	11.46±1.23	24.11±0.56	35.30±1.25	18.30±1.26	12.99±0.30
CMLRSSC [36]	50.10±3.13	33.42±2.27	39.23±1.87	45.08±4.49	34.90±1.35	25.36±1.85
PKMLRSSC [36]	54.79±3.38	36.84±1.46	41.21±1.46	45.44±2.11	37.86±1.46	28.26±2.24
DiMSC [27]	53.34±0.67	35.98±0.13	40.35±0.48	**48.20±0.15**	34.35±0.48	25.98±0.53
LMSC [28]	48.33±0.33	30.79±0.24	37.68±0.19	43.05±0.81	33.19±0.15	23.30±0.25
CoMSC [47]	54.50±2.13	37.35±2.41	41.63±1.56	48.12±1.36	36.25±0.56	28.74±1.12
COMVSC [48]	47.00±1.23	31.57±1.14	37.71±0.85	47.52±0.56	31.26±0.45	22.19±0.52
Ours	**60.74±2.98**	**41.15±2.36**	**45.74±1.92**	46.23±1.66	**45.29±2.29**	**34.86±2.42**

Table 7: Results (mean ± standard deviation) on the FERET dataset (best results in bold)

Method	ACC	NMI	F-score	Recall	Precision	ARI
GMVNMF [41]	7.00±0.20	49.00±0.42	4.92±0.41	5.75±0.33	4.60±0.10	4.11±0.35
ECRMSC [10]	24.64±0.45	61.67±1.20	4.54±1.23	5.82±0.85	4.20±0.25	2.80±0.86
CMLRSSC [36]	23.21±0.65	65.88±0.24	4.95±0.32	5.38±0.35	4.58±0.30	4.51±0.32
PKMLRSSC [36]	29.06±0.83	68.69±0.38	9.61±0.63	10.72±0.66	8.71±0.61	9.18±0.63
DiMSC [27]	29.82±0.50	68.41±0.60	9.32±1.57	10.44±1.78	8.51±1.40	8.95±1.58
LMSC [28]	27.93±0.72	68.01±0.94	8.28±0.31	8.87±0.63	7.48±0.39	7.89±0.41
CoMSC [47]	30.00±1.56	68.66±1.47	9.94±1.69	10.67±1.21	9.12±0.96	9.53±1.58
COMVSC [48]	30.07±0.17	68.57±0.29	10.51±0.56	12.28±0.85	8.70±0.34	10.05±0.98
Ours	**30.69±0.76**	**69.34±0.33**	**10.60±0.56**	**12.48±0.80**	**9.22±0.48**	**10.16±0.56**

In order to better understand the relevant classification results of the visual dataset, we display the boxplot of some datasets (ORL_mtv and COIL20) in Fig. 5. As shown in Fig. 5, the final classification results of our proposed method are relatively stable. Besides, the confusion matrix of the datasets ORL_mtv and COIL20 is shown in Fig. 6. It can be seen that the data on the main diagonal represents the number of correct clustering for each category. Where, the darker the color on the main diagonal is, the more samples that are clustered correctly by our proposed algorithm. By observing the distribution on the main diagonal, we can clearly see that the classification results are almost consistent with the classification results in Table 2 and Table 4.

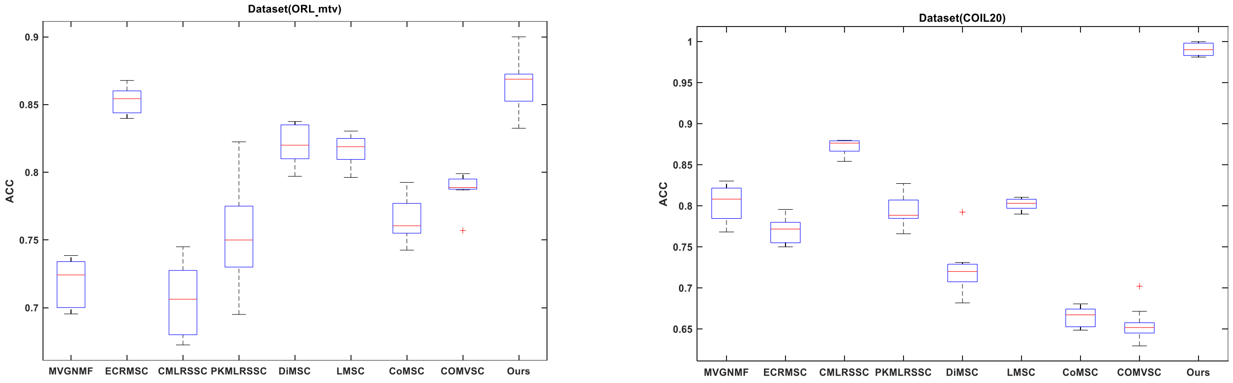

Figure 5: The boxplot of ORL_mtv and COIL20.

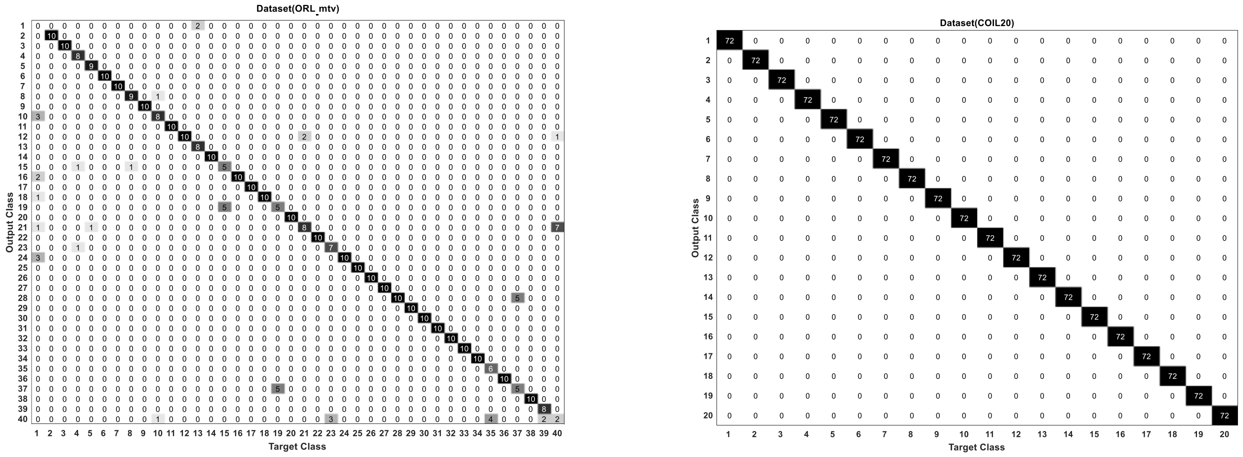

Figure 6: Confusion matrix of the datasets (ORL_mtv and COIL20), where the numbers in the black box on the main diagonal represent the number of samples that are correctly classified by our algorithm.

4.6 Similarity Matrix Selection

There are many ways to define the similarity matrix through the *k*-nearest neighbor graph. In this experiment, three different definitions are used, which include the binary weighting, heat-kernel weighting and dot-product weighting. The binary is generally the measure of similarity between samples. As shown in Eq. (3) and Eq. (4), they are the measures of similarity by heat kernel weighting and dot-product weighting between samples [39], respectively. After conducting a large number of experiments, it is found that the best results can be obtained by using the dot-product of real datasets. Various numbers of *k*-nearest neighbors for different datasets need to be specified for optimal clustering.

According to the results shown in Fig. 7, we select k=12 in our experiments. If someone wants to have the best performance of clustering, k can be selected at the range from 3 to 12.

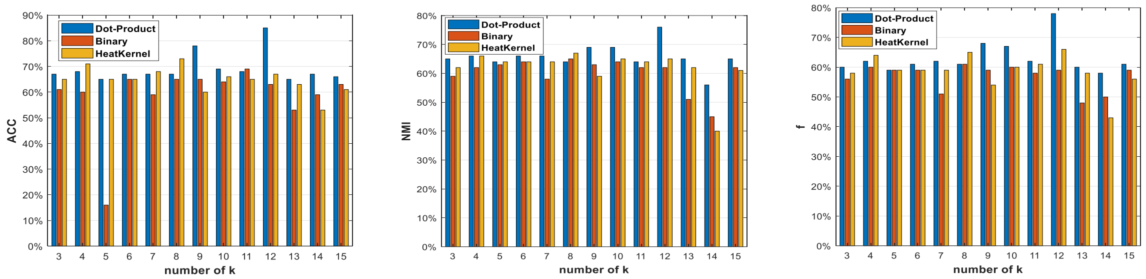

Figure 7: The performance of the proposed RLMDOM in the 3-Sources datasets (the same effect can be witnessed in other datasets). The parameter k with different weighting schemes (dot-product, binary and heat kernel). ACC, NMI and f were only used for comparison purposes and generally, k was chosen as 12.

4.7 Convergence and Algorithm Complexity Analysis

The way of minimizing the objective function is an iterative process. Herein, the most important prerequisite is to consider the convergence of the proposed algorithm. Actually, our method is found to have a faster convergence speed than others, as shown in Fig. 8. As seen, for most of the datasets, optimal results are achieved after 4 iterations in our proposed approach**.**

The complexity analysis of the proposed method is detailed in four main sub-problems, $V^{(v)}$, $U^{(v)}$, $E_x^{(v)}$ and $E_u^{(v)}$ as follows. For the sub-problem $V^{(v)}$, as Eq. (16a) is a standard Sylvester equation, according to Refs. [27][42], the algorithm complexity of update $V^{(v)}$ is $O(n^3)$, n is the number of samples. For the sub-problem $U^{(v)}$, as seen from Eqs. (19a), (19b) and (19c), this equals to an optimal sub-problem for the direct derivation of Quadratic polynomial. Therefore, its algorithmic complexity is mainly owning to the computation of the inverse of the matrix, which is $O(n^3)$. For updating $E_x^{(v)}$ and $E_u^{(v)}$, which is the L_{21} norm and nuclear norm optimization problem, thus the complexity will be $O(n^3)$. Through the stepwise analysis of the above four sub-problems, the complexity of the proposed method can be derived as $O(tn_vn^3)$, where t is the total number of iterations, and n_v is the number of views in each dataset. For comparison, the computational complexity terms of CMLRSSC, KMLRSSC, DiMSC and LMSC are found to be $O(tn_vn^3)$ [36], $O(tn_vn^3)$ [36], $O(tn_vn^3)$ [27] and $O(t(d^3 + n^3))$ [28], respectively, where d is the total number of dimensions of multi-view features. As can be seen, in theory the proposed method actually has a very comparable computational complexity as other peers, though it has produced significantly improved classification accuracy as validated in six datasets.

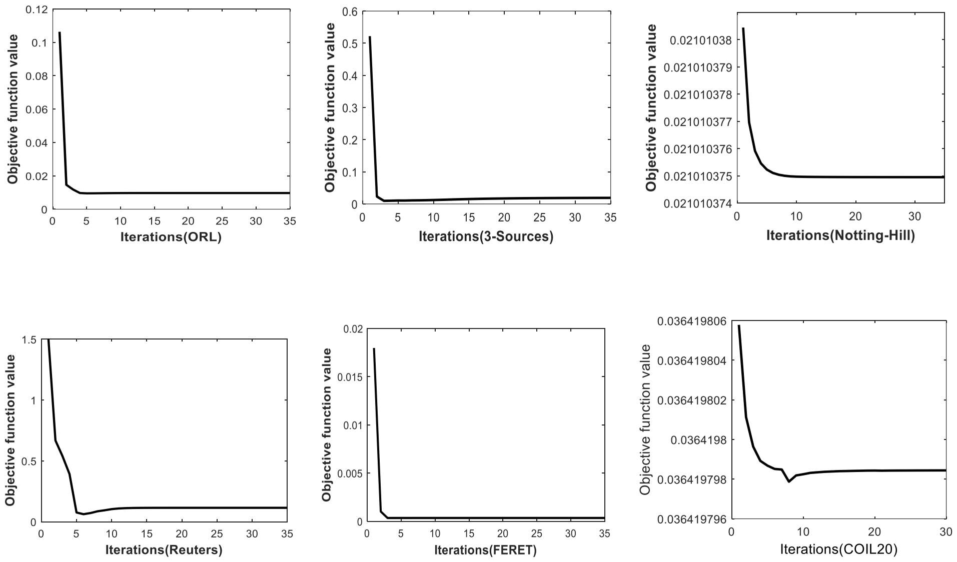

Figure 8: Convergence plots for ORL_mtv, 3-Sources, Notting-Hill, Reuters, COIL20 and FERET datasets are shown separately under different number of iterations.

Table 8: Comparison of execution times of related MSC algorithms (in seconds) (best results in bold and second best results underlined)

	ORL_mtv	COIL20	3-sources	Notting-Hill	Reuters	FERET
GMVNMF [41]	75.58	848.52	9.61	2976.11	9.22	80.88
ECRMSC [10]	117.13	384.48	30.98	423.76	31.41	812.67
CMLRSSC [36]	**2.12**	**13.16**	0.60	182.72	5.14	164.39
PKMLRSSC [36]	2.45	36.72	**0.33**	27.84	2.91	65.96
DiMSC [27]	25.16	469.54	4.19	161.38	16.93	260.95
LMSC [28]	33.90	442.67	7.84	466.72	85.81	829.00
CoMSC [47]	11.40	37.65	1.91	**23.21**	**2.63**	**21.19**
COMVSC [48]	22.79	146.60	11.20	100.92	43.03	236.03
Ours	12.51	38.37	2.94	26.77	6.01	21.34

As for the algorithm running time analysis, we show the running times of all the compared methods on the six real-word datasets in Table 8. Although the running time of our proposed method is not the best compared with other latest MSC algorithms, it is still comparable. For example, the running time of the proposed algorithm is quite good on Notting-Hill, Reuters and FERET datasets. In particular, the runtimes on FERET is only 0.15 seconds more than the best one. The running time of PKMLRSSC is particularly influenced by the sample dimension of the input data. It can be seen that 3-sources and ORL_mtv datasets have fewer sample dimensions, so PKMLRSSC consumes less time. Notting-Hill and FERET datasets have relatively higher sample dimensions, so PKMLRSSC consumes much more time. However, the time consumption of the proposed algorithm on all datasets is very stable, which is unaffected much by dataset size, sample feature dimensions, number of views, etc. However, the running time of other algorithms on all datasets fluctuates greatly. The proposed approach is the most efficient one among the nine approaches as shown in Table 8, due mainly to its fast convergence.

5 Conclusions and Future Work

In this paper, an efficient and effective algorithm, RLMDOM, is proposed for MVC, which features sparse low-rank subspace multi-view spectral clustering based on NMF and AWML. Benefitted from DiMSC and LMSC, our approach fully takes into account robust low-rank decomposition, hidden noise information, trade-off between the importance of the views and the diversity of the information fusion between different views for data fusion. In addition, we tested on six public datasets, including five multi-view datasets and one single-view datasets. In comparison with eight state-of-the-art algorithms, our approach has produced significantly improved results, and it further benefits from a fixed value for all datasets for the trade-off parameters of the entire objective function, i.e. λ_1= 0.001, λ_v= 0.001 and γ =1 . Besides, the fast convergence speed improves the working efficiency. Higher dimensional images are the trend of future development, and tensor is the most powerful theoretical support to solve higher dimensional images. Since our method can only deal with two-dimensional matrices, some useful information will be lost in sparse representation, such as view structure, interrelation between samples and intra-class discrimination. The research of high dimensional tensor optimization theory is our focus in the future.

Acknowledgements

This work is supported in part by the Science and Technology Project of Guangdong Province (no. 2019A050513011), Guangzhou Science and Technology Plan Project (no. 202002030386), National Nature Science Foundation of China (no. U1701266), Guangdong Graduate Education Innovation Project (no. 2020XSLT16), and Guangdong Provincial Key Laboratory of Intellectual Property and Big Data under Grant (no. 2018B030322016).

References

[1] E. Elhamifar and R. Vidal, “Sparse subspace clustering,” Proc. of the IEEE Conf. Computer Vision and Pattern Recognition (CVPR), pp. 2790-2797, 2009.

[2] Y.-M. Kim, M.-R. Amini, C. Goutte, P. Gallinari, “Multi-view clustering of multilingual documents,” Proc. of the Int. Conf. on Research and Development in Information Retrieval (SIGIR), pp. 821-822, 2010.

[3] Y. Shi, G. Li, Q. Cao, et al. “Face hallucination by attentive sequence optimization with reinforcement learning,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 42, no. 11, pp. 2809-2824, 2019.

[4] X. Cai, F. Nie, and H. Huang, “Multi-view k-means clustering on big data,” In Proc. 23rd Int. Joint Conf. on Artificial Intelligence, pp. 2598-2604, 2014.

[5] M. Zhang, Y. Yang, F. Shen, H. Zhang, Y. Wang, “Multi-view feature selection and classification for Alzheimer’s disease diagnosis,” Multimedia Tools Appl., vol. 76, pp. 10761-10775, 2017.

[6] V. Ha, J. Ren, X. Xu, S. Zhao, G. Xie, V. Masero, “Deep Learning Based Single Image Super-resolution: A Survey,,” International Journal of Automation and Computing, vol. 16, pp. 413-426, 2019.

[7] K. Kanjani, “Parallel non negative matrix factorization for document clustering,” Dept. Elect. Comput. Eng., Texas A&M Univ., College Station, TX, USA, Tech. Rep. CPSC-659, 2007.

[8] X. Wang, L. Zhen, X. Gao, C. Zhang, et al. “Multi-view subspace clustering with intactness-aware similarity,” Pattern Recognition, vol. 88, pp. 50-63, 2019.

[9] W. Zhu, J. Lu, and J. Zhou. “Structured general and specific multi-view subspace clustering,” Pattern Recognition, vol. 93, pp. 392-403, 2019.

[10] X. Wang, X. Guo, Z. Lei, C. Zhang, and S. Z. Li, “Exclusivity-consistency regularized multi-view subspace clustering,” Proc. of the IEEE Conf. on CVPR, pp. 923-931, 2017.

[11] N. Chen, J. Zhu, F. Sun, and E. P. Xing, “Large-margin predictive latent subspace learning for multiview data analysis,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 34, no. 12, pp. 2365-2378, 2012.

[12] G. Tzortzis and A. Likas, “Kernel-based weighted multi-view clustering,” in Proc. IEEE 12th Int. Conf. Data Mining, Brussels, Belgium, pp. 675-684, Dec. 2012.

[13] Y. Jiang, F. Chung, S. Wang, and Z. Deng, “Collaborative fuzzy clustering from multiple weighted views,” IEEE Trans. Cybern., vol. 45, no. 4, pp. 688-701, 2015.

[14] G. Zhang, C. Wang, D. Huang, and W. Zheng, “Multi-view collaborative locally adaptive clustering with Minkowski,” Expert Systems with Applications, vol. 86, pp. 307-320, 2017.

[15] G. Zhang, C. Wang, D. Huang, and W. Zheng, “TW-Co-k-means: Two-level weighted collaborative k-means for multi-view clustering,” Knowledge-Based Systems, vol. 150, pp. 127-138, 2018.

[16] C. Wang, J. Lai, and P. S. Yu, “Multi-view clustering based on belief propagation,” IEEE Trans. Knowl. Data Eng., vol. 28, no. 24, pp. 1007-1021, 2016.

[17] D. Cai, X. He, X. Wu, and J. Han, “Non-negative matrix factorization on manifold,” in Proc. Int Conf. on Data Mining, 2008.

[18] K. Zhan, F. Nie, J. Wang, and Y. Yang, “Multiview consensus graph clustering,” IEEE Trans. on Image Processing, vol. 28, no. 3, pp. 1261-1270, 2019.

[19] X. Wang, T. Zhang, and X. Gao, “Multiview clustering based on non-negative matrix factorization and pairwise measurements,” IEEE Trans. Cybern., vol. 49, no. 9, pp. 3333-3346, 2018.

[20] K. Zhan, J. Shi, J. Wang, and F. Tian, “Graph-regularized concept factorization for multi-view document clustering,” J. Vis. Commun. Image R., vol. 48, pp. 411-418, 2017.

[21] G. Liu, Z. Lin, and S. Yan, “Robust recovery of subspace structures by low-rank representation,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 35, no. 1, pp. 171-184, 2013.

[22] Y. Wang, L. Wu, and X. Lin, “Multiview spectral clustering via structured low-rank matrix factorization,” IEEE Trans. on Neural Networks and Learning Systems, vol. 29, no. 10, pp. 4833-4843, 2018.

[23] P. O. Hoyer, “Non-negative matrix factorization with sparseness constraints,” Journal of Machine Learning Research, vol. 5, pp. 1457-1469, 2004.

[24] Y. Wang, X. Lin, and L. Wu, “Robust subspace clustering for multi-view data by exploiting correlation consensus,” IEEE Trans. on Image Processing, vol. 24, no. 11, pp. 3939-3949, 2015.

[25] L. Zong, X. Zhang, L. Zhao, and H. Yu, “Multi-view clustering via multi-manifold regularized non-negative matrix factorization,” Neural Networks, vol. 88, pp. 74-89, 2017.

[26] A. Gretton, O. Bousquet, A. Smola, and B. Scholkopf. “Measuring statistical dependence with hilbert-schmidt norms,” Algorithmic learning theory, vol. 3734, pp. 63-77, 2005.

[27] X. Cao, C. Zhang, and H. Fu, “Diversity-induced multi-view subspace clustering,” Proc. of the IEEE Conf. CVPR, pp. 586-594, 2015.

[28] C. Zhang, H. Fu, and Q. Hu, “Generalized latent multi-view subspace clustering,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 42, no. 1, pp. 86-99, 2018.

[29] Z. Lin, M. Chen, and Y. Ma, “The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices,” UIUC Tech. Rep. UILU-ENG-09-2215, 2009.

[30] R. H. Bartels and G. Stewart. Solution of the matrix equation AX+ XB= C. Communications of the ACM, pp. 820-826, 1972.

[31] S. Boyd and L. Vandenberghe, “Convex optimization,” Cambridge, U.K.: Cambridge Univ. Press, 2004.

[32] F. Nie, H. Huang, and X. Cai, “Efficient and robust feature selection via joint L_21-norms minimization,” In Proc. 24th Annual Conf. on Neural Information Processing Systems, 2010.

[33] J. Cai, E. J Candes, and Z. Shen, “A singular value thresholding algorithm for matrix completion,” SIAM Journal on Optimization, vol. 20, no. 4, pp. 1956-1982, 2010.

[34] X. Cao, C. Zhang, and C. Zhou, “Constrained multi-view video face clustering,” IEEE Trans. on Image Proc., vol. 24, no. 11, pp. 4381-4393, 2015.

[35] P. J. Phillips, A. Martin, C. L. Wilson, et al. “An introduction evaluating biometric systems,” Computer, vol. 33, no. 2, pp. 56-63, 2000.

[36] M. Brbic, and I. Kopriva, “Multi-view low-rank sparse subspace clustering,” Pattern Recognition, vol. 73, no. 1, pp. 247-258, 2018,

[37] A. Strehl, and J. Ghosh, “Cluster ensembles-a knowledge reuse framework for combining multiple partitions,” the J. of Machine Learning Research (JMLR), vol. 3, no. 3, pp. 583-617, 2003.

[38] M. L. Zhang, Z. H. Zhou. “ML-KNN: A lazy learning approach to multi-label learning,” Pattern Recognition, vol. 40, no. 7, pp. 2038-2048, 2007.

[39] D. Cai, X. He, J. Han, et al. “Graph regularized nonnegative matrix factorization for data representation,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 33, no. 8, pp. 1548-1560, 2011.

[40] P. J. Phillips, A. Martin, C. L. Wilson, et al. "An introduction evaluating biometric systems," Computer, vol. 33, no. 2, pp. 56-63, 2000.

[41] D. Hidru, A. Goldenberg, "EquiNMF: Graph regularized multiview nonnegative matrix factorization," arXiv preprint arXiv: 1409.4018, 2014.

[42] R. H. Bartels and G. Stewart, "Solution of the matrix equation AX+ XB= C," Communications of the ACM, vol. 15, no. 9, pp.820-826, 1972.

[43] J. Gou, Y. Xue, H. Ma, et al. "Double graphs-based discriminant projections for dimensionality reduction," Neural Computing and Applications, vol. 32, no. 7, pp. 17533-17550, 2020.

[44] J. Gou, Y. Yang, Z. Yi, et al. "Discriminative globality and locality preserving graph embedding for dimensionality reduction," Expert Systems with Applications, vol. 144, 113079, 2020.

[45] G. Y. Zhang, X. W. Chen, Y. R. Zhou, et al. "Kernelized multi-view subspace clustering via auto-weighted graph learning," Applied Intelligence, pp. 1-16, 2021.

[46] H. Yu, X. Wang, G. Wang, et al. "An active three-way clustering method via low-rank matrices for multi-view data, Information Sciences," vol. 507, pp. 823-839, 2020.

[47] J. Liu, X. Liu, Y. Yang, et al. "Multiview subspace clustering via co-training robust data representation," IEEE Transactions on Neural Networks and Learning Systems, pp. 1-13, 2021.

[48] P. Zhang, X. Liu, J. Xiong, et al. "Consensus one-step multi-view subspace clustering," IEEE Transactions on Knowledge and Data Engineering, pp. 1-14, 2020.