SRAGL-AWCL: A Two-step Multi-view Clustering via Sparse Representation and Adaptive Weighted Cooperative Learning

Junpeng Tan [1], Zhijing Yang [1], Yongqiang Cheng [2], Jielin Ye [1], Bing Wang [2], Qingyun Dai [3]

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

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

[3] School of Electronic and Information, Guangdong Polytechnic Normal University, Guangzhou, China

ABSTRACT

Sparse representation and cooperative learning are two representative technologies in the field of multi-view spectral clustering. The former can effectively extract features of multiple views via removing redundant information contained in each view, whilst the latter can incorporate the diversity of each view. However, neither of them is adequate in preserving the internal geometric features of data. General approaches rarely consider the correlation between the similarities of the internal graph structures of individual views. To achieve the optimal global feature learning, a novel two-step multi-view spectral clustering strategy is proposed in this paper, where the sparse representation by adaptive graph learning is combined with the adaptive weighted cooperative learning. In the first step, matrix factorization by manifold regularization is proposed, which can strengthen the sparse features clustering discrimination of samples of each view. Specifically, the synchronization optimization method by introducing adaptive graph learning can better retain its complete internal structure of each view. This ensures the structural correlation between views through the usage of the sparse matrix and the optimal graph similarity matrix. In the second step, the adaptive weighted cooperative learning is performed on each view to get a global optimized matrix. In the meantime, the graph learning is also performed on the global matrix which ensures that the global matrix can be associated with various view features. Both of our experiment results shown on several multi-view datasets and single-view datasets indicate that our proposed approach has achieved significantly better performance than the current algorithms.

Keywords: Multi-view clustering, Sparse Representation (SR), Adaptive Graph Learning (AGL), Adaptive Weighted Cooperative Learning (AWCL), Global Optimized Matrix.

1 INTRODUCTION

Multi-view clustering (MVC) has become a hot topic due to the needs of analyzing the ever-rising ubiquitous multi-view data. MVC expresses an intuitive meaning for clustering from multiple angles and it fuses these angles to obtain a model that is more effective than single-view clustering. With its development, multi-view clustering can be regarded as using multiple features of samples extracted from different ways, such as local binary patterns (LBP) [1], 2D Gabor Wavelets (GABOR) [2] and histograms of oriented gradients (HOG) [3] etc. Features from different views take into account the characteristics of different angles of the same sample, which reflects its diversity. The diversity can then be used among these features to complement the information in order to better characterize the sample, hence MVC is currently one of the most used methods for tackling this problem. MVC can be divided into supervised MVC [4], semi-supervised MVC [5] and unsupervised MVC [6] according to the need of the associated class labels. Since unsupervised MVC does not need to annotate sample data, it is more convenient than other two types of clustering methods, therefore, it saves time and effort. The most representative unsupervised MVC methods are sparse representation (SR) [7], spectral clustering (SC) [8], graph learning (GL) [9] and consensus graph clustering [10]. Hybrid methods have also been used recently by combining multiple multi-view learning methods to obtain better results. Zhang et al [11] introduced an unsupervised clustering method which simultaneously learned fuzzy *k*-means and non-negative SC with side information. This method considers the internal structural information of each view and the diversity information between views. Preserving the structural information of each view and discriminating the diversity between them is another important research topic in MVC task. In the early research of MVC algorithm, researchers paid much attention to this. Cai et al [12] proposed a SR method based on non-negative matrix factorization (NMF). By adding the manifold regularization (MR) to preserve the local structural features of the data, it can improve the properties of the clustering within the class and keep its structural integrity within the class. Unfortunately, due to the way NMF handles the original data, some useful negative information is discarded to some extent. Hence, NMF has some limitations in data processing, which is also one of the challenges our paper aims to tackle with.

Further studies show that views are not independent and often inseparable in the process of MVC. Therefore, exploring the information connection between views has become another research hotspot. Feng et al [13] proposed an unsupervised multi-view link learning method, called adaptive unsupervised multi-view feature selection (AUMFS). This approach only considers the similarity learning among views on a cluster label matrix. Zhan et al [14] introduced a new method to leverage the problem of multi-view clusters by a joint approach which uses both adaptive structure concept factorization and optimization of the similarity matrix to deal with the data and the relationship information among views. Zhao et al [15] and Wang et al [16] suggested two models of adaptive similarity structure by using adaptive weighted decomposition of each view. In general, these methods simply weight each view to merge the various views without considering their complementarity and diversity comprehensively. Consequently, these approaches lead to uncertainty in term of the useful information inevitably. As a matter of fact, each view can be trained collaboratively to get a global matrix in which the diversity information of each view is contained. Hence, it ensures the effectiveness of clustering and makes full use of the diversity information between views. This is our motivation employing cooperative learning (CL) to solve the multi-view fusion problem.

The MVC based on CL approach has become an important direction of research in this field recently. You et al [39] introduced a global discriminant analysis method to handle the differences between views. Kumar et al [17] proposed a method that utilizes manifold learning to preserve the local structural features of each view, and transformed each view into a global matrix through model learning. Although it used the idea of cooperative representation, they did not consider the redundancy of data, nor the diversity of similarity matrix between views during process. Consequently, the existence of noise in each view can easily cause misjudgment of information in the fusion process. To solve the problem of data redundancy and cooperative representation, Wen et al [40] used sparse matrix as the learning graph for adaptive cooperative graph learning (ACGL). The method has some robustness to the noise in each view, and in particular it is helpful for discovering the internal structure of the noisy data. Brbic et al [18] proposed a method through learning a joint subspace representation by constructing an affinity matrix shared between views and using the importance of low-rank and sparse constraints in affinity matrix construction. Even though this method embodied its robustness at a certain level, it failed to fully preserve the internal structural integrity of each view.

To conclude, the MVC algorithms based on SR and CL have solved some problems that exist in features fusion. However, further study is still required into the usage of optimal extraction and fusion of specific information in the entire view. The issues associated with the existing methods detailed above can be summarized as follows:

1. In the SR of the original data, NMF requires non-negative input data and non-negative constraints on the basis matrix, which causes some useful information contained in negative input data being filtered out during decomposition.

2. It lacks effective methods to extract view-specific information. Existing methods have difficulties to obtain the optimal internal structural features, as the data often contains some redundant noise.
3. The diversity of views and the globality among view cannot be fully considered by independently using SR, AGL or CL.

In order to address these three issues, we propose an adaptive multi-view spectral clustering (MVSC) method. As shown in Fig. 1, the proposed method can be divided into two steps: 1) sparse representation with adaptive graph learning (SRAGL); 2) adaptive weighted cooperative learning (AWCL). Fig. 1 also illustrates the specific operation process of each step. The purpose of the first step (SRAGL) is to extract the optimal sparse matrix by SR and AGL. Redundant information from the original matrix is removed and the geometry of each view is preserved to the maximum possible extent. Additionally, the first step solves issues (1) and (2). The objective of the second step (AWCL) is to merge the diversity of views with the globality between views. In this step, it is crucial that the view with specific structure is effectively fused to make it more discriminative between different clusters. As shown in the right-hand side of Fig.1, step 2 will provide more discriminative representation. This is because AWCL fully considers the different advantages of each view, and then obtains an optimal global matrix. The above two steps closely link with SR, AGL and CL to form a final solution to tackle issue (3).

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

(1) A unique sparse matrix decomposition method is proposed to relax the non-negative constraint of the NMF basis matrix, so that the sparse matrix obtained by matrix decomposition contains more useful information.

(2) A direct derivative method is introduced to efficiently optimize the basis matrix and the sparse matrix. Consequently, it increases the similarity discrimination between samples, in the meantime, a quadratic processing is proposed for the coefficient matrix of the SR.

(3) A new two-steps algorithm is proposed to solve view specific information learning and fusion simultaneously. Firstly, we applied AGL to optimize the similarity matrix and preserve the internal structural features of each view by combining manifold learning. The integration of AGL and SR can extract the specific information of each view. After that, an AWCL fusion method is proposed to effectively fuse the view's diversity information. The specific information for different views is obtained by using adaptive weighted method to learn a global matrix for fusion. In order to make a complete global matrix data structure, a manifold learning is further applied to the global matrix.

The remaining sections of the paper are organized as follows: related works will be introduced in Section 2; the proposed method based on the adaptive SR and AWCL method is introduced in Section 3; the proposed optimization process is outlined in Section 4; Section 5 details the experimental results and analysis; and finally, Section 6 concludes the paper with prospective future work.

2 RELATED WORKS

2.1 Graph Learning

Many GL based methods are used to improve MVC algorithms [38] for MVC [20]. Among these methods, the similarity matrix to evaluate the similarity between samples can often be solved by applying simplest distance metric. However, it is sensitive to the initial graph input and therefore, the initialization process can critically impact the entire MVC performance. Assuming all the elements in the similarity matrix $S \in R^{n\times n}$ are non-negative, we can get the relevant property of the Laplacian matrix L as follows [21][22].

Theorem 1: The multiplicity c of 0 as an eigenvalue of the Laplacian matrix L is equal to the number of components in the similarity matrix $S \in R^{n\times n}$.

A very important constraint on the similarity measure matrix can be obtained from Theorem 1. If the constraint condition $\text{rank}(L) = n - c$ is satisfied, then the similarity matrix S is a key neighbor assignment and the data points have been split into c clusters [23]. According to the relevant theoretical proof, a conclusion can be drawn - when the sum of the first c top smallest eigenvalues of the Laplacian matrix L is equal to 0 and the constraint on $\text{rank}(L) = n - c$ is satisfied. Hence, $\sum_{i=1}^{c} \lambda_i = 0$, where the parameter λ_i is the i-th smallest eigenvalue of Laplacian matrix L. Therefore, according to Fan's theorem [24], we can get:

$$\sum_{i=1}^{c} \lambda_i = \min_{F} tr(F^T L F)$$

$$\text{s.t.} F \in R^{n\times c}, F^T F = I \tag{1}$$

where $(\cdot)^T$denotes the matrix transpose, I represents the identity matrix and $F^T = [f_1, f_2, \cdots, f_n]$ is the eigenvector matrix of the Laplacian matrix $L($ $L = D - [(S^T + S)/2]$, where $D = \sum_{i=1}^{n} (S^T + S)_{ii}/2)$. The theoretical proof of Eq. (1) can be derived from [25] and [26]. However, Zhan [20] found that it had a trivial solution of S, so Eq. (1) saw previous improvements. He added L_{21}-norm regularization to smooth the elements of the similarity matrix $S \in R^{n\times n}$ as well as the constraint that the sum of each column of the similarity matrix $S \in R^{n\times n}$ was one. This GL model can be expressed as follows:

$$\min_{S,F} \sum_{i,j=1}^{n} \left\|f_i - f_j\right\|_2^2 s_{ij} + \beta \|S\|_F^2$$

$$\text{s.t.} \forall \text{ j}, \mathbf{1}^T s_j = 1, s_j \geq 0 \tag{2}$$

where the parameter β is the regularization parameter of S, $\mathbf{1}^T$ is the transpose of unit row vector and s_j is the j-th column of S. f_i is the i-th eigenvector of the Laplacian matrix L. It can be seen that this model is superior in terms of retaining local features.

2.2 Graph Regularized Non-negative Matrix Factorization

Cai et al [12] proposed a method to combine NMF and MR. This method can decompose the data matrix into two non-negative basis matrix and a sparse matrix using NMF. In general, we use spectral non-negative sparse matrices for SC to achieve SR for the data matrix. By adding MR, this may preserve its internal structural features during the process of NMF. The NMF base on MR is the following objective function:

$$\min_{U,V} \|X - UV\|_F^2 + \lambda tr(V^T L V)$$

$$\text{s.t.} U \geq 0, V \geq 0 \tag{3}$$

where $X = [x_1, x_2, \cdots, x_n] \in R^{m \times n}$ is the data matrix, $U = u_{ij} \in R^{m \times n}$ is the basis matrix, $V \in R^{n \times n}$ is the coefficient matrix, and λ is the regularization parameter controlling the smoothness of the SR.

3 THE PROPOSED METHOD

In this section, the related symbols of this paper will firstly be briefed. For multi-view data, we can represent the input dataset as $X = [X^{(1)}, X^{(2)}, \cdots, X^{(v)}, \cdots, X^{(n_v)}]$, where n_v is the number of input data views, $X^{(v)}$ donates the v-th view data, and $X^{(v)} \in R^{m \times n}$, where m and n represent the number of features in each sample data and the number of samples, respectively. $X^{(v)} \in R^{m \times n}$ has the same data dimension for all n_v views. $W^{(v)} \in R^{m \times n}$ is used to represent the v-th view basis matrix of the SR, and $V^{(v)} \in R^{n \times n}$ is the v-th view coefficient matrix of the SR. We use $L^{(v)} \in R^{n \times n}$ to donate the v-th view Laplacian matrix for the coefficient matrix of SR, and $L^{(v)} = D^{(v)} - S^{(v)}, S^{(v)} = (S^{(v)} + (S^{(v)})^T)/2$, $D^{(v)} = \sum_i^n s_{ii}^{(v)}$, $S^{(v)}$ is the v-th view similarity graph matrix. L_* is the Laplacian matrix of the global matrix V^* in AWCL.

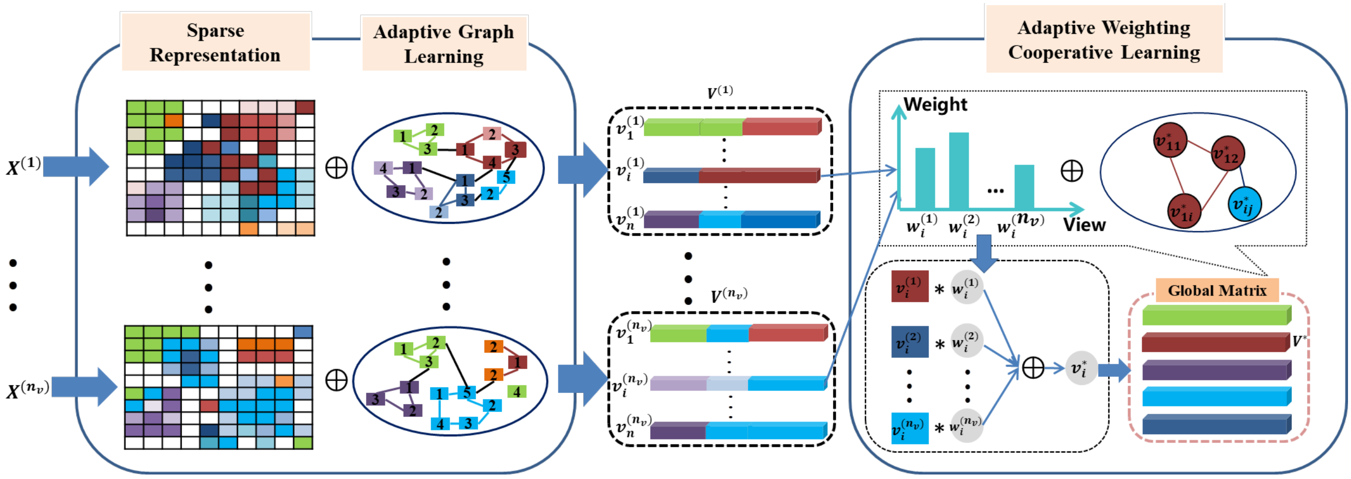


Figure 1: The flow chart of the proposed algorithm. The algorithm consists of a two-step strategy. The box on the left illustrates the first step (SRAGL), which can obtain the sparse features of each view. The box on the right illustrates the second step (AWCL), which results in a global matrix integrating all the view information.

3.1 Sparse Representation by Adaptive Graph Learning

The redundant data features existed in the original data can be reduced via SR by introducing MR to smooth the decomposed data. Cai et al [12] proposed a method to regulate the graph by using NMF for the data representation. This method not only reduces the dimensionality of the data but also preserves the local structural features of the sparse data. On the other hand, NMF is sensitive to the input dataset, which have to be non-negative. In order to solve this problem, we remove the non-negative constraint on the basis matrix. The improved model can be represented as follows:

$$\min_{V^{(v)}} \sum_{v=1}^{n_v} \left\| X^{(v)} - W^{(v)} V^{(v)} \right\|_F^2 + \lambda^{(v)} tr((V^{(v)})^T L^{(v)} V^{(v)})$$

$$\text{s.t. } V^{(v)} \geq 0 \tag{4}$$

where $\lambda^{(v)}$ is the v-th view regularization parameter. The greater $\lambda^{(v)}$ is, the more important the second part of this model. The first part of this model represents the SR. By using the idea of least squares, the original data is decomposed into two matrices, one of which is the basis matrix and the other is the sparse matrix. This model can remove the redundant information by SR. Notably, the MR is used in the second part of the objective function (4). $tr((V^{(v)})^T L^{(v)} V^{(v)}) \equiv \sum_{i,j=1}^{n} \left\| v_i^{(v)} - v_j^{(v)} \right\|_2^2 s_{ij}^{(v)}$ is the manifold learning of each view, which preserves the internal structural features. More details can be found in [12].

In most MVC algorithms, the similarity matrix often remains unchanged during the whole process. Thus, this will affect the performance of the cluster to some extent. Since the internal structural features between different views and the sparse matrix obtained by SR of the original data should be optimized each time, we obtain an optimal sparse matrix through a continuous optimization process. The redundant information of the original data matrix can be minimized significantly, and at the same time an optimal similarity matrix can be obtained to preserve the local geometric features of the original data. Consequently, we have to adaptively learn the optimal similarity matrix as the follows:

$$\begin{aligned} \min_{W^{(v)},V^{(v)}S^{(v)}} & \sum_{v}^{n_v} \left\| X^{(v)} - W^{(v)} V^{(v)} \right\|_F^2 + \lambda^{(v)} tr((V^{(v)})^T L^{(v)} V^{(v)}) \\ & + \sum_{v=1}^{n_v} \sum_{i,j=1}^{n} \left\| f_i^{(v)} - f_j^{(v)} \right\|_2^2 s_{ij}^{(v)} + \lambda^{(v)} \left\| S^{(v)} \right\|_F^2 \\ & \mathrm{s.t.}\ \forall \mathrm{j}\ \mathbf{1}^T s_j^{(v)} = 1, s_j^{(v)} \geq 0, V^{(v)} \geq 0 \end{aligned} \tag{5}$$

where $s_{ij}^{(v)}$ is the (i,j)-th element of $S^{(v)}$, $f_i^{(v)}$ is the i-th eigenvector of the Laplacian matrix $L^{(v)}$ for the *v*-th view. $\lambda^{(v)}$ denotes the regularization parameter of each similarity matrix and MR. The third and fourth parts of the objective function represent the adaptive learning of the similarity matrices. The structural information of the view is preserved by learning the internal structures accordingly.

3.2 Adaptive Weighted Cooperative Learning

For MVC to exhibit high quality performance, not only its structural characteristics of each view's data features but also its fusion characteristics need to be considered. Therefore, the full use of the diversity information between views needs to be implemented and effective integration of the complementary information of each view needs to be enabled to improve the clustering performance. Considering the independent nature between the views, the advantage of the diversity between various views can be integrated via the idea of CL. The GL is simultaneously used to smooth the global matrix and preserve the structural characteristics of the global matrix. According to Fan's theorem [24], we can use the following objective function to represent this model:

$$\min_{w_j^{(v)},v_j^*,Y} \sum_{j=1}^{n} \left\| v_j^* - \sum_{v=1}^{n_v} w_j^{(v)} v_j^{(v)} \right\|_F^2 + \eta tr(Y^T L_* Y)$$

$$\text{s.t.}\ \mathbf{1}^T v_j^* = 1, \sum_{v=1}^{n_v} w_j^{(v)} = 1, Y \in R^{n\times c}, Y^T Y = I \tag{6}$$

where $v_j^* = [v_{1j}^*, v_{2j}^*, \cdots v_{nj}^*]^T$, η is the trade-off parameter for MR, $Y^T = [y_1, y_2, \cdots, y_n]$ is the eigenvector matrix of the Laplacian matrix L_* for the global matrix. In addition, the matrix normalization constraints can be added to the fusion matrix to make the global matrix smoother. Hence, Eq. (6) can be rewritten as follows:

$$\min_{w_j^{(v)},v_j^*,Y} \sum_{j=1}^{n} \left\| v_j^* - \sum_{v=1}^{n_v} w_j^{(v)} v_j^{(v)} \right\|_F^2 + \eta \sum_{i,j=1}^{n} \left\| y_i - y_j \right\|_F^2 v_{ij}^*$$

$$\text{s.t.}\ \mathbf{1}^T v_j^* = 1, \sum_{v=1}^{n_v} w_j^{(v)} = 1, Y \in R^{n\times c}, Y^T Y = I \tag{7}$$

From Eq. (7), we can find out that the model displays the ability to learn the internal structures of the global matrix and partially contribute to the improvement of the clustering performance.

4 OPTIMIZATION

In this section effective iterative methods to solve Eqs. (5) and (7) are proposed, including specific details of the optimization of these two parts.

4.1 Graph Learning Optimization

In this optimization sub-module, four parameters in Eq. (5) need to be updated, i.e. $S^{(v)}, V^{(v)}, W^{(v)}, f_i^{(v)}$. The augmented Lagrange multiplier method is used to optimize this objective function. More details are provided as below.

1) The first part involves updating each single view similarity graph $S^{(v)}$ by fixing other three variables. Thus, the following objective function can be obtained:

$$\min_{s_{ij}^{(v)}} \sum_{v=1}^{n_v} \sum_{i,j=1}^{n} \left\| f_i^{(v)} - f_j^{(v)} \right\|_2^2 s_{ij}^{(v)} + \lambda^{(v)} \left\| S^{(v)} \right\|_F^2$$

$$\mathrm{s.t.} \forall \mathrm{j}\ \mathbf{1}^T s_j^{(v)} = 1, s_j^{(v)} \geq 0 \tag{8}$$

let $g_{ij}^{(v)} = \left\| f_i^{(v)} - f_j^{(v)} \right\|_2^2$, then Eq. (8) will become:

$$\min_{s_{ij}^{(v)}} \sum_{v=1}^{n_v} \sum_{i=1}^{n} g_{ij}^{(v)} * s_{ij}^{(v)} + \lambda^{(v)} \sum_{i=1}^{n} (s_{ij}^{(v)})^2$$

$$\mathrm{s.t.} \forall \mathrm{j}\ \mathbf{1}^T s_j^{(v)} = 1, s_j^{(v)} \geq 0 \tag{9}$$

Eq. (9) can be further simplified as:

$$\min_{s_j^{(v)}} \left\| s_j^{(v)} + \frac{1}{2\lambda^{(v)}} g_j^{(v)} \right\|_2^2$$

$$\mathrm{s.t.} \forall \mathrm{j}\ \mathbf{1}^T s_j^{(v)} = 1, s_j^{(v)} \geq 0 \tag{10}$$

where $\left[g_{1j}^{(v)}, g_{2j}^{(v)}, \cdots, g_{nj}^{(v)}\right]^T$ is denoted by $g_j^{(v)}$. This objective function's Lagrangian function is then presented as the follows:

$$\mathrm{L}\left(s_j^{(v)}, \alpha, \rho\right) = \left\| s_j^{(v)} + \frac{1}{2\lambda^{(v)}} g_j^{(v)} \right\|_2^2 - \alpha\left((s_j^{(v)})^T \mathbf{1} - 1\right) - \rho s_j^{(v)} \tag{11}$$

where α and ρ are the Lagrangian multipliers.

According to Karush-Kuhn-Tucker (KKT) condition [27], we can easily get the optimal solution of Eq. (11):

$$s_j^{(v)} = \left(-\frac{g_j^{(v)}}{2\lambda^{(v)}} + \alpha\right)_+ \tag{12}$$

where the symbol $+$ denotes greater than 0.

2) The second updating parameter $f_i^{(v)}$ is obtained by fixing the other three variables. We can optimize the variable $F^{(v)}$, and $(F^{(v)})^T = (f_i^{(v)}, f_2^{(v)}, \cdots, f_n^{(v)})$, and then Eq. (5) can be expressed as:

$$\min_{F^{(v)}} tr((F^{(v)})^T L^{(v)} F^{(v)})$$

$$\mathrm{s.t.} F^{(v)} \in R^{n \times c}, (F^{(v)})^T F^{(v)} = I. \tag{13}$$

This objective function can be solved by calculating the eigenvectors of $L^{(v)}$.

3) The third updating parameter is each single view $V^{(v)}$ by fixing other three variables. Therefore, we can get the following objective function:

$$\min_{V^{(v)}} \left\| X^{(v)} - W^{(v)} V^{(v)} \right\|_F^2 + \lambda^{(v)} tr((V^{(v)})^T L^{(v)} V^{(v)})$$

$$\text{s.t. } V^{(v)} \geq 0. \tag{14}$$

Eq (14) demonstrates that a simple quadratic function can be easily achieved. Hence, we can solve it by applying the derivation directly.

$$V^{(v)} = \frac{W^{(v)} X^{(v)} + \lambda^{(v)} S^{(v)} V^{(v)}}{(W^{(v)})^T W^{(v)} V^{(v)} + \lambda^{(v)} D^{(v)} V^{(v)}}. \tag{15}$$

There is a constraint in Eq. (5) that is $V^{(v)} \geq 0$. To ensure the effectiveness of this constraint, we add the square to variable $V^{(v)}$. Firstly, the sparse matrix $V^{(v)}$ can be interpreted as similarity measurement matrix between data samples. The absolute values of coefficient matrix elements represent the similarity between samples. Besides, the square has the effect of polarization, increasing the value of the number greater than 1 and decreasing the value of the number smaller than 1. This shows that two samples within the same category have larger weights and the sample weights between different classes will be smaller. To a certain extent, it can play a role in preserving the characteristics of the global structure. Therefore, the update of $V^{(v)}$ can be further expressed as follows:

$$V^{(v)} := (V^{(v)})^2 \tag{16}$$

4) By fixing the other three parameters, the fourth updating parameter $W^{(v)}$ is obtained in each single view. We can use the direct derivation method to optimize this variable. Thus, we get the following solution:

$$W^{(v)} = X^{(v)} * inv\left(V^{(v)}\right). \tag{17}$$

The optimization process of the SR and AGL model is illustrated in Algorithm 1.

Algorithm 1 Update parameters of SRAGL, $V^{(v)}, S^{(v)}, W^{(v)}, F^{(v)}$.

```
Input: Dataset $X = [X^{(1)}, X^{(2)}, \cdots, X^{(v)}, \cdots, X^{(n_v)}]$ and cluster number $c$.
Output: the SR of each view, $V^{(v)}, v \in [1, n_v]$.
Initialize: Each view graph $S^{(v)}$ is initialized with Eq. (8) by substituting $X^{(v)}$ into $f$, and using each initial graph $S^{(v)}$ to obtain the initial $F^{(v)}$ by Eq. (11). $V^{(v)} = I^{n \times n}, W^{(v)} = I^{m \times n}$.
for $v \in [1, n_v]$ do
    while $t$<= inter（inter=30 in our experiments）
        update $s_j^{(v)}$ by using Eq. (12).
        update $F^{(v)}$ by using Eq. (13).
        update $V^{(v)}$ by using Eqs. (15) and (16).
        update $W^{(v)}$ by using Eq. (17).
    end while
end for
```

4.2 Adaptive Weighted Cooperative Learning Optimization

By learning SRAGL, we firstly obtain sparse matrices for individual views. Next, we get a global matrix with complementary information between the views via fusing the sparse matrix of each view. Finally, the objective function of the optimization model in Eq. (7) will be expressed as follows:

$$\mathrm{L}\left(w_j^{(v)}, v_j^*, y_j\right) = arg\ min \sum_{j=1}^{n} \left\| v_j^* - \sum_{v=1}^{n_v} w_j^{(v)} v_j^{(v)} \right\|_F^2 + \eta \sum_{i,j=1}^{n} \left\| y_i - y_j \right\|_F^2 v_{ij}^*$$

$$\text{s.t. } \mathbf{1}^T v_j^* = 1, \sum_{v=1}^{n_v} w_j^{(v)} = 1, Y \in R^{n \times c}, Y^T Y = I. \quad (18)$$

There are three main variables in Eq. (18), namely, $w_j^{(v)}$, v_j^*, y_j, which need to be optimized separately.

1) Update v_j^*: when optimizing the variable v_j^*, we fix $w_j^{(v)}$, y_j, and Eq. (18) can be rewritten as follows:

$$\min_{v_j^*} \sum_{j=1}^{n} \left\| v_j^* - \sum_{v=1}^{n_v} w_j^{(v)} v_j^{(v)} \right\|_F^2 + \eta \sum_{i,j=1}^{n} \left\| y_i - y_j \right\|_F^2 v_{ij}^*$$

$$\text{s.t. } \mathbf{1}^{\boldsymbol{T}} v_j^* = 1. \quad (19)$$

According to [20], different columns of v^* are independent, therefore, each column can be solved separately. Hence, the objective function is represented by the following equation:

$$\min_{v_j^*} \sum_{j=1}^{n} \left\| v_j^* - \sum_{v=1}^{n_v} w_j^{(v)} v_j^{(v)} \right\|_F^2 + \eta \sum_{i,j=1}^{n} \left\| y_i - y_j \right\|_F^2 v_{ij}^* + \gamma (\mathbf{1}^T v_j^* - 1)^2 \quad (20)$$

where γ is the Lagrangian multiplier.

Since the variable y_i is fixed, denote $\left\|y_i - y_j\right\|_F^2$ as h_{ij}. Finally, Eq. (20) can be simplified as follows:

$$\min_{v_j^*} \left\|v_j^* - (\frac{\eta}{2}h_j - \sum_{v=1}^{n_v} w_j^{(v)} v_j^{(v)})\right\|_2^2 + \gamma(\mathbf{1}^T v_j^* - 1)^2 \tag{21}$$

where $h_j = [h_{1j}, h_{2j}, \cdots, h_{nj}]$ and the process of solving Eq. (19) is similar to that of Eq. (14). The variable v_j^* can be optimized as follows:

$$v_j^* = \sum_{i=1}^{n_v} w_j^{(v)} v_j^{(v)} - \frac{\eta}{2}h_j + \gamma \tag{22}$$

2) Update y_i: we must fix $w_j^{(v)}$ and $v_j^{(v)}$, the objective function becomes

$$\min_{Y} tr(Y^T L_* \mathrm{Y})$$

$$\mathrm{s.t.}\, Y \in R^{n\times c}, Y^T Y = I \tag{23}$$

the solution of Eq. (23) is the same as Eq. (11). It calculates the eigenvectors of L_*.

3) Update $w_j^{(v)}$: this sub-problem involves fixing v_j^* and Y. Consequently Eq. (18) becomes

$$\min_{w_j^{(v)}} \sum_{j=1}^{n} \left\|v_j^* - \sum_{v=1}^{n_v} w_j^{(v)}\, v_j^{(v)}\right\|_F^2$$

$$\mathrm{s.t.} \sum_{v=1}^{n_v} w_j^{(v)} = 1 \tag{24}$$

According to [20], Eq. (24) can be simplified as follows:

$$\min_{w_j^{(v)}} \sum_{j=1}^{n} \left\|v_j^* - \sum_{v=1}^{n_v} w_j^{(v)}\, v_j^{(v)}\right\|_F^2 = \min_{W_j} \sum_{j=1}^{n} W_j^T Z_j^T Z_j W_j, \tag{25}$$

where $Z_j = \left[z_j^{(1)}, z_j^{(2)}, \cdots, z_j^{(n_v)}\right], z_j^{(v)} = v_j^* - v_j^{(v)}, W_j = \left[w_j^{(1)}, w_j^{(2)}, \cdots, w_j^{(n_v)}\right]$. Then, the Lagrangian function of Eq. (24) is given by

$$L\left(W_j, \emptyset\right) = \sum_{j=1}^{n} W_j^T Z_j^T Z_j W_j - \emptyset(1 - w_j^T \mathbf{1}) \tag{26}$$

where $\emptyset$ is the Lagrangian multiplier.

The direct derivation is used to solve Eq. (26) by setting its derivative as 0. We can get the following equation:

$$\frac{\partial L(W_j,\emptyset)}{\partial W_j} = Z_j^T Z_j W_j - \emptyset\mathbf{1} = 0. \tag{27}$$

By combining constraint $W_j^T\mathbf{1} = 1$ with Eq. (27), we can get the final solution W_j as

$$W_j = \frac{(Z_j^T Z_j)^{-1}\mathbf{1}}{\mathbf{1}^T(Z_j^T Z_j)^{-1}\mathbf{1}}. \tag{28}$$

The above steps stop when the difference between the two iterations is smaller than a threshold, or the maximum number of iterations INTER is reached. The optimization process of the AWCL model is summarized as in Algorithm 2.

Algorithm 2：Update parameters iterative algorithm of AWCL, $v_j^*, w_j^{(v)}$.

```
Input: different view sparse matrix $V_j = \left[v_j^{(1)}, v_j^{(2)}, \cdots, v_j^{(n_v)}\right]$.
Output: a global matrix $V^*$.
Initialize: Each element of $W_j$, $\forall j$, is set to $1/n_v$, eps=0.001, and let $V^* = \sum_{v=1}^{n_v} V^{(v)}/n_v$, $V^* = V^* - \sum_{i=1}^{n} V_{ii}^*$.
while abs((obj_new - obj_old)/obj_old) > eps do
    for $i$<=INTER（INTER=30 in our experiments）
        update $v_j^*$ by solving Eq. (22).
        update $Y$ by solving Eq. (23).
        update $W_j$ by solving Eq. (28).
        obj_old = obj_new;
        obj_new = sum(sum($V^*$- sum($\sum_{v=1}^{n_v}\sum_j^n(w_j^{(v)} * v_j^{(v)})$))$^2$);
    end for
end while
```

In summary, we individually optimize SRAGL and AWCL modules. Through the optimization process, the raw data of each view is sparsely represented, which is then fed to the AWCL as input. A global matrix with complementary information is then obtained, which is used as the final input to the SC. The overall flow of the algorithm is given in Algorithm 3.

Algorithm 3: The overall flow of the algorithm SRAGL-AWCL.

```
Input: Data matrix $X = \left[X^{(1)}, X^{(2)}, \cdots, X^{(v)}, \cdots, X^{(n_v)}\right]$, $\lambda^{(v)}$.
Initialization: Each view graph $S^{(v)}$ is initialized with Eq. (2) by substituting $X^{(v)}$ into $f$, and
    using each initial graph $S^{(v)}$ to obtain initial $F^{(v)}$ by Eq. (11). Let $V^{(v)} = I^{n\times n}, W^{(v)} = I^{m\times n}$,
    $V^* = \sum_{v=1}^{n_v} V^{(v)}/n_v$, $V^* = V^* - \sum_{i=1}^{n} V_{ii}^*$, $\lambda^{(v)} = 0.01$.
Output: The features of data points to c clusters
```

```
for $i$<=INTER
    update $V^{(v)}$ by using Eqs. (15) and (16).
end for
while not converge do
    update $v_j^*$ by solving Eq. (22).
    update $W_j$ by solving Eq. (28).
    if $i$ > INTER
        break.
    end if
end while
$Z = |V^* + (V^*)^T|/2$.
Apply SC to the affinity matrix $Z$.
```

5 EXPERIMENTS RESULTS AND ANALYSIS

5.1 Experimental Setting

5.1.1 Multi-view Datasets

1) UCI_Digits:[1] It includes ten classes of handwritten numbers in this dataset, which consists of ten numbers from 0 to 9. In this experiment, five views are used for each point: the first view is the 216-D profile-correlation feature; the second is the 76-D Fourier-coefficient feature; the third is 64-D Karhunen-Loeve-coefficient feature; the fourth is 240-D intensity-averaged feature in windows; and the last view is 47-D morphological feature.

2) ORL_mtv dataset:[2] This is a generic dataset used by many multi-view algorithms. There are 10 different gray scale facial images of 40 distinct subjects, which are taken under different conditions, such as: different illuminations; different facial expressions; and facial details. In this experiment, three recognized views are used to evaluate the algorithm.

3) Notting-Hill:[3] This dataset is an image dataset extracted from a movie called 'Notting Hill', which contains 4660 images in 76 tracks [28][29]. Since the images in each track are very similar, we consider each track as a class. In this experiment, the first 20 tracks of the dataset are selected for a total of 1206

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

[2] http://www.cl.cam.ac.uk/research/dtg/attarchive/facedatabase.html.

[3] https://bitbucket.org/chengjuzhou/constrained-multi-view-video-face-clustering/src/master/CX_VFC/data/NH/.

samples. Here three views features are used to represent each facial image: the first view is the 6750-D Gabor feature; the second view is 3304-D LBP feature; and the third view is the 2000-D gray feature.

4) Caltech-101:[4] The image dataset includes 101 categories of images [30]. We choose the six general categories of objects and obtain 1439 images. These six objects include the following: faces, motorbikes, dollar bill, Garfield, stop sign, and Windsor chair. We also use three views features to represent each image: the first view is 160-D Gabor feature; the second view is 324-D hog feature; and the third view is 236-D LBP feature.

5.1.2 Single-view Datasets

1) COIL-20: This dataset is a globally used dataset by single view clustering algorithms. It has 32x32 gray scale images of 20 objects viewed from various angles. In total, there are 1440 object image samples in the dataset [31].

2) USPS: This is also a commonly used digital image dataset by single view clustering algorithms. This dataset contains a large number of samples, with 9298 digital image samples, and its size is 16x16, which contains numbers from 0 to 9. The information of all the datasets is summarized in Table 1.

Tabel 1: Statistic of the Datasets

Datasets	Instance	View	Class
UCI_Digits	2000	5	10
ORL_mtv	400	3	40
Notting-Hill	1206	3	20
Caltech-101	1439	3	6
COIL-20	1440	1	20
USPS	9298	1	10
ORL_32x32	400	1	40
Yale_32x32	165	1	15

5.2 Compared Methods

[4] http://www.vision.caltech.edu/Image_Datasets/Caltech101/.

5.2.1 Multi-view Compared Methods

RMKMC: This is a multi-view method of k-means clustering for big data [32]. The algorithm adds L_{21}-norm into sparse factorization to obtain robust result. According to the authors' suggestion, the parameter $log_{10}\gamma$ is adjusted in the range of [0.1, 2] with interval 0.2 in order to obtain the best result.

CRSC: This is a classical MVC method [17] of which its main contribution is to carry out the local structure reservation of each view and the implementation of the global view via co-regularizing the clustering hypotheses. For the key parameter λ as recommended in the paper, we set it as 0.01.

AMGL: The highlight of this approach is the use of automatic weighted multi-view GL [19]. It obtains a new representation of each data point by summing up different Laplacian matrices from different views.

MVGL: This is MVC with GL [20]. It involves automatic adjustment of the similarity matrix of each view to get a good result as well as cooperative learning to obtain a global similarity matrix.

MVCF: This is adaptive structure concept factorization for MVC [14]. The contribution of this algorithm is to use concept decomposition for the original data, which is different from NMF. The similarity matrix in this algorithm is updated according to the algorithm optimization. The default parameter values are used for this algorithm.

MVNMF: This is MVC involving the jointing of NMF [37]. The main idea is to sparse represent the row dataset by fusing each view sparse matrix into a global matrix by NMF and related constraints. According to the requirements of the paper, we set λ=0.01.

KPMLRSSC: Kernel pairwise multi-view low-rank sparse subspace clustering [18] uses linear decomposition to achieve sparse matrices. The addition of the nuclear norm and L_1-norm constraint to a sparse matrix makes the process more robust. The parameters are set to the default values according to the paper.

KCMLRSSC: Kernel centroid-based multi-view low-rank sparse subspace clustering [18] is an algorithm model that is the same as KPMLRSSC in terms of sparse decomposition and related constraints. However, this method also obtains a global sparse matrix via computing between views.

5.2.2 Single-view Compared Methods

GRNMF: This method is about data representation based on graph regularized NMF [12]. This paper is the first one to combine NMF with manifold learning to preserve the local structural features of the data.

DSRMR: This robust unsupervised feature selection method includes dual self-representation and MR [33]. It uses L_{21}-norm for robust outliers. Likewise for the similarity matrix, the dual update method is used to optimize the variable.

GSR_SFS: This is a method based on factor self-representation for sparse feature selection [34]. The difference between GSR_SFS and DSRMR is that a traditional fixed similarity matrix is used to solve the model.

GLOSS: This is a sparse subspace learning feature selection method based on local structural feature preservation [35]. It can simultaneously realize feature selection and subspace learning.

RSR: This is an unsupervised feature selection method performed by regularized self-representation [36], which uses the L_{21}-norm to characterize the representation coefficient matrix.

5.3 Evaluation Matrices and Complexity Analysis

5.3.1 Evaluation Matrices

In order to evaluate the experimental results, several globally adopted evaluation criteria are applied, including accuracy (ACC), normalized mutual information (NMI), F1-score, precision, and adjusted Rand index (ARI). F1-score and precision can be defined as:

$$\text{Precision} = \frac{TP}{TP+FP}, \tag{29}$$

$$\text{Recall} = \frac{TP}{TP+FN} \tag{30}$$

$$\text{F1–score} = 2\frac{Precision\cdot Recall}{Precision+Recall} \tag{31}$$

where true positive (TP) is the number of item pairs that are in the same cluster and belong to the same class. False positive (FP) is the number of item pairs that are in the same cluster but belong to different classes and false negative (FN) is the number of item pairs that are in different clusters but belong to the same class. Another complex index is ARI, which can be defined as follows:

$$\text{ARI} = \frac{\sum_{ij}\binom{n_{ij}}{2} - [\sum_i\binom{a_i}{2}\sum_j\binom{b_j}{2}]/\binom{n}{2}}{\frac{1}{2}\left[\sum_i\binom{a_i}{2}+\sum_j\binom{b_j}{2}\right] - [\sum_i\binom{a_i}{2}\sum_j\binom{b_j}{2}]/\binom{n}{2}} \tag{32}$$

where n_{ij} represents the number of classes belonging to both class i and cluster j, a_i and b_j are the numbers of class i and cluster j, respectively.

5.3.2 Complexity Analysis

In this section, we theoretically analyze the computational complexity of the proposed SRAGL-AWCL to evaluate the computational efficiency. Our approach can be divided into two steps, namely the SRAGL and AWCL. The first step is to calculate the complexity of the objective function (5). According to [20] and the derivative of a square function, we know the complexity of SRAGL is $O(t_1 n_v(3n^2 + cn^2))$, where $t_1, n_v,$ c and n are the number of iterations in SRAGL , number of views, number of clusters and number of samples, respectively. Since the highest power is 2 and has some constant terms in it, the complexity of SRAGL can be simplified to $O(t_1 n_v n^2)$. The second step is to calculate the complexity of AWCL. According to [20], we can get the complexity of AWCL as $O(t_2(n_v^2 n + n_v^3 + cn_v)n)$, where t_2 is the number of iterations of AWCL. According to the analysis of the complexity of SRAGL and AWCL, the complexity of SRAGL-AWCL can be expressed as $O(t_1 n_v n^2 + t_2(n_v^2 n + n_v^3 + cn_v)n)$. Since normally $n_v \ll n$ and $c \ll n$, the complexity of SRAGL-AWCL is $O(tn_v^2 n^2)$, where t is the number of the total iterations. We can also find out the algorithmic complexity of RMKMC [32], MVGL [20], MVCF [14], and MVNMF [37] as $O(tn_vn)$, $O(tn_vn^2)$, $O(tn_vn^2)$, and $O(tn_vmn)$, respectively, where m denotes the size of each sample. By comparing the complexity of these algorithms, the lowest algorithmic complexity is a linear polynomial, whilst the most algorithmic complexity is a quadratic polynomial of n. The computation times of each algorithm are listed in Tables 2-6 (the last column) for different datasets. It can be seen that the time consumption of the algorithms with relatively lower algorithmic complexity is greater in some cases, for example, the time consumption of RMKMC in multi-view dataset ORL-mtv and Notting-Hill. This is because the number of iterations, t, can determine the time consumption of the whole algorithm to a certain extent. Besides, the time consumption of the same algorithm may vary on different datasets, mainly due to the different number of iterations (such as MVGL in datasets COIL20 and UCI_Digits) needed in the process. On the other hand, the size of the sample data also determines the running time of the algorithms (such as MVNMF in datasets ORL_mtv and UCI_Digits). Overall, the time consumed by our proposed algorithm is relatively low compared to other algorithms, which shows the good efficiency of our algorithm.

5.4 Clustering Results

In this section we will evaluate the clustering performance of our algorithms from two aspects, multi-view and single view clustering. It is worth noting that all datasets are sampled in the same processing mode during comparison experiments, including dimensions of input data, number of samples, number of views and pre-processing operations. Since the proposed method is based on an unsupervised SC algorithm, it does not require labeled or annotated training data. We use the existing implementations published by the corresponding references and the results of the MVC algorithm which are shown in Tables 2-5. The results of single view clustering are shown in Table 6 and Fig. 2. Some results of the comparison methods are taken directly from the corresponding references listed in the table.

For multi-view datasets, the proposed method is superior to all similar comparison methods in some evaluation matrices, i.e. ACC, F1-score, precision and ARI. Notably, the most obvious improvement is observed on the ORL_mtv dataset. Compared to the second-best experimental result (KPMLRSSC), we can find out that there is an increase about 6%, 3%, 7%, 5% and 7% in each of the evaluation indexes, i.e. ACC, NMI, F1-score, precision and ARI. On the contrary, the result of the RMKMC algorithm is the worst in the ORL_mtv dataset. ACC is around 20% and precision is only 9%. By comparing SRAGL-AWCL to KPMLRSSC and RMKMC, we can conclude that our proposed method has the advantage of using the SR method with AGL for each view. More specifically, the strategy utilized in the first-step of our method is not only better at removing redundant information, but also enhances the discrimination between samples and continuously optimizes the graph structure features of each view. Finally, we can obtain each view feature matrix with a complete internal feature structure.

Our algorithm also obtains great improvement on dataset Notting-Hill. Especially, the evaluation indicators of ACC and precision are improved by 4% and 5% from the second-best results, respectively. Furthermore, as shown in Table 4, the experimental results of some compared methods are unsatisfactory in some evaluation indicators, such as the second-best result MVGL in NMI (20.63%) and ARI (27.03%). From the experimental results obtained by the methods AMGL and MVGL, it is observed that the proposed SRAGL-AWCL method with a novel matrix decomposition outperforms all other comparison algorithms, which further justifies our two-step strategy. Additionally, the AWCL module also adaptively weights each view and makes good use of complementary information between views to merge into global feature information - this improves classification performance. In order to prove the generality of our proposed algorithm compared with other algorithms, we conducted some simple processing of the Catech-101 dataset, using the feature extraction algorithms (GABOR, HOG and LBP) of the images in this dataset. Our algorithm has slightly lower NMI (0.22%) than MVGL shown in Table 5. The reason is that the dataset UCI_Digits is simply a handwritten digital picture, which only contains fewer feature points. Additionally, the constraint of the original SR model (that all values of the sparse subspace matrix are greater than zero) will filter out some information during matrix decomposition. It could significantly affect the final performance of the cluster for dataset with a small number of characteristics in the sample, such as UCI_digit. On the other hand, MVGL directly integrates various view features from similarity measures without the SR processing through each view, which allows it to preserve complete feature information.

For single-view results, the experimental results on COIL20, USPS, Yale_32x32 and ORL_32x32 datasets are shown in Table 6 and Fig. 2. These results emphasize that our algorithm has generally better clustering effect on single-view datasets than baseline methods. The multi-view comparison methods are shown in Table 6, and the single-view comparison methods in Fig. 2. From the results shown in Table 6, important details can be extracted: SRAGL and SRAGL-AWCL demonstrate the same results in the single-view dataset COIL20. In another word, AWCL is not essential when it comes to the single-view dataset. In terms of the evaluation index NMI, the proposed method is very close to the second-best comparison method MVGL – it is only 0.41% higher. The reason is because most samples in COIL20 are with some simple structures and the required features for clustering are relatively simple. On the other hand, after the newly proposed SR model for feature extraction is used, the information content of each view feature becomes simpler. Fig. 2 provides the summary of the additional experiment results to compare our method with other single-view benchmark datasets, i.e. Yale_32x32, ORL_32x32, USPS and COIL20. It can be seen that our method has tremendously improved the performance against indexes of ACC and NMI. For example, our method has shown great improvement (ACC is 83%, NMI is 87%) compared to other algorithms results on the USPS dataset. Better results have been also achieved on Yale_32x32 (ACC is 51%, NMI is 59%) and ORL_32x32 (ACC is 62%, NMI is 81%) datasets.

Table 2: Results on ORL_mtv dataset (the best result is bolded)

Method	ACC(%)	NMI(%)	F1-score(%)	Precision(%)	ARI(%)	Time(second)
RMKMC	20.00±0.00	53.43±0.00	16.22±0.00	9.72±0.00	12.94±0.00	11.65±0.25
CRSC	72.07±2.52	86.50±1.78	63.75±3.55	59.17±3.44	62.85±3.64	6.54±0.21
AMGL	74.15±0.98	89.62±0.89	59.59±3.32	47.37±3.66	58.40±3.44	3.39±0.28
MVGL	75.25±0.00	88.82±0.00	48.72±0.00	34.76±0.00	47.05±0.00	3.62±0.22
MVCF	67.00±0.00	81.08±0.00	50.79±0.00	43.59±0.00	49.46±0.00	8.74±0.14
MVNMF	72.70±0.40	89.10±0.10	66.00±0.44	57.90±0.65	65.00±0.46	73.56±0.24
KPMLRSSC	76.70±2.83	89.46±1.46	70.30±3.41	66.17±3.93	69.57±3.49	2.04±0.36
KCMLRSSC	74.13±4.41	88.56±1.72	67.88±4.56	63.10±5.21	67.08±4.69	1.39±0.23
SRAGL	79.30±2.03	90.76±0.52	70.87±1.85	63.89±2.75	70.12±1.91	**1.21±0.12**
SRAGL-AWCL	**82.50±1.57**	**92.78±0.78**	**77.00±2.63**	**71.85±3.76**	**76.43±2.71**	3.44±1.25

Table 3: Results on Notting-Hill dataset (the best result is bolded)

Method	ACC(%)	NMI(%)	F1-score(%)	Precision(%)	ARI(%)	Time(second)
RMKMC	39.39±0.00	51.70±0.00	23.21±0.00	13.25±0.00	12.08±0.00	53.00±0.27
CRSC	60.94±0.85	76.94±1.24	52.60±1.80	59.36±1.82	49.36±1.90	23.88±0.52
AMGL	75.75±2.18	87.23±0.83	62.48±1.30	47.82±1.12	58.57±1.43	41.70±0.46
MVGL	81.76±0.00	88.83±0.00	71.76±0.00	64.81±0.00	69.31±0.00	39.35±0.47
MVCF	66.50±0.00	78.37±0.00	56.77±0.00	59.94±0.00	53.60±0.00	55.63±1.24
MVNMF	73.72±1.68	82.11±1.03	54.68±1.80	47.09±5.33	50.47±2.30	355.93±0.84
KPMLRSSC	63.84±2.98	78.49±1.58	56.04±2.70	63.26±2.18	53.04±2.82	110.25±1.58
KCMLRSSC	62.87±3.76	77.82±1.61	54.56±2.59	62.07±3.44	51.48±2.75	38.93±0.25
SRAGL	77.88±2.97	86.35±1.35	70.15±2.79	68.50±2.63	67.77±2.95	**8.91±0.12**
SRAGL-AWCL	**85.57±4.61**	**89.18±1.79**	**76.27±4.50**	**73.50±4.71**	**74.35±4.86**	32.56±0.34

Table 4: Results on Caltech-101 dataset (the best result is bolded)

Method	ACC(%)	NMI(%)	F1-score(%)	Precision(%)	ARI(%)	Time(second)
RMKMC	52.33±0.00	14.05±0.00	56.27±0.00	50.58±0.00	20.59±0.00	**3.41±0.39**
CRSC	28.07±0.65	12.70±1.34	30.04±0.79	48.62±0.88	6.81±0.79	9.76±0.06
AMGL	55.46±0.10	8.37±0.98	59.37±0.53	43.11±0.83	12.72±1.86	65.71±0.37
MVGL	55.46±0.00	20.63±0.00	62.40±0.00	51.02±0.00	27.03±0.00	48.73±0.56
MVCF	37.46±0.00	17.84±0.00	36.92±0.00	52.41±0.00	11.84±0.00	43.97±0.05
MVNMF	55.52±0.00	16.91±0.31	57.78±0.00	40.64±0.00	24.55±0.50	59.25±0.16
KPMLRSSC	20.08±1.39	11.01±1.32	19.95±1.57	46.15±3.20	3.02±1.72	53.47±1.29
KCMLRSSC	20.08±0.59	10.16±0.92	20.00±1.18	46.04±1.72	2.99±0.96	53.78±1.56
SRAGL	55.79±0.28	47.67±0.45	62.97±0.45	80.08±4.21	39.81±4.09	8.57±0.25
SRAGL-AWCL	**56.29±0.23**	**49.04±1.84**	**63.14±0.38**	**83.08±4.21**	**41.71±0.22**	9.96±0.14

Table 5: Results on UCI_Digits dataset (the best result is bolded)

Method	ACC(%)	NMI(%)	F1-score(%)	Precision(%)	ARI(%)	Time(second)
RMKMC	47.05±0.00	55.95±0.00	45.32±0.00	35.18±0.00	37.28±0.00	**42.85±0.37**
CRSC	75.23±3.87	72.79±2.64	68.21±3.47	66.45±3.74	64.58±3.90	278.68±0.46
AMGL	81.51±5.02	87.84±2.12	80.68±3.98	74.00±4.70	78.32±4.51	182.54±0.28
MVGL	85.35±0.00	**90.31±0.00**	84.69±0.00	78.66±0.00	82.85±0.00	341.60±1.28
MVCF	85.03±0.00	77.90±0.00	74.12±0.00	72.86±0.00	71.20±0.00	290.47±1.23

MVNMF	71.88±1.81	73.62±1.16	65.77±1.37	62.29±1.11	61.75±1.51	54.96±0.56
KPMLRSSC	78.35±5.16	76.98±1.73	72.08±3.16	70.43±4.68	68.90±3.61	181.02±1.25
KCMLRSSC	80.31±3.56	77.42±1.41	72.90±2.72	71.74±3.96	69.84±3.10	179.44±1.58
SRAGL	81.01±3.19	86.69±1.67	81.01±2.79	75.99±3.21	78.75±3.14	77.31±0.21
SRAGL-AWCL	**86.90±2.48**	90.09±2.02	**85.28±1.77**	**79.21±2.57**	**83.51±2.02**	208.21±0.50

Table 6: Results on COIL20 dataset (the best result is bolded)

Method	ACC(%)	NMI(%)	F1-score(%)	Precision(%)	ARI(%)	Time(second)
RMKMC	27.64±0.00	55.32±0.00	29.75±0.00	18.55±0.00	23.71±0.00	7.62±0.52
CRSC	64.00±0.96	76.87±0.35	60.29±0.76	59.91±0.99	58.13±0.81	10.76±0.26
AMGL	80.12±4.30	91.72±2.38	77.54±3.15	67.12±4.39	76.18±3.38	21.23±0.11
MVGL	87.08±0.00	95.95±0.00	84.84±0.00	75.57±0.00	83.96±0.00	**4.50±0.48**
MVCF	61.88±0.00	74.19±0.00	55.58±0.00	50.46±0.00	53.03±0.00	18.46±0.27
MVNMF	78.04±3.40	88.83±1.56	74.60±2.78	69.55±4.43	73.18±2.97	6.44±0.06
KPMLRSSC	62.88±3.36	75.75±1.68	58.76±2.72	56.24±3.06	56.62±2.88	15.48±0.51
KCMLRSSC	63.16±4.68	76.35±2.25	59.63±3.82	57.21±4.18	57.44±4.05	9.74±0.25
SRAGL	**90.76±0.00**	**96.38±0.00**	**89.60±0.00**	**85.44±0.00**	**89.04±0.00**	6.35±0.25
SRAGL-AWCL	**90.76±0.00**	**96.38±0.00**	**89.60±0.00**	**85.44±0.00**	**89.04±0.00**	9.54±0.47

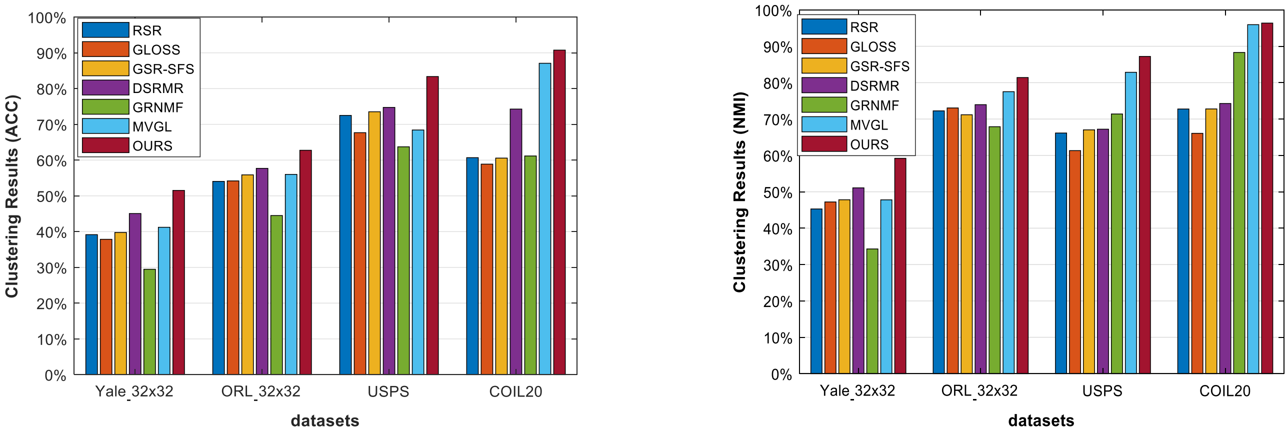


Figure 2: The comparison clustering results achieved by our method and other benchmark ones on single-view datasets.

5.5 Parameters λ and η Analysis

In the proposed model, we use SC of sparse matrices to obtain the final evaluation value. There are mainly two parameters λ and η in the two models SRAGL and SRAGL-AWCL that are affecting the performance. The λ is a regularization parameter for data smoothing and similarity matrix learning for sparse matrix in each view, and $\lambda = \lambda^{(v)}$ in all views. The η is a regularization parameter for learning the global Laplacian matrix L_*. In our experiments, an initial value of η is assigned to 1. After that, we update η by the following regulation: if the sum of the top c smallest eigenvalues of the global Laplacian matrix L_* equals to zero, we set $\eta = 1$; if the sum of the top c smallest eigenvalues of the global Laplacian matrix L_* is smaller than zero, we increase η; and if the sum of the top c smallest eigenvalues of the global Laplacian matrix L_* is greater than zero, we decrease η. An important fact that the parameter λ has a large influence on the initial value has been proved by our experiments. The specific details of this parameter are shown in Fig. 3. Three most representative datasets are selected from the multi-view dataset and the single-view dataset in order to show the effect of the initial value of the parameters λ on the clustering results. From Fig. 3, we can find out that when the value of λ is assigned to 0.01 or 0.001, it produces the best results. Thus, the initial value of parameter λ is set to 0.01 in the majority of the datasets except one single-view dataset USPS whose parameter λ is set to 0.001 in order to get the best results. However, the second best result can be achieved when setting λ to 0.01 for the dataset USPS in Fig. 3. Therefore, $\lambda = 0.01$ is normally chosen as the initial value of the parameter.

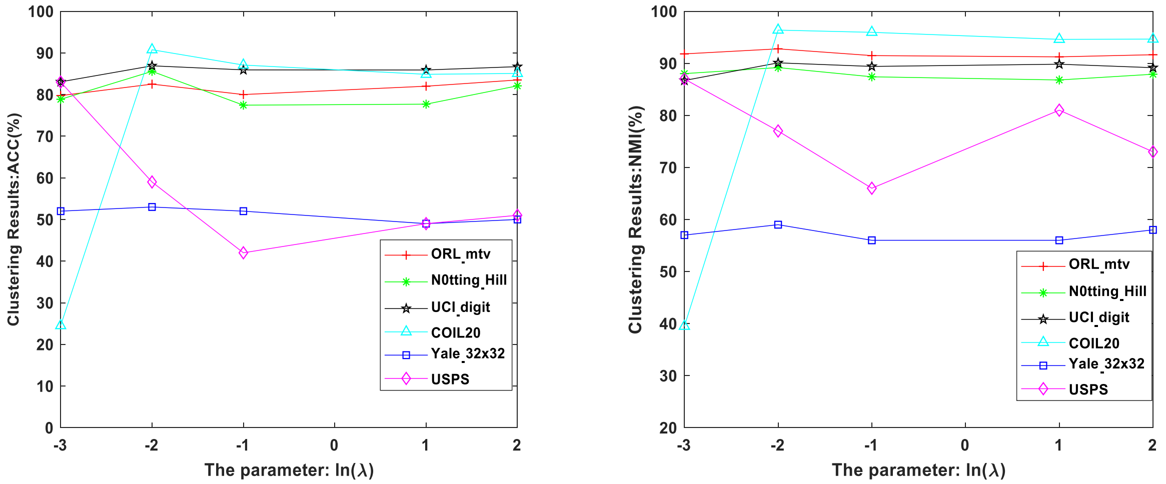
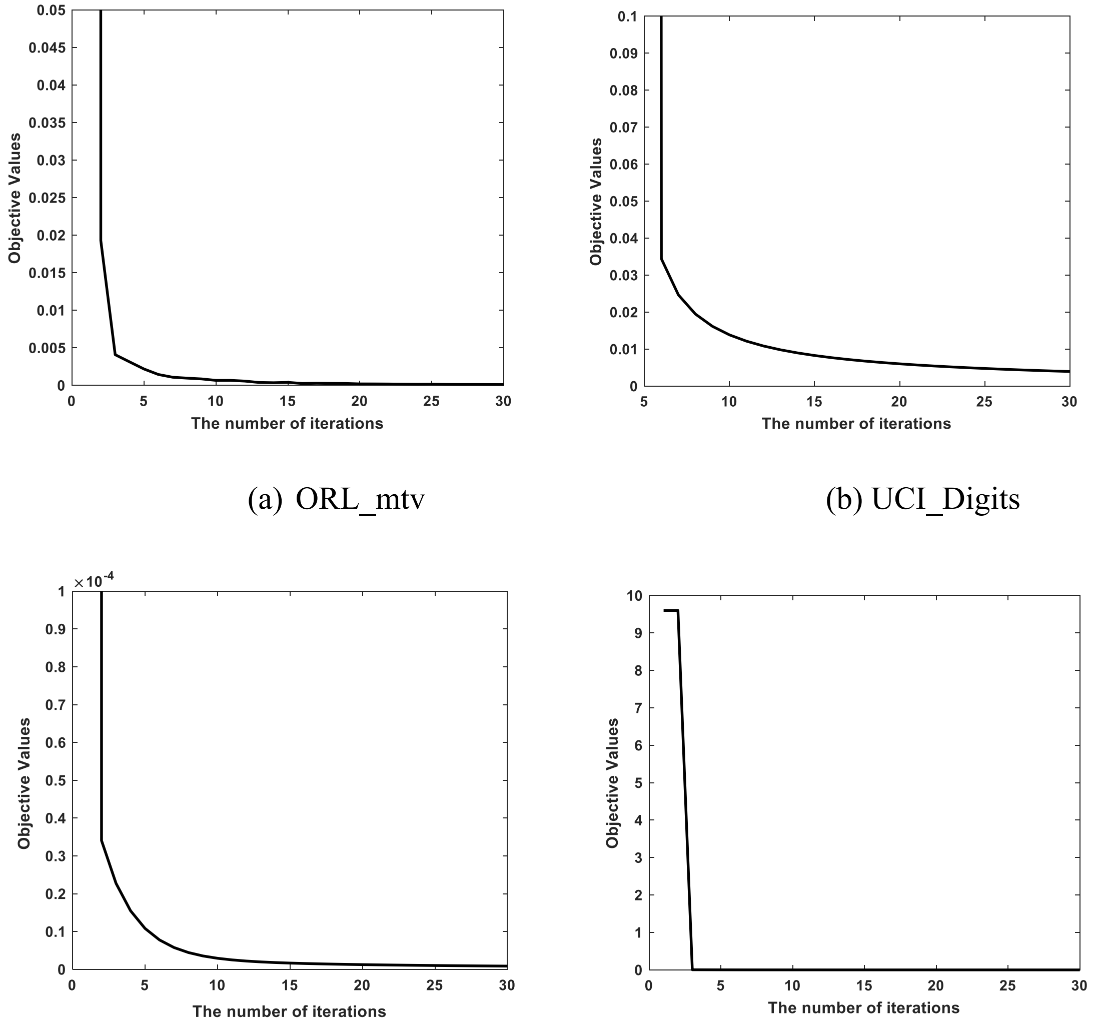


Figure 3: The effect of the initial value of the parameter λ on the overall clustering result. The abscissa represents the logarithm of the parameter ($\ln(\lambda)$). The left is the clustering indicator ACC, and right is the clustering indicator NMI.

5.6 Convergence Analysis

The convergence of Algorithm 3 can be proved theoretically by following [5]. To demonstrate the performance of the convergence with respect to SRAGL-AWCL, we use some examples illustrated in Fig. 4. In order to clearly show the convergent results, only the convergence graphs of the four real datasets are given. For other datasets, the algorithm has converged after one or two iterations. In Fig. 4, the x-axis and the y-axis are used to denote the number of iterations and the corresponding objective function cost respectively. As the number of iterations increases, the value of the objective function constantly decreases as shown in Fig 4 and the value of the objective function is also trending to be stable after several iterations. This result demonstrates the effectiveness of SRAGL-AWCL convergence. To ensure the generality, the maximum number of iterations is set to 30.



(a) ORL_mtv

(b) UCI_Digits

(c)Notting-Hill (d) COIL20

Figure 4: The convergent results of Algorithm 3 on the datasets ORL_mtv, UCI_Digits, Notting-Hill and COIL20.

6 CONCLUSION

In this paper, a novel method has been proposed to integrate SR with CL adaptively for MVSC. Both theoretical derivations and the pseudocode of algorithms have been given in detail. The convergence of the proposed method has also been theoretically proved.

The performance of the proposed algorithm has been tested on a wide range of benchmark datasets including both multi-view datasets and single-view datasets against evaluation index ACC, NMI, F1-score, precision, ARI and computational time. The results have shown that our algorithm performed much better than others on all different datasets. Specifically, our algorithm outperformed all the current multi-view methods and single-view methods on the ORL dataset, with increases at about 6%, 3%, 7%, 5% and 7% in the multi-view ORL_mtv dataset on each evaluation index. This is attributed to the preservation of the internal structural features and the fusion of complementary information between the views in global matrix. The performance in the dataset Notting-Hill, especially the evaluation indicators ACC and precision, has shown significant improvement by 4% and 5% compared to the second best results,. It is also worth noting that the specification parameters in the model are relatively stable in all experiments.

Our research will focus on the generalization of the algorithms in the near future by taking into account wider ranges of parameter adaptation, incomplete image recovery, the overall spatial structure and the information extraction of high dimensional data. Finally, the computational complexity of our algorithms is currently relatively high compared to the best results, and we are aiming to tackle in the future in order to improve our algorithm's time efficiency.

Acknowledgements

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

References

[1] T. Ahonen, A. Hadid, M. Pietikainen, “Face description with local binary patterns: Application to face recognition,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 28, pp. 2037-2041, 2006.

[2] S. Tai, “Image Representation Using 2D Gabor Wavelets,” IEEE Trans. Pattern Anal. Mach. Intell., pp. 959-971, 1996

[3] N. Dalal and B. Triggs, “Histograms of oriented gradients for human detection,” Proc. of the IEEE Conf. Computer Vision and Pattern Recognition (CVPR), pp. 886-893, 2005.

[4] S. Tulsiani, T. Zhou, A. A. Efros, and J. Malik, “Multi-view supervision for single-view reconstruction via differentiable ray consistency,” Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 2626-2634, 2017.

[5] F. Nie, G. Cai , Jing Li , and X. Li, “Auto-weighted multi-view learning for image clustering and semi-supervised classification,” IEEE Trans. on Image Processing, vol. 27, no. 3, pp. 1501-1511, March 2018.

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

[7] X. Wang, W. Liu, J. Li, and X. Gao. “A novel dimensionality reduction method with discriminative generalized eigen-decomposition,” Neurocomputing, vol. 173, pp. 163-171, 2016.

[8] 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, pp. 4833-4843, Oct. 2018.

[9] K. Zhan, C. Niu, C. Chen, et al. “Graph structure fusion for multiview clustering,” IEEE Trans. on Knowledge and Data Engineering, vol. 31, pp. 1984-1993, 2018.

[10] K. Zhan, F. Nie, J. Wang, et al. “Multiview consensus graph clustering,” IEEE trans. on image process., vol. 28, pp. 1261-1270, 2019.

[11] R. Zhang, F. Nie, M. Guo, and X. Wei, “Joint learning of fuzzy *k*-means and nonnegative spectral clustering with side information,” IEEE Trans. on Image Processing, vol. 28, pp. 2152-2162, 2018.

[12] D. Cai, X. He, J. Han, et al. “Graph regularized nonnegative matrix factorization for data representation,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 33, pp. 1548-1560, 2011.

[13] Y. Feng, J. Xiao, Y. Zhuang, et al. “Adaptive unsupervised multi-view feature selection for visual concept recognition,” Proc. of the Asian Conf. on Computer Vision. Springer, Berlin, Heidelberg, vol. 7724, pp. 343-357, 2012.

[14] K. Zhan, J. Shi, J. Wang, et al. “Adaptive structure concept factorization for multiview clustering,” Neural Computation, vol. 30, pp. 1080-1103, 2018.

[15] Y. Zhao, X. You, S. Yu, et al. “Multi-view manifold learning with locality alignment,” Pattern Recognit., vol. 78, pp. 154-166, 2018.

[16] X. Wang, Z. Lei, X. Guo, et al. “Multi-view subspace clustering with intactness-aware similarity,” Pattern Recognit., vol. 88, pp. 50-63, 2019.

[17] A. Kumar, P. Rai, and H. Daumé, “Co-regularized multi-view spectral clustering,” Proc. of the Advances in Neural Information Processing Systems 24 (NIPS), pp. 1413-1421, 2011.

[18] M. Brbic, and I. Kopriva, “Multi-view low-rank sparse subspace clustering,” Pattern Recognit., vol. 73, pp. 247-258, 2018.

[19] F. Nie, J. Li, X. Li, “Parameter-free auto-weighted multiple graph learning: a framework for multiview clustering and semi-supervised classification,” Proc. of the International Joint Conference on Artificial Intelligence (AAAI), pp. 1881-1887, 2016.

[20] K. Zhan, C. Zhang, J. Guan, and J. Wang, “Graph learning for multiview clustering,” IEEE Trans. Cybern., vol. 48, no. 10, pp. 2887-2895, Oct. 2018.

[21] B. Mohar, Y. Alavi, G. Chartrand, and O. Oellermann, “The Laplacian spectrum of graphs,” Graph Theory Combinatorics Appl., vol. 2, no. 12, pp. 871-898, 1991.

[22] F. R. Chung, “Spectral graph theory,” Providence, RI, USA: Amer. Math. Soc., 1997.

[23] R. Tarjan, “Depth-first search and linear graph algorithms,” SIAM J. Comput., vol. 1, no. 2, pp. 146-160, 1972.

[24] K. Fan, “On a theorem of Weyl concerning eigenvalues of linear transformations I,” Proc. Nat. Acad. Sci. USA, vol. 35, no. 11, pp. 652-655, 1949.

[25] M. L. Overton and R. S. Womersley, “On the sum of the largest eigenvalues of a symmetric matrix,” SIAM J. Matrix Anal. Appl., vol. 13, no. 1, pp. 41-45, 1992.

[26] P. K. Chan, M. D. F. Schlag, and J. Y. Zien, “Spectral K-way ratio-cut partitioning and clustering,” IEEE Trans. Comput.-Aided Design Integr. Circuits Syst., vol. 13, no. 9, pp. 1088-1096, Sep. 1994.

[27] S. Boyd and L. Vandenberghe, “Convex Optimization, Cambridge. U.K.: Cambridge Univ, 2004.

[28] B. Wu, Y. Zhang, B.-G. Hu, and Q. Ji, “Constrained clustering and its application to face clustering in videos,” Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 3507-3514, 2013.

[29] X. Cao, C. Zhang, C. Zhou, H. Fu, and H. Foroosh, “Constrained multi-view video face clustering,” IEEE Trans. Image Process., vol. 24, no. 11, pp. 4381-4393, Nov. 2015.

[30] L. Fei-Fei, R. Fergus, and P. Perona, “Learning generative visual models from few training examples: An incremental Bayesian approach tested on 101 object categories,” Comput. Vis. Image Understand., vol. 106, no. 1, pp. 59-70, 2007.

[31] S. A. Nene, S. K. Nayar, and H. Murase, “Columbia object image library (COIL-20),” Dept. Comput. Sci., Columbia Univ., New York, NY, USA, Tech. Rep. CUCS-005-96, 1996.

[32] X. Cai, F. Nie, and H. Huang, “Multi-view K-means clustering on big data,” Proc of the International Joint Conf. on Artificial Intelligence (IJCAI), vol. 23, pp. 2598-2604, 2013.

[33] C. Tang, X. Liu, M. Li, et al, “Robust unsupervised feature selection via dual self-representation and manifold regularization,” Knowledge-Based Systems, vol. 145, pp.109-120, 2018.

[34] W. He, X. Zhu, D. Cheng, R. Hu, S. Zhang, “Unsupervised feature selection for visual classification via feature-representation property,” Neurocomputing, vol. 236, pp. 5-13, 2017.

[35] N. Zhou, Y. Xu, H. Cheng, J. Fang, W. Pedrycz, “Global and local structure preserving sparse subspace learning: an iterative approach to unsupervised feature selection,” Pattern Recognit., vol. 53, pp. 87-101, 2016.

[36] P. Zhu, W. Zuo, L. Zhang, Q. Hu, S.C.K. Shiu, “Unsupervised feature selection by regularized self-representation,” Pattern Recognit., vol. 48, pp. 438-446, 2015.

[37] J. Liu, C Wang, J Gao, and J Han, “Multi-view clustering via joint nonnegative matrix factorization,” Proc. of the SIAM International Conf. on Data Mining, 2013.

[38] R. Zhu, F. Dornaika, Y. Ruichek. “Joint graph based embedding and feature weighting for image classification,” Pattern Recognit., vol. 93, pp.458-469, 2019.

[39] X. You, J. Xu, W. Yuan, et al. “Multi-view global component discriminant analysis for cross-view classification,” Pattern Recognit., vol. 92, pp. 37-51, 2019.

[40] J. Wen, Y. Xu, and H. Liu. “Incomplete multiview spectral clustering with adaptive graph learning,” IEEE Trans. on cybernetics, vol. 50, pp. 1418 – 1429, 2018.