A Unified Collaborative Multi-kernel Fuzzy Clustering for Multiview Data

Shan Zeng, Xiuying Wang, *Member, IEEE,* Hui Cui, Chaojie Zheng, David Feng, *Fellow, IEEE*

***Abstract*—Clustering is increasingly important for multiview data analytics and current algorithms are either based on the collaborative learning of local partitions or directly derived global clustering from multi-kernel learning. In this paper, we innovate a clustering model that unifies the local partitions and global clustering in a collaborative learning framework. We firstly construct a common multi-kernel space (CMKS) from a set of basis kernels to better reflect clustering information of each individual view. Then, considering that joint local partitions would conform to the global clustering, we fuse the local partitions and global clustering guidance as a single objective function in accordance with fuzzy clustering form. The collaborative learning strategy enables the mutual and interactive clustering from local partitions and global clustering. The validation was performed over two synthetic and four public databases and the clustering accuracy was measured by NMI and RI. The experimental results demonstrated that the proposed algorithm outperformed the related state-of-the-art algorithms in comparison which included multitask, multi-kernel and multiview clustering approaches.**

***Index Terms*— Multiview data, Fuzzy clustering, Collaborative learning, Multi-kernel space**

I. Introduction

Clustering is essential to unsupervised image segmentation and interpretation, content-based retrieval, computer vision and visual analytics by classifying the image contents into disjointed groups based on predefined similarity criteria. The conventional clustering algorithms, such as k-means [1], fuzzy c-means (FCM) [2-3], spectral clustering [4], maximum entropy clustering (MEC) [5] and possibilistic fuzzy c-means (PFCM) [6-7] receive tremendous popularity due to their simplicity and rapid computation. As the rapid advance of imaging and data acquisition techniques, nowadays, more complicated multimodality data are generated on a daily basis from areas such as healthcare, finance, social network and scientific research. Multiple representations and descriptors become indispensable to reflect the embedded information for these complicated multiview data. The conventional clustering methods mainly for single attribute data or single view data are becoming less feasible for these multiview data.

The multiview clustering was initially addressed by Bickel *et al*. [8]. The early collaborative clustering for multi-view included EM-based method [9] and spectral clustering algorithms [10-11]. Based on k-means clustering, Chen et al. [12] proposed a weighted clustering algorithm for multiview data to automatically compute weights for each individual view and for the variables as well. Liu *et al*. [13] modeled multiview data as a tensor and developed a new tensor-based framework for the integration of heterogeneous multiview data in the context of spectral clustering. More recently, new learning technologies [14-19] have been introduced for multi-view data analysis and processing. When seeking to achieve consistent and common conclusions, these learning models also take into account the differences among multiple views. The collaborative learning outcomes provide important guidance for multi-view clustering algorithm [20].

Soft or fuzzy clustering mechanism is particularly suitable for multi-view data to extract the latent common partition from a fusion of fuzzy clustering results of each individual view. According to the learning strategies for latent partition, the fuzzy multi-view clustering methods can be classified into two categories, collaborative learning with adaptive fusion of the partition from an individual view, and multi-kernel learning with fusion based on distance measurements.

In the first category, a collaborative fuzzy clustering (CoFC) algorithm was firstly proposed [21] based on the standard FCM method. This method extracted the common structure from separate subsets of patterns that collaborated by exchanging information of local partition matrices. Pedrycz and the team continued to further improve the CoFC algorithm in [22-23], and their systematic research provided the foundation for collaborative multi-view learning. Cleuziou *et al*. proposed a centralized model for multi-view clustering, called Co-FKM [24], in which a penalty term was introduced to reduce the disagreement between the partitions on different views. Then, in order to identify the importance of each view and enhance the clustering performance, Jiang *et al*. [25] proposed a weighted view collaborative fuzzy c-means (WV-Co-FCM) algorithm based on Shannon entropy. In these algorithms, for each individual view, the partition was achieved on the basis of collaborative learning with adaptive-fusion. However, the final clustering result was obtained by simple integration or voting

S. Zeng is with College of Mathematics and Computer Science, Wuhan Polytechnic University, Wuhan, China (e-mail: zengshan1981@whpu.edu.cn).

X. Wang (e-mail: xiu.wang@sydney.edu.au), H. Cui and C. Zheng are with the Biomedical & Multimedia Information Technologies (BMIT) research group, The University of Sydney, Australia.

D. Feng is with the BMIT research group, The University of Sydney and with the Med-X Research Institute, Shanghai Jiao Tong University.

This work is supported by Hubei natural science foundation (NSF) (Grant No. 2017CFB500) and NSFC (Grant No. 61303116).

mechanism. In addition, the global guidance was not considered in the clustering process of each view. At last, these algorithms generally adopted a uniform distance measure, such as the Euclidean distance. However, in the complicated multiview data, the distance measurements of different views may vary significantly and thereby a single distance measure is far from sufficient.

The second category of multiview clustering methods are fundamentally based on the fact that kernel methods provide an intuitive way to merge and integrate different types of data, and the features of each view can be processed into a kernel matrix [26]. A multiple kernel interval-valued fuzzy C-means clustering (MKIFCM) algorithm [27] was based on the kernel learning method and the interval-valued fuzzy sets. Huang et al. [28] proposed a multiple kernel fuzzy c-means (MKFCM) algorithm that extended the FCM algorithm with a multiple kernel learning setting. Chen et al. [29] used the multiple kernel fuzzy c-means algorithm for image segmentation, in which, different pixel information represented by different kernels was combined in the kernel space to produce a new kernel. By incorporating multiple kernels and automatically adjusting the kernel weights, these two MKFC algorithms were more immune to ineffective kernels and irrelevant features. Baili *et al*. [30] extended the kernel fuzzy c-means clustering algorithm to an adaptive clustering model and proposed a fuzzy c-means algorithm with Multiple Kernels (FCM-MK). Tzortzis *et al*. [31] introduced the multi-kernel mechanism into multiview clustering. Combining the features of each view as a vector by multi-kernel mapping, these methods overlooked uniqueness of an individual view and omitted the collaborative learning between each view in the clustering process. Since multiple-kernel clustering is highly related to identification of fuzzy model, the fuzzy clustering algorithm was extended to an adaptive clustering model by the recursive processing strategy [32-34] for stream data clustering. The model was demonstrated on different applications including monitoring of the waste water treatment process.

As discussed above, clustering methods based on collaborative learning do not utilize global clustering estimation as guidance for searching optimal clustering results. Comparatively, the algorithms based on multi-kernel directly derive the global solution from the vector of combined features of different views, yet without collaborative learning. In this paper, we propose a collaborative multi-kernel fuzzy clustering (CoMK-FC) model that unifies global clustering guidance with the local partitions in a single objective function. We introduce multi-kernel learning for each view and construct a common multi-kernel space. Then, the global clustering can be directly derived from the composite kernel space, and the local partition is generated from the collaborative learning of these composite kernels. Hence, this strategy is able to waive issue of the uniform distance measurement in the collaborative learning process. Further, based on the hypothesis that when the clustering achieves an optimal solution, the joint local partitions would conform to the global clustering, we fuse the local partitions and global clustering guidance as a unified objective function.

The rest of this paper is organized as follows. In Section II, we briefly review some multiview fuzzy clustering algorithms in the latest advance, such as the Co-FKM algorithm, the WV-Co-FCM algorithm and the MKFC algorithm. In Section III, our method, CoMK-FC, is proposed. In Section IV, the experimental results of our method are compared with the state-of-the-art algorithms to prove the effectiveness of CoMK-FC. Finally, this paper concludes in Section V.

II. Related Work

In this section, we briefly review some related work on collaborative learning and multi-kernel learning for fuzzy clustering.

A. Collaborative Fuzzy Clustering

For a given sample set $X=\{x_1,\cdots,x_N\}$ with N sample elements, we assume that X is described in K different views. Correspondingly, for x_j, a vector $x_j^{(k)}$ is defined to represent its kth view features. To cluster the sample set X into C classes, the fuzzy partition matrix of N samples belonging to C classes is represented as $U^{(k)}=[u_{ij,k}]_{C\times N}$ in the k th view, where $u_{ij,k}\in[0,1]$ $(1\le i\le C, 1\le j\le N)$ denotes the fuzzy membership degree of the sample x_j belonging to the cluster i of kth view. $u_{ij,k}$ should satisfy the following two constraints:

$$\sum_{i=1}^{C} u_{ij,k}=1 \tag{1}$$

$$0\le u_{ij,k}\le 1, 1\le i\le C, 1\le j\le N, 1\le k\le K \tag{2}$$

Pedrycz *et al*. [21] defines the objective function of the collaborative fuzzy clustering (CoFC) algorithm as:

$$\begin{aligned} J_{CoFC}(k)=&\sum_{i=1}^{C}\sum_{j=1}^{N} u_{ij,k}^{m} d_{ij,k}^{2} \\ &+\sum_{k'=1}^{K}\alpha_{k,k'}\sum_{i=1}^{C}\sum_{j=1}^{N}(u_{ij,k}-u_{ij,k'})^{m} d_{ij,k}^{2} \end{aligned} \tag{3}$$

where $d_{ij,k}^2=\| x_j^{(k)}-v_i^{(k)}\|^2$ denotes the distance between x_j and $v_i^{(k)}$ in k th view, $v_i^{(k)}$ denotes the prototype of cluster i of k th view and $\alpha_{k,k'}$ is the collaborative coefficient between the k th view and k' th view, and $m=2$ is the fuzzification degree.

There are two terms in the objective function as defined in (3): the first term can be regarded as an individual clustering mechanism for each view by standard FCM algorithm, and the second term is a collaborative learning process among different views which was implemented by exchanging information from local partition matrices. The CoFC algorithm is an iterative process to minimize the objective function $J_{CoFC}(k)$. Fixing fuzzification degree at 2 as in (3) makes the optimization process possible; however, the clustering results may not be satisfactory.

B. Co-FKM Algorithm

On the basis of CoFC algorithm, Cleuziou *et al.* [24] proposed a collaborative approach, namely Co-FKM, by using conventional FCM framework. In this method, a specific partition was obtained in each view, and then a penalty term was introduced to reduce the disagreement between partitions from the different views. The objective function of the Co-FKM algorithm was defined as:

$$J_{Co-FKM}(U,V)=\sum_{k=1}^{K}\sum_{i=1}^{C}\sum_{j=1}^{N}u_{ij,k}^{m}d_{ij,k}^{2}+\eta\frac{1}{K-1}\sum_{k'=1,k'\neq k}^{K}\sum_{i=1}^{C}\sum_{j=1}^{N}(u_{ij,k'}^{m}-u_{ij,k}^{m})d_{ij,k}^{2}$$
$$=\sum_{k=1}^{K}\sum_{i=1}^{C}\sum_{j=1}^{N}\tilde{u}_{ij,k,\eta}d_{ij,k}^{2} \quad (4)$$

where $\tilde{u}_{ij,k,\eta}=(1-\eta)u_{ij,k}^{m}+\frac{\eta}{K-1}\sum_{k'=1,k'\neq k}^{K}u_{ij,k'}^{m}$ presents the weighted mean of the usual fuzzy memberships $u_{ij,k}^{m}$ obtained from each view, η is a parameter used to control the penalty correlated to the disagreement. The second term of the first equation of this objective function is a penalty term for the co-learning for any pair of two different views (k,k'). Since the lower the value of $(u_{ij,k'}^{m}-u_{ij,k}^{m})$ is, the lower the disagreement, it can be considered as a divergence between partitions from the different views. In order to merge each view partition $u_{ij,k}$ and obtain the global clustering result $\hat{u}_{ij}$, Co-FKM defined the geometric mean of $u_{ij,k}$ for each view as $\hat{u}_{ij}$ where

$$\hat{u}_{ij}=\sqrt[K]{\prod_{k=1}^{K}u_{ij,k}} \quad (5)$$

and assigned x_j to the i th cluster when $\hat{u}_{ij}$ is maximized.

Co-FKM improved the performance of the multiview clustering, and as indicated by (4), Co-FKM considered that each view contributed equally for clustering, which might not be always appropriate in particular when the views have different importance.

C. WV-Co-FCM Algorithm

Jiang *et al.* [25] firstly introduced the Shannon entropy to identify the importance of each view and proposed a weighted multiview collaborative fuzzy c-means (WV-Co-FCM) algorithm. The objective function of WV-Co-FCM was defined as

$$J_{WV-Co-FCM}(U,V,W)=\sum_{k=1}^{K}w_k\left[\sum_{i=1}^{C}\sum_{j=1}^{N}u_{ij,k}^{m}d_{ij,k}^{2}+\Delta_k\right]+\lambda\sum_{k=1}^{K}w_k\ln w_k$$
$$\Delta_k=\sum_{j=1}^{N}\alpha_{ij,k}\sum_{i=1}^{C}u_{ij,k}(1-u_{ij,k}^{m-1})-\sum_{j=1}^{N}\beta_{ij,k}\sum_{i=1}^{C}u_{ij,k}(1-u_{ij,k}^{m-1}) \quad (6)$$
$$s.t.\ u_{ij,k}\in[0,1],w_k\in[0,1],\sum_{k=1}^{K}w_k=1,\sum_{i=1}^{C}u_{ij,k}=1,1\le j\le N,1\le k\le K$$

where w_k is the weight for the kth view, $\alpha_{ij,k}=\eta d_{ij,k}^{2}$, $\beta_{ij,k}$ has four cases including $\beta_{ij,k}=\frac{\eta}{K-1}\sum_{k'=1,k'\neq k}^{K}d_{ij,k'}^{2}$, $\beta_{ij,k}=\frac{\eta}{K}\sum_{k=1}^{K}d_{ij,k}^{2}$, $\beta_{ij,k}=\eta\min_{k'\neq k}\{d_{ij,k'}^{2}\}$ and $\beta_{ij,k}=\eta\sqrt[K-1]{\prod_{k'\neq k}d_{ij,k'}^{2}}$. The two parameters η and λ where $0<\eta<1$, $\lambda>0$ were used to adjust the penalty corresponding to the partition disagreement and the weight of each view respectively.

In order to obtain the final global clustering result, the summation of each weighted fuzzy partition matrix for each view was adopted as

$$U=\sum_{k=1}^{K}w_kU_k \quad (7)$$

where U_k is the fuzzy membership matrix for the k th view.

D. Clustering based on multiple kernel fuzzy C-means

The multiple kernel fuzzy C-means (MKFCM) [28-30] is a multiple kernel learning algorithm which extends the conventional fuzzy C-means algorithm. A non-negative linear expansion of the bases in kernel space was defined as

$$\psi(x)=\sum_{k=1}^{M}\omega_k\psi_k(x) \quad (8)$$

where ψ_k is mapping function, ω_k is the weight of ψ_k and M is the number of mapping functions.

The objective function of MKFCM algorithm was defined as below:

$$J_{MKFCM}(w,U,V)=\sum_{i=1}^{C}\sum_{j=1}^{N}u_{ij}^{m}\|\psi(x_j)-v_i\|^2=\sum_{i=1}^{C}\sum_{j=1}^{N}u_{ij}^{m}\|\sum_{k=1}^{M}\omega_k\psi_k(x_j)-v_i\|^2 \quad (9)$$
$$s.t.\ u_{ij}\in[0,1],\omega_k\in[0,1],\sum_{k=1}^{M}\omega_k=1,\sum_{i=1}^{C}u_{ij}=1,1\le j\le N,1\le k\le M$$

where v_i denotes the prototype of the i th cluster in the implicit feature space, $\omega=(\omega_1,\omega_2,\cdots,\omega_M)^T$ is a weight vector, $U\in\mathbb{R}^{N\times C}$ denotes the fuzzy partition matrix whose elements are the membership u_{ij}, and V is a prototype matrix where each row corresponds to a cluster prototype.

Tzortzis *et al.* [31] proposed a kernel-based weighted multiview clustering (MVKKM) algorithm. This algorithm firstly assigned a kernel mapping for each view, and then defined the kernel combination as the following (10) to take advantage of all views:

$$\kappa_{com}=\sum_{k=1}^{K}w_k^{b}\kappa^{(k)},w_k\ge0,\sum_{k=1}^{K}w_k=1,b>1 \quad (10)$$

where w_k is the weight of the kth view, $\kappa^{(k)}$ denotes the kernel function of the kth view and b is an exponent.

As reviewed above, there has been systematic research on collaborative learning and multiple kernel learning for fuzzy clustering. Based on conventional FCM, CoFC initialized collaborating local partition matrices to extract common structure. While contributing to ease optimization process, fixed fuzzification degree was considered a major reason leading to less satisfactory clustering. To improve the performance of multiview clustering, Co-FKM introduced a penalty term to reduce disagreement between each view. However, treating different views with equal importance, this method may not

always achieve clustering convergence, in particular when some views are noisy and not reliable. WV-Co-FCM was then proposed with the entropy regularization term to adjust the weights of different views. WV-Co-FCM mainly used these weights to produce final clustering consensus and ignored their influence in the clustering procedure. These collaborative learning algorithms often adopt uniform distance measure for each view, which may raise difficulty when dealing with complicated data with large differences in data structures of different views. More importantly, these collaborative learning based algorithms might be trapped into local optimization.

Based on the multiple kernel learning strategies, the algorithms such as MKFCM and MVKKM derived global clustering by the multi-kernel mapping during the clustering process. The MKFCM algorithm was more immune to ineffective kernels and irrelevant features by incorporating multiple kernels and automatically adjusting the kernel weights. However, MKFCM neither considered the uniqueness of each view, nor the collaborative learning between each view in the clustering process.

Our hypothesis is that a clustering model, which unifies the global guidance with local partitions into a collaborative framework, will achieve optimal clustering.

III. Our Method

In this paper, we propose a collaborative multi-kernel fuzzy clustering (CoMK-FC) algorithm for multiview data, which takes into account both local partition and global clustering for an optimal ultimate solution. As illustrated in Fig. 1, based on a common multi-kernel space, our algorithm incorporates the advantages of collaborative learning and multiple kernels learning. In the proposed model, we not only consider the collaborative learning between each view, but also consider the global clustering guiding each view partition in the clustering process.

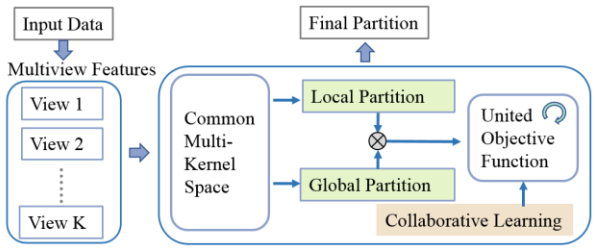


Fig. 1. Schematic framework of our proposed model

A. Common Multi-Kernel Space

How to effectively evaluate the distance between data items of each view and fuse the clustering results from each view are two important while challenging factors in multiview clustering.

Kernel-based clustering provides an effective mechanism for non-linear analysis and modeling. Different kernel functions have their specific characteristics. For instance, Gaussian kernel has merits on extracting local information while in the contrary, polynomial kernel exhibits strong capabilities of presenting global features. In order to effectively measure complex data structures of each view, we construct a common multi-kernel space on the basis of MKFCM algorithm [28] which is more immune to ineffective kernels and irrelevant features via automatically adjusting the kernel weights. The common multi-kernel space is constructed as shown in Fig. 2.

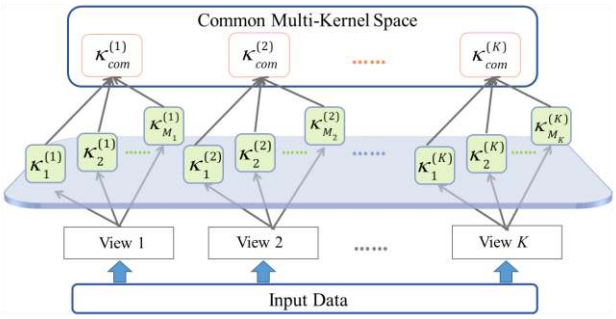


Fig. 2. Schematic diagram of the common multi-kernel space construction

In the k th view, considering a set of M_k implicit mapping functions $\Phi^{(k)}=\{\phi_1^{(k)},\phi_2^{(k)},\ldots,\phi_{M_k}^{(k)}\}$, each mapping function $\phi_h^{(k)}$ ($h=1,2,\ldots,M_k$) maps the L-dimension data $x_j^{(k)}$ as a vector $\phi_h^{(k)}(x_j^{(k)})$ in its feature space with L_h dimensions. The set $\{\kappa_1^{(k)},\kappa_2^{(k)},\ldots,\kappa_{M_k}^{(k)}\}$ denotes the Mercer kernels corresponding to these implicit mapping functions respectively as explicit function $\kappa_h^{(k)}(x_j^{(k)},x_{j'}^{(k)})=\phi_h^{(k)}(x_j^{(k)})^T\phi_h^{(k)}(x_{j'}^{(k)})$. For simplification, when there is no ambiguity in index of $x_j^{(k)}$ in $\phi_h^{(k)}(x_j^{(k)})$ and $\kappa_h^{(k)}(x_j^{(k)},x_{j'}^{(k)})$, the view index k will be omitted, i.e. $\phi_h^{(k)}(x_j)$ and $\kappa_h^{(k)}(x_j,x_{j'})$.

To combine these kernels and ensure that the resulting kernel still satisfies the Mercer condition, we propose a convex combination of these feature maps $\phi_{com}^{(k)}$ as below:

$$\phi_{com}^{(k)}(x_j)=\sum_{h=1}^{M_k}\sqrt{(\omega_h^{(k)})^{b_k}}\,\phi_h^{(k)}(x_j),\omega_h^{(k)}\geq 0,b_k\geq 1 \qquad (11)$$

where $\omega_h^{(k)}\geq 0$ is the weight of the h th mapping function $\phi_h^{(k)}$ and b_k is an exponent.

As these implicit mappings do not necessarily have same dimensionality, we construct a new set of independent mappings with identical dimensionality for the kth view $\Psi^{(k)}=\{\psi_1^{(k)},\psi_2^{(k)},\ldots,\psi_{M_k}^{(k)}\}$ from the original mappings $\Phi^{(k)}$ such that

$$\psi_1^{(k)}(x)=\begin{bmatrix}\phi_1^{(k)}(x)\\0\\\vdots\\0\end{bmatrix},\psi_2^{(k)}(x)=\begin{bmatrix}0\\\phi_2^{(k)}(x)\\\vdots\\0\end{bmatrix},\ldots,\psi_{M_k}^{(k)}(x)=\begin{bmatrix}0\\0\\\vdots\\\phi_{M_k}^{(k)}(x)\end{bmatrix} \qquad (12)$$

Each of these mappings projects $x_j^{(k)}$ into a L-dimension vector $L=\sum_{h=1}^{M_k}L_h$. These new mappings form a new set of orthogonal bases since

$$\psi_h^{(k)}(x_j)^T\psi_{h'}^{(k)}(x_{j'})=\begin{cases}\kappa_h^{(k)}(x_j,x_{j'}), & h=h'\\0, & h\neq h'\end{cases} \qquad (13)$$

By such, the feature spaces of these new mappings have the same dimension and their linear combination can be well defined. A non-negative linear expansion of the bases in $\Psi^{(k)}$ is defined as

$$\psi_{com}^{(k)}(x)=\sum_{h=1}^{M_k}\sqrt{(\omega_h^{(k)})^{b_k}}\psi_h^{(k)}(x) \tag{14}$$

Therefore, the corresponding composite kernel $\kappa_{com}^{(k)}$ in the k th view is defined as:

$$\kappa_{com}^{(k)}=(\omega_1^{(k)})^{b_k}\kappa_1^{(k)}+(\omega_2^{(k)})^{b_k}\kappa_2^{(k)}+\cdots+(\omega_{M_k}^{(k)})^{b_k}\kappa_{M_k}^{(k)} \tag{15}$$

where $\omega_h^{(k)}\geq 0$ represents the weight of $\kappa_h^{(k)}$ and $b_k>1$ is a coefficient analogous to the fuzzy coefficient m in (4). The regulation on weights is $\sum_{h=1}^{M_k}\omega_h^{(k)}=1$.

B. Objective Clustering Function

On the basis of the constructed common multi-kernel space, we take into account both the local partition from collaborative learning and global clustering from multi-kernel learning and define our objective function as below:

$$J\ (U,V,W)=J_{local}(U,V,W)+\alpha\cdot J_{global}(U,V,W) \tag{16}$$

where α is a trade-off parameter, $J_{local}(U,V,W)$ denotes the local partition from each view partition and $J_{global}(U,V,W)$ is the global clustering obtained from multi-kernel learning serving as a guidance. As we consider that the local partition and the global clustering are equally important for the ultimate clustering, we set $\alpha=1$ in the model.

1) Local partition function

Local partition for each individual view is achieved by collaborative learning and defined on the common multi-kernel space as

$$\begin{aligned}J_{local}(U,V,W)&=\sum_{k=1}^{K}\sum_{i=1}^{C}\sum_{j=1}^{N}\tilde{u}_{ij,k,\eta}\,\|\psi_{com}^{(k)}(x_j)-v_i^{(k)}\|^2\\ \tilde{u}_{ij,k,\eta}&=(1-\eta)u_{ij,k}^m+\frac{\eta}{K-1}\sum_{k'=1,k'\neq k}^{K}u_{ij,k'}^m\end{aligned} \tag{17}$$

Similar to the algorithm proposed in [25], we introduce the Shannon entropy to identify the importance of each view. The local partition function in (17) is then defined as:

$$\begin{aligned}J_{local}(U,V,W)=&\sum_{k=1}^{K}w_k\sum_{i=1}^{C}\sum_{j=1}^{N}\tilde{u}_{ij,k,\eta}\,\|\psi_{com}^{(k)}(x_j)-v_i^{(k)}\|^2\\&+\lambda\sum_{k=1}^{K}w_k\ln w_k\end{aligned} \tag{18}$$

where w_k is the weight of the k th view, $v_i^{(k)}$ denotes the prototype of cluster i for the k th view.

2) Global clustering function

On the basis of common multi-kernel space, we can directly derive the global clustering which serves as global guidance for local partitions. We define the global clustering function in the common kernel space as:

$$J_{global}(U,V,W)=\sum_{i=1}^{C}\sum_{j=1}^{N}u_{ij}^m\,\|\Psi_{com}(x_j)-\tilde{v}_i\|^2 \tag{19}$$

where u_{ij} denotes the fuzzy membership degree of the sample x_j belonging to cluster i, and u_{ij} is assumed to satisfy the constraint $\sum_{i=1}^{C}u_{ij}=1, 1\leq j\leq N$. $\tilde{v}_i$ is the prototype of cluster i and Ψ_{com} is the mapping function of the global kernel that is a combination kernel on top of the common multi-kernel space:

$$\Psi_{com}(x)=\sum_{k=1}^{K}\sqrt{w_k^b}\psi_{com}^{(k)}(x) \tag{20}$$

where w_k is the weight of the k th view; b is an exponent and set to be 1 in this paper. Equation (20) is then written as

$$\Psi_{com}(x)=\sum_{k=1}^{K}w_k^{\frac{1}{2}}\psi_{com}^{(k)}(x) \tag{21}$$

Note that in (20) and (21), for simplicity, we continue to use $\psi_{com}^{(k)}(x)$ to represent its orthogonal expansion, same as defined in (12). Accordingly, the global kernel can be represented as below:

$$\begin{aligned}\kappa_{com}&=\langle\Psi_{com}(x),\Psi_{com}(x)\rangle\\&=w_1\kappa_{com}^{(1)}+w_2\kappa_{com}^{(2)}+\cdots+w_K\kappa_{com}^{(k)}\end{aligned} \tag{22}$$

where $\kappa_{com}^{(k)}$ is the common kernel for the kth view as (15). The update equations of prototypes $\tilde{v}_i$ can be derived by Lagrangian optimization as below:

$$\tilde{v}_i=\frac{\sum_{j=1}^{N}u_{ij}^m\Psi_{com}(x_j)}{\sum_{j=1}^{N}u_{ij}^m},\quad i=1,2,\ldots,C \tag{23}$$

and then the global clustering function is rewritten as:

$$\begin{aligned}J_{global}(U,V,W)&=\sum_{i=1}^{C}\sum_{j=1}^{N}u_{ij}^m\left(\begin{array}{l}\Psi_{com}(x_j)^T\Psi_{com}(x_j)\\\qquad -2\Psi_{com}(x_j)^T\tilde{v}_i+\tilde{v}_i^T\tilde{v}_i\end{array}\right)\\&=\sum_{i=1}^{C}\sum_{j=1}^{N}u_{ij}^m\left(\begin{array}{l}\kappa_{com}(x_j,x_j)-2\sum_{j_1=1}^{N}\hat{u}_{ij_1}\kappa_{com}(x_j,x_{j_1})\\\qquad +\sum_{j_1=1}^{N}\sum_{j_2=1}^{N}\hat{u}_{ij_1}\hat{u}_{ij_2}\kappa_{com}(x_{j_1},x_{j_2})\end{array}\right)\end{aligned} \tag{24}$$

where $\hat{u}_{ij}=u_{ij}^m\Big/\sum_{j=1}^{N}u_{ij}^m$ is the normalized membership. By substituting (22) into (24), the objective function of the global clustering can be further rewritten as

$$\begin{aligned}J_{global}(U,V,W)&=\sum_{k=1}^{K}w_k\sum_{i=1}^{C}\sum_{j=1}^{N}u_{ij}^m\left(\begin{array}{l}\kappa_{com}^{(k)}(x_j,x_j)-2\sum_{j_1=1}^{N}\hat{u}_{ij_1}\kappa_{com}^{(k)}(x_j,x_{j_1})\\\qquad +\sum_{j_1=1}^{N}\sum_{j_2=1}^{N}\hat{u}_{ij_1}\hat{u}_{ij_2}\kappa_{com}^{(k)}(x_{j_1},x_{j_2})\end{array}\right)\\&=\sum_{k=1}^{K}w_k\sum_{i=1}^{C}\sum_{j=1}^{N}u_{ij}^m\,\|\psi_{com}^{(k)}(x_j)-v_i^{(k)}\|^2\end{aligned} \tag{25}$$

where $v_i^{(k)}=\sum_{j=1}^{N}u_{ij}^m\psi_{com}^{(k)}(x_j)\Big/\sum_{j=1}^{n}u_{ij}^m$ is the prototype of cluster i in the kth view.

The global clustering u_{ij} can be directly derived from the multi-kernel fuzzy clustering. This derived global clustering

provides important guidance for local partition to avoid being trapped into local optimization. However, we cannot ensure it is the optimal partition. According to the common fusion strategy in collaborative fuzzy clustering, the global clustering is considered as the weighted summation of the local partitions from each individual view [25] as (26)

$$u_{ij}^{m} = \sum_{k=1}^{K} w_k u_{ij,k}^{m} \tag{26}$$

We can estimate the global clustering by using (26) which is obtained by collaborative learning of local partitions.

3) Fusion of local partition and global clustering

With the definitions of local partition (18) and global clustering (25), our objective clustering function as defined in (16) can be then expressed as below:

$$\begin{aligned} J\ (U,V,W) &= J_{local}(U,V,W) + \alpha \cdot J_{global}(U,V,W) \\ &= \sum_{k=1}^{K} w_k \left[\sum_{i=1}^{C}\sum_{j=1}^{N} \tilde{u}_{ij,k,\eta} \| \psi_{com}^{(k)}(x_j) - \overline{v}_i^{(k)} \|^2 \right] \\ &\quad + \sum_{k=1}^{K} w_k \sum_{i=1}^{C}\sum_{j=1}^{N} u_{ij}^{m} \| \psi_{com}^{(k)}(x_j) - \overline{v}_i^{(k)} \|^2 + \lambda \sum_{k=1}^{K} w_k \ln w_k \end{aligned} \tag{27}$$

where $\tilde{u}_{ij,k,\eta} = (1-\eta)u_{ij,k}^{m} + \frac{\eta}{K-1} \sum_{k'=1,k'\neq k}^{K} u_{ij,k'}^{m}$ and $\psi_{com}^{(k)}(x)$ is defined as (14). It should be noted that $\overline{v}_i^{(k)}$ represents the prototype of cluster i in the kth view based on the hypothesis that the global clustering and each view partition are able to achieve consistency for an optimal clustering.

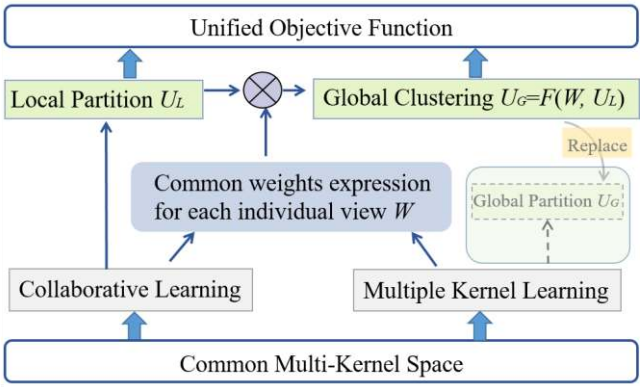


Fig. 3. Schematic derivation of the unified objective function.

In our formulation, we estimate the global clustering by substituting (26) into (27) and with further mathematical simplification, our method CoMK-FC aims at minimizing the objective function which is defined as

$$\begin{aligned} J(U,V,W) &= \sum_{k=1}^{K} w_k \sum_{i=1}^{C}\sum_{j=1}^{N} (\tilde{u}_{ij,k,\eta} + \sum_{k=1}^{K} w_k u_{ij,k}^{m}) \| \psi_{com}^{(k)}(x_j) - \overline{v}_i^{(k)} \|^2 \\ &\quad + \lambda \sum_{k=1}^{K} w_k \ln w_k \\ &= \sum_{k=1}^{K} w_k \sum_{i=1}^{C}\sum_{j=1}^{N} \tilde{u}_{ij,k} \| \psi_{com}^{(k)}(x_j) - \overline{v}_i^{(k)} \|^2 + \lambda \sum_{k=1}^{K} w_k \ln w_k \\ s.t.\ u_{ij,k} &\in [0,1], w_k \in [0,1], \sum_{k=1}^{K} w_k = 1, \sum_{i=1}^{C} u_{ij,k} = 1, 1 \le j \le N, 1 \le k \le K \end{aligned} \tag{28}$$

where $\tilde{u}_{ij,k} = \tilde{u}_{ij,k,\eta} + \sum_{k=1}^{K} w_k u_{ij,k}^{m}$.

In our model as illustrated in Fig. 3, based on a common kernel space, collaborative learning is performed to extract local partition from each individual view, while a higher level of multi-kernel learning is conducted to directly derive global clustering. We unify the local partitions and global clustering as a uniform model under our hypothesis that the clustering achieves optimal solution when the joint local partitions conform to the global clustering.

C. Clustering Optimization

By minimizing the objective function in (28), our algorithm can simultaneously search the optimal weight vector $W = [w_1, \cdots, w_K]$, partition matrix $U = [U^{(1)}, \cdots, U^{(K)}]$, and cluster prototypes $V = [V^{(1)}, \cdots, V^{(K)}]$ for each view.

1) Partition Matrix Optimization

We define the distance function between data x_j and cluster prototype $\overline{v}_i^{(k)}$ in the k*th* view as $d_{ij,k}^2 = (\psi_{com}^{(k)}(x_j) - \overline{v}_i^{(k)})^T (\psi_{com}^{(k)}(x_j) - \overline{v}_i^{(k)})$. When $V^{(k)}$ and w_k are fixed, the distances are constants. With constraint $\sum_{i=1}^{C} u_{ij,k} = 1$ and by Lagrangian optimization, the minimization of $J(U,V,W)$ in (28) is equivalent to the optimization of $u_{ij,k}$ as below:

$$\begin{aligned} &J(u_{ij,k}, \alpha_{j,k}) = \\ &\sum_{k=1}^{K} w_k \sum_{i=1}^{C}\sum_{j=1}^{N} ((1-\eta+w_k)u_{ij,k}^{m} + \sum_{k'=1,k'\neq k}^{K} (\frac{\eta}{K-1} + w_{k'})u_{ij,k}^{m}) d_{ij,k}^2 \\ &\quad - \sum_{k=1}^{K}\sum_{j=1}^{N} \alpha_{j,k} (\sum_{i=1}^{C} u_{ij,k} - 1) + \lambda \sum_{k=1}^{K} w_k \ln w_k \end{aligned} \tag{29}$$

From (29), we can obtain the optimal value of $u_{ij,k}$ by setting $\partial J\ (u_{ij,k}, \alpha_{j,k}) / \partial u_{ij,k} = 0$ and $\partial J\ (u_{ij,k}, \alpha_{j,k}) / \partial \alpha_{j,k} = 0$. Thus, we have the following equations

$$\begin{aligned} &\frac{\partial J\ (u_{ij,k}, \alpha_{j,k})}{\partial u_{ij,k}} = m w_k (1-\eta+w_k) u_{ij,k}^{m-1} d_{ij,k}^2 \\ &\quad + \sum_{k'=1,k'\neq k}^{K} m w_{k'} (\frac{\eta}{K-1} + w_k) u_{ij,k}^{m-1} d_{ij,k'}^2 - \alpha_{j,k} = 0 \\ &\Rightarrow \frac{\partial J(u_{ij,k}, \alpha_{j,k})}{\partial \alpha_{j,k}} = \sum_{i=1}^{C} u_{ij,k} - 1 = 0 \end{aligned} \tag{30}$$

From (30), the solution for $u_{ij,k}$ is:

$$u_{ij,k} = \left(\frac{\alpha_{j,k}}{m w_k (1-\eta+w_k) d_{ij,k}^2 + \sum_{k'=1,k'\neq k}^{K} m w_{k'} (\frac{\eta}{K-1} + w_k) d_{ij,k'}^2} \right)^{\frac{1}{m-1}} \tag{31}$$

and with the constraint $\sum_{i=1}^{C} u_{ij,k} = 1$, we can eliminate $\alpha_{j,k}$ and further obtain the closed-form solution of the alternative primal memberships for *k*th view as:

$$u_{ij,k}=\frac{1}{\sum_{h=1}^{C}\left[\frac{w_k(1-\eta+w_k)d_{ij,k}^2+\sum_{k'=1,k'\neq k}^{K}w_{k'}(\frac{\eta}{K-1}+w_k)d_{ij,k'}^2}{w_k(1-\eta+w_k)d_{hj,k}^2+\sum_{k'=1,k'\neq k}^{K}w_{k'}(\frac{\eta}{K-1}+w_k)d_{hj,k'}^2}\right]^{\frac{1}{m-1}}} \quad (32)$$

2) Weight Optimization

For the kth view, suppose that the memberships $U^{(k)}=[u_{ij,k}]_{C\times N}$ are fixed, we seek to derive the optimal prototypes and weights. By setting $\partial J\ (U,V,W)/\partial v_i^{(k)}=0$, we have

$$\frac{\partial J(U,V,W)}{\partial v_i^{(k)}}=-2\sum_{j=1}^{N}\tilde{u}_{ij,k}(\psi^{(k)}(x_j)-\overline{v}_i^{(k)})=0 \quad (33)$$

Thus, when $U^{(k)}$ are fixed, (33) reaches the local minimum if and only if $\overline{v}_i^{(k)}$ meets the following condition:

$$\overline{v}_i^{(k)}=\sum_{j=1}^{N}\tilde{u}_{ij,k}\psi^{(k)}(x_j)\Big/\sum_{j=1}^{N}\tilde{u}_{ij,k},\quad i=1,2,...,C \quad (34)$$

where $\tilde{u}_{ij,k}=(1-\eta+w_k)u_{ij,k}^m+\sum_{k'=1,k'\neq k}^{K}(\frac{\eta}{K-1}+w_{k'})u_{ij,k'}^m$.

From (34), it can be seen that these cluster prototypes are in the implicit kernel-induced feature space, and the prototypes cannot be directly computed. In order to solve this problem, we substitute $\overline{v}_i^{(k)}$ in our distance function with its solution as (34), then we can derive the explicit expression as below:

$$\begin{aligned}d_{ij,k}^2&=\|\psi_{com}^{(k)}(x_j)-\overline{v}_i^{(k)}\|^2\\&=\psi_{com}^{(k)}(x_j)^T\psi_{com}^{(k)}(x_j)-2\psi_{com}^{(k)}(x_j)^T\overline{v}_i^{(k)}+\overline{v}_i^{(k)T}\overline{v}_i^{(k)}\\&=\kappa_{com}^{(k)}(x_j,x_j)-2\sum_{j_1=1}^{N}\hat{u}_{ij_1,k}\kappa_{com}^{(k)}(x_j,x_{j_1})+\sum_{j_1=1}^{N}\sum_{j_2=1}^{N}\hat{u}_{ij_1,k}\hat{u}_{ij_2,k}\kappa_{com}^{(k)}(x_{j_1},x_{j_2})\end{aligned} \quad (35)$$

where $\kappa_{com}^{(k)}=(\omega_1^{(k)})^{b_k}\kappa_1^{(k)}+(\omega_2^{(k)})^{b_k}\kappa_2^{(k)}+\cdots+(\omega_{M_k}^{(k)})^{b_k}\kappa_{M_k}^{(k)}$, $\hat{u}_{ij,k}=\tilde{u}_{ij,k}\Big/\sum_{j=1}^{N}\tilde{u}_{ij,k}$. Thus, when the memberships are fixed, the distance can be obtained without implicitly computing cluster prototypes.

For the kth view, suppose that $U^{(k)}$ and $d_{ij,k}^2$ are fixed and by the Lagrangian optimization, the minimum of $J(U,V,W)$ in (28) can be solved by finding the following optimization problem about w_k:

$$\begin{aligned}J(w_k,\beta)=&\sum_{k=1}^{K}w_k\sum_{i=1}^{C}\sum_{j=1}^{N}\tilde{u}_{ij,k,\eta}d_{ij,k}^2+\sum_{k=1}^{K}w_k\sum_{i=1}^{C}\sum_{j=1}^{N}u_{ij}^m d_{ij,k}^2\\&+\lambda\sum_{k=1}^{K}w_k\ln w_k-\beta(\sum_{i=1}^{K}w_k-1)\end{aligned} \quad (36)$$

where $\tilde{u}_{ij,k,\eta}=(1-\eta)u_{ij,k}^m+\frac{\eta}{K-1}\sum_{k'=1,k'\neq k}^{K}u_{ij,k'}^m$, $u_{ij}^m=\sum_{k=1}^{K}w_k u_{ij,k}^m$.

From (36), we can obtain the optimal value of w_k by setting $\partial J(w_k,\beta)/\partial w_k=0$ and $\partial J(w_k,\beta)/\partial\beta=0$. Thus, we have the following equations:

$$\frac{\partial J\ (w_k,\beta)}{\partial w_k}=\sum_{i=1}^{C}\sum_{j=1}^{N}(\tilde{u}_{ij,k,\eta}+u_{ij}^m)d_{ij,k}^2+\lambda(\ln w_k+1)-\beta=0 \quad (37)$$

$$\frac{\partial J\ (w_k,\beta)}{\partial\beta}=\sum_{k=1}^{K}w_k-1=0 \quad (38)$$

From (37), w_k is obtained as below:

$$w_k=\exp(\frac{\beta}{\lambda})\cdot\exp(-\frac{\sum_{i=1}^{C}\sum_{j=1}^{N}(\tilde{u}_{ij,k,\eta}+u_{ij}^m)d_{ij,k}^2+\lambda}{\lambda}) \quad (39)$$

With the constraint (38), we have

$$\exp(\frac{\beta}{\lambda})=1\Big/\sum_{h=1}^{K}\exp(-\frac{\sum_{i=1}^{C}\sum_{j=1}^{N}(\tilde{u}_{ij,h,\eta}+u_{ij}^m)d_{ij,h}^2+\lambda}{\lambda}) \quad (40)$$

Thus, we can eliminate β and further obtain the closed-form solution of the alternative optimal weight for the kth view as

$$w_k=\frac{\exp(-\sum_{i=1}^{C}\sum_{j=1}^{N}(\tilde{u}_{ij,k,\eta}+u_{ij}^m)d_{ij,k}^2\Big/\lambda)}{\sum_{h=1}^{K}\exp(-\sum_{i=1}^{C}\sum_{j=1}^{N}(\tilde{u}_{ij,h,\eta}+u_{ij}^m)d_{ij,h}^2\Big/\lambda)} \quad (41)$$

D. Algorithm and Complexity and Convergence Analysis

1) Algorithm

The proposed model is summarized as Algorithm 1.

2) Analysis of the computational complexity and convergence

There are two main parts which may affect the computational complexity of CoMK-FC, given the input N samples, C clusters, K views, M kernel matrices for each view and T_1 training iterations of MKFC, T_2 training iterations of CoMK-FC. The first part is the initialization which is to construct the combined-kernel expression for each view by MKFC. Since the computational complexity of MKFC is $o(N^2CM)$ per iteration, the time complexity of this part is $o(N^2CMKT_1)$. Note that this doesn't include the construction of the kernel matrices. The second part is for the alternating optimization. The computational complexity of this part is $o(KT_2+NCKT_2+CKT_2)$. Thus, the overall cost for our method is $o(N^2CMKT_1+KT_2+NCKT_2+CKT_2)$. It is noted that the view number K and clustering number C are far smaller than samples number N, the computational complexity of CoMK-FC can be rewritten as $o(\overline{K}N^2)$ where $\overline{K}=CMKT_1$ is a constant. And it shows that the computational cost focuses on the first part. In practical applications, we will try to use a simple combination of kernel functions to express each view according to the actual conditions.

The CoMK-FC algorithm is initialized with the partition matrix and weight for each view, followed by iteratively updating the distance matrices with fixed partition matrices and view weights until the change of objective function per iteration falls below a given threshold. In this way, the objective function

(28) is minimized interactively.

For the k th view, it is supposed that partial minimization is achieved in the τ th iteration and according to (32), we have

$$J(U^{(k,\tau+1)},V^{(k,\tau)},W^{(k,\tau)}) \le J(U^{(k,\tau)},V^{(k,\tau)},W^{(k,\tau)}) \quad (42)$$

Similarly, according to (34) and (41), we can obtain

$$J\ (U^{(k,\tau+1)},V^{(k,\tau+1)},W^{(k,\tau)}) \le J(U^{(k,\tau+1)},V^{(k,\tau)},W^{(k,\tau)}) \quad (43)$$

$$J(U^{(k,\tau+1)},V^{(k,\tau+1)},W^{(k,\tau+1)}) \le J(U^{(k,\tau+1)},V^{(k,\tau+1)},W^{(k,\tau)}) \quad (44)$$

and the following relationship is obtained as

$$J(U^{(k,\tau+1)},V^{(k,\tau+1)},W^{(k,\tau+1)}) \le J(U^{(k,\tau)},V^{(k,\tau)},W^{(k,\tau)}) \quad (45)$$

It is shown that $J(U,V,W)$ is a decreasing function in each iterative computing. Therefore, our algorithm can subsequently converge to a local optimal solution corresponding to different initializations.

Algorithm 1 Collaborative multi-kernel fuzzy clustering (CoMK-FC)

Input: Given a set of N data points $X=\{x_i\}_{i=1}^{N}$ in K views, a basis set of kernel functions $\{\kappa_k\}_{k=1}^{M}$, the number of kernels M_k and the exponent b_k for each view, the number of clusters C, the fuzzy index m, the parameter η and λ, the termination criterion ζ and T, and initialization of each view partition matrix $U^{(k)}=\{u_{ij,k}\}_{i,j=1}^{C,N}$ and weights $\{w_k\}_{k=1}^{K}$.

Output: The overall partition matrix $U=\{u_{ij}\}_{i,j=1}^{C,N}$ and the weights $\{w_k\}_{k=1}^{K}$ for each view.

1. Perform MKFC on each view to get the partition matrix $U^{(k)}$ and combined-kernel $\kappa_{com}^{(k)}$.

2. **Procedure** CoMK-FC (Data *X*, Number *C*, combined-kernel $\{\kappa_{com}^{(k)}\}_{k=1}^{K}$)

3. The partition matrix $U^{(k)}=\{u_{ij,k}\}_{i,j=1}^{C,N}$ from MKFC as initial membership matrix

4. **Repeat**

5. Calculate

$$\tilde{u}_{ij,k,\eta}=(1-\eta)u_{ij,k}^{m}+\frac{\eta}{K-1}\sum_{k'=1,k'\neq k}^{K}u_{ij,k'}^{m},\ u_{ij}^{m}=\sum_{k=1}^{K}w_k u_{ij,k}^{m},\ \tilde{u}_{ij,k}=\tilde{u}_{ij,k,\eta}+u_{ij}^{m},$$

$$\hat{u}_{ij,k}=\tilde{u}_{ij,k}\Big/\sum_{j=1}^{N}\tilde{u}_{ij,k}$$

6. Calculate distances $D^{(k)}=\{d_{ij,k}^{2}\}_{i,j,k=1}^{C,N,K}$ by (35)

7. Update weights $\{w_k\}_{k=1}^{K}$ by (41)

8. Update partition matrix $U^{(k)}=\{u_{ij,k}\}_{i,j=1}^{C,N}$ ($k=1,\cdots,K$) by (32)

9. **Until** $|J^{(t)}-J^{(t-1)}|\le\zeta$ or the number of iterations $t>T$

10. **Return** $U^{(k)}=\{u_{ij,k}\}_{i,j=1}^{C,N}$ ($k=1,\cdots,K$) and $\{w_k\}_{k=1}^{K}$

11. Calculate the overall partition matrix $U=\{u_{ij}\}_{i,j=1}^{C,N}$ by $u_{ij}^{m}=\sum_{k=1}^{K}w_k u_{ij,k}^{m}$.

12. **End procedure**

IV. Experimental Results and Discussion

The performance of our proposed model was evaluated by comparing with 6 algorithms including: 1) Co-FKM [24]; two multitask clustering algorithms: 2) CombKM [35] and 3) Co-clustering [36]; and three weighting multiview clustering methods: 4) the multiple kernel fuzzy clustering algorithm (refer to as MKFCM) [28-29], 5) two-level variable weighting multiview clustering algorithm (TW-k-means) [12] and 6) weighted view collaborative fuzzy c-means algorithm (WV-Co-FCM) [25]. The methods were evaluated over six data resources including two synthetic datasets, Brodatz texture images [37], two public datasets from UCI machine learning repository [38] and Corel3400 image databases [39].

A. Evaluation Measurements

Clustering results were evaluated by two performance indices which are the normalized mutual information (NMI) [40] and the rand index (RI) [41]. NMI is defined as

$$NMI=\sum_{i=1}^{C}\sum_{j=1}^{C}n_{i,j}\log(N\cdot n_{i,j}/n_i\cdot n_j)\Big/\sqrt{(\sum_{i=1}^{C}n_i\log\frac{n_i}{N})(\sum_{j=1}^{C}n_j\log\frac{n_j}{N})} \quad (46)$$

where N denotes the total number of samples in dataset, n_i is the number of data points belonging to class i, n_j is the number of data points belonging to cluster j and $n_{i,j}$ is the number of samples belonging to class i and cluster j.

Assuming that R and Q are two partitions of a dataset by two different clustering algorithms, RI is defined as

$$RI=\frac{a+d}{a+b+c+d} \quad (47)$$

where a denotes the number of any two samples belonging to the same class in R and to the same cluster in Q, b denotes the number of any two samples belonging to the same class in R and to different clusters in Q, c denotes the number of any two samples belonging to different classes in R and to the same cluster in Q, d denotes the number of any two samples belonging to different classes in R and to different clusters in Q.

The range of NMI and RI values are from 0 to 1. And a value of 1 denotes that the clustering results match the given category labels perfectly. To evaluate the performances of fuzzy clustering algorithms with NMI and RI, we converted the fuzzy membership degrees to hard assignments by assigning each data to the cluster with the highest membership degree.

B. Parameter settings

To evaluate the clustering performance of the methods for comparison, we constructed basis kernel mappings with different kernel functions. As shown in Table I, we selected a set of commonly used kernels in our experiments.

As shown in Table II, for all the experiments, the grid search strategy combined with NMI and RI was used to obtain the optimal parameters for each algorithm. And the results of each algorithm in this paper were obtained based on these optimal parameters as described in Table II. Since the performance of these clustering methods depends on the initial values, the experimental results are shown in terms of the mean and standard deviations of NMI and RI for 20 runs of each algorithm with different initializations and corresponding optimal parameters.

C. Experimental results on Synthetic Datasets

1) Synthetic Dataset 1

In this experiment, a three-dimensional synthetic dataset was used for validation and comparison of our algorithm and other four clustering methods for multiview data. This synthetic dataset has two encircling ring-shapes including 1024 sample points in the black ring, and 2025 samples in the blue ring (as shown in Fig. 4(a)). In our experiment, we constructed multiview data by projecting the 3-D synthetic dataset onto x-y and y-z subspaces. Fig. 4(b)-(c) shows the two-view data, i.e., view-1 and view-2, respectively. Based on these two-view data, we aim to cluster these samples into two groups, the blue group and the black group.

TABLE I.
BASIS KERNEL FUNCTIONS

Kernel Type	Kernel Functions $\kappa(x_1, x_2)$	Parameters Settings	Kernels Number
Linear	$x_1^T x_2$; $1+x_1^T x_2$	-	2
Polynomial	$(x_1^T x_2 + a)^b$	b =2,3,4, a : [0.5:0.5:2.5]	15
Tangent Hyperbolic	$\tanh(c1 + c2 * (x_i^T x_j))$	$c1$: [0.1:0.1:0.5], $c2$ =1.0; $c1$ =0.5, $c2$: [0.2:0.2:1].	10
Histogram Intersection	$\min\{x_i, x_j\}$	-	1
Gaussian	$\exp\left(-\frac{\| x_i - x_j \|^2}{2\sigma^2}\right)$	$\sigma = \min_{i,j}(\frac{-\| x_i - x_j \|^2}{\log(\lambda)})$ λ =0.005	Unlimited
Hermite Orthogonal Polynomial[42]	$\prod_{h=1}^{d}\sum_{k=0}^{n} He_k(x_{i,h}) He_k(x_{j,h})$	$He_0(x)=1, He_1(x)=x$ $He_{k+1}(x) = xHe_k(x) - nHe_{k-1}(x)$	Unlimited

d is the dimension of feature vector, n is a member of natural number set.

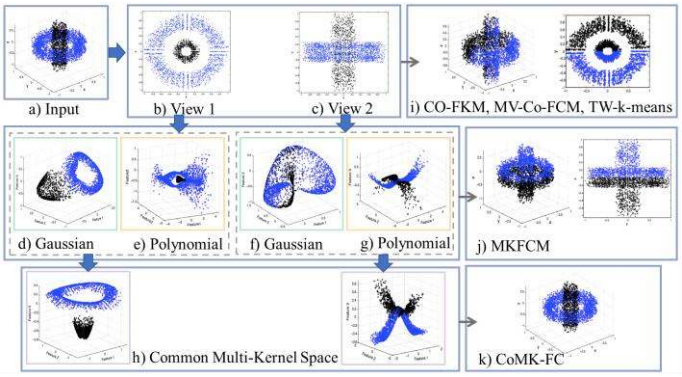


Fig. 4. Analysis on the synthetic dataset 1. (d),(e),(f),(g) are kernel types.

The visualizations of data patterns in kernel spaces by method in [43] are given in Fig. 4(d)-(h). Fig. 4(d)-(g) present the visualizations of view-1 and view-2 features in Gaussian and polynomial kernel spaces respectively. From Fig. 4(d)-(g), we can see that the separation of view-1 in Gaussian kernel spaces is better than that in polynomial kernel spaces, while the separation of view-2 in polynomial kernel spaces is better than that in Gaussian kernel spaces. Fig. 4(h) is the visualization of view-1 and view-2 features in common multi-kernel space with $\kappa_{com} = w_1\kappa_1 + w_2\kappa_2$ where κ_1 is a Gaussian kernel, κ_2 is a polynomial kernel and $w_1 = 0.7, w_2 = 0.3$ for view-1, $w_1 = 0.3, w_2 = 0.7$ for view-2. As demonstrated in Fig. 4(h), the separation of view-1 and view-2 in common kernel space is better than that in a single kernel space.

As displayed in Fig. 4(i)-(j), the clustering results of the comparison methods, Co-FKM, MV-Co-FCM and TW-k-means, did not converge to the expected clustering groups. The major reason is that these methods used the Euclidean distance as the distance measure function which is not suitable to cluster the synthetic non-spherical data. MKFCM was introduced aiming to improve the clustering of non-spherical data. However, this method tended to divide the dataset equally as same as the conventional FCM algorithm and resulted in the clustering result as shown in Fig. 4(j). On the contrary, based on the common multi-kernel space, our method clustered the dataset into the expected groups as shown in Fig. 4(k).

2) Synthetic Dataset 2

The two-rainbow shaped synthetic dataset is a typical dataset in the studies of non-convex clustering analysis. As shown in Fig. 5(a), this synthetic dataset has two intertwining rainbows including 763 sample points in the black rainbow, 585 samples in the blue rainbow and 80 noise samples denoted by red spots. And we aim to cluster these samples into two groups. In our experiments, based on three scale parameters σ in the range of [0.05, 0.45] with the interval of 0.15, we calculated spectral features with Gaussian similarity measurement of the multiview data (shown in Fig. 5(d)-(f)). Two-dimensional spectral features were selected for each view where the eigenvectors correspond to the first two minimal eigenvalues.

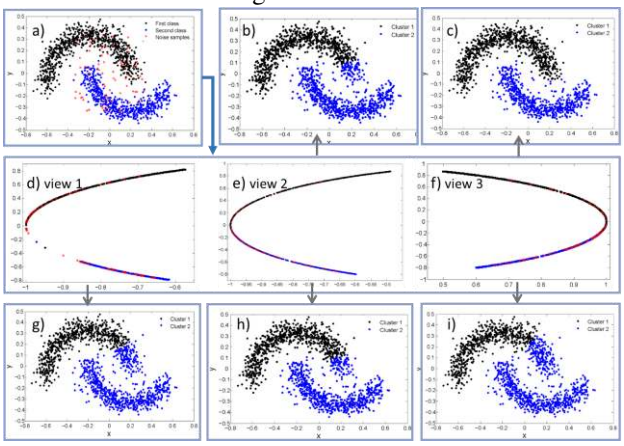


Fig. 5. Analysis on synthetic dataset 2. (a) Original data, (b)-(c) The clustering results of WV-Co-KFCM and CoMK-FC, respectively, (d)-(f) Spectral features of three views corresponding to scale parameters 0.05, 0.30 and 0.45, respectively, (g)-(i) The clustering results of Co-FCM, MKFCM and WV-Co-FCM, respectively. The results of cluster 1 and 2 are plotted in black and blue colors; noise is in red.

In our experiments, one kernel mapping was used for each view due to the simplicity of the two-dimensional features in our multiview data. In order to validate the contribution of the global clustering guidance in our method, we omit the term corresponding to the global clustering guidance in our objective function, i.e. set $\alpha = 0$ in (16), and namely WV-Co-KFCM.

As shown in Fig. 5(b), WV-Co-KFCM failed to obtain a satisfactory result without the global clustering guidance. And the clustering results of Co-FCM, MKFCM, and WV-Co-FCM failed to cluster the dataset into the expected groups as shown in Fig. 5(g)-(i). Our method outperformed the other methods in

comparison as demonstrated by the result of CoMK-FC as shown in Fig. 5(c). The main reason is that the three methods in comparison tended to cluster the samples equally. However, when the numbers of samples in different classes are unbalanced, these methods might wrongly classify some samples from the larger classes into the smaller classes in order to obtain an equally divided classification.

TABLE II.
OPTIMAL PARAMETER VALUES FOR SIX ALGORITHMS ON DIFFERENT DATASETS

Algorithms	Parameter Ranges for Grid Search	Optimal Parameters from Heuristic Analysis				
		Synthetic	Brodatz	MF	IS	Corel
MKFCM	m : [1.03:0.01:3] b :{1:0.05:2, 3:1:7}	$m=2$ $b=2$	$m=1.1$ $b=2$	$m=1.3$ $b=1.25$	$m=1.03$ $b=1.25$	$m=1.08$ $b=6$
Co-clustering	m: $fix(d/2)$, λ : [100:50:1000] μ : [100:50:1000]	-	$\lambda=500$ $\mu=400$	$\lambda=650$ $\mu=500$	$\lambda=850$ $\mu=700$	$\lambda=300$ $\mu=100$
TW-k-means	λ : [10:5:120] η : [1:1:40]	-	$\lambda=95$ $\eta=35$	$\lambda=30$ $\eta=7$	$\lambda=70$ $\eta=40$	$\lambda=40$ $\eta=20$
Co-FKM	m : [1.03:0.01:3] η : [0:0.01: $(K-1)/K$]	$m=1.25$ $\eta=0.42$	$m=1.25$ $\eta=0.42$	$m=1.2$ $\eta=0.42$	$m=1.2$ $\eta=0.42$	$m=1.08$ $\eta=0.42$
WV-Co-FCM	m : [1.03:0.01:3], $x\in[-6:1:8]$ η : [0:0.01: $(K-1)/K$], λ : { e^x }	$m=2$ $\eta=0.3$ $\lambda=10$	$m=1.25$ $\eta=0.6$ $\lambda=5000$	$m=1.03$ $\eta=0.3$ $\lambda=5000$	$m=1.1$ $\eta=0.6$ $\lambda=3000$	$m=1.03$ $\eta=0.3$ $\lambda=100$
CoMK-FC	m : [1.03:0.01:3], $x\in[-6:1:8]$ η : [0:0.01: $(K-1)/K$], λ : { e^x }	$m=1.2$ $\eta=0.3$ $\lambda=10$	$m=1.1$ $\eta=0.6$ $\lambda=400$	$m=1.08$ $\eta=0.3$ $\lambda=2500$	$m=1.25$ $\eta=0.35$ $\lambda=3500$	$m=1.1$ $\eta=0.3$ $\lambda=600$

* d is the dimension of feature vector, K is the number of views.

D. *Experimental results on Brodatz Texture Images*

To validate the robustness of the CoMK-FC algorithm against the noise, impulse salt-and-pepper noise (SPN) with various density levels [0%, 10%] was added to the Brodatz texture images [37]. Our CoMK-FC was compared with other six related multiview or multi-task clustering algorithms on these noise corrupted datasets.

To extract the multiview features from the texture images, we adopted the Gabor filter [44] in the experiments. We firstly constructed the filter bank with various orientations and frequencies. Then, the corresponding features were extracted from each pixel of the texture image by the filter bank. The detailed information of each view is shown in Table III. Our method was executed with 25 Gaussian kernel functions for each view in the experiments and the values of parameters for each algorithm are listed in Table II.

TABLE III
THE COMPOSITION OF EACH VIEW OF THE BRODATZ TEXTURE IMAGE AND BRODATZ TEXTURE IMAGE WITH NOISE

View	View generation by Gabor filter	Dimension	Cluster	Size
1	10 features generated with five orientations and two scales starting from 0.4	10	7	4096
2	15 features generated with five orientations and three scales starting from 0.5	15		
3	30 features generated with six orientations and five scales starting from 0.6	30		
4	40 features generated with eight orientations and five scales starting from 0.25	40		
		95		

The experimental results over the Brodatz texture images deteriorated by SPN with different noise levels are plotted in Fig. 6 and Fig. 7 with regards to NMI and RI measurements respectively. These experimental results demonstrated that our proposed CoMK-FC algorithm was robust against noise and outperformed the other algorithms in comparison. Our method steadily ranked the first and Co-FKM ranked the second. It is interesting to see that although WV-Co-FCM is an improved version of Co-FKM, it resulted in less accurate clustering than Co-FKM. For NMI measures as shown in Fig. 6, the accuracy of CoMK-FC ranged from 0.7456 for clean dataset to 0.4013 when the image was corrupted by SPN with 10% density. In comparison, the clustering accuracy of Co-FKM was close to the accuracy of CoMK-FC when the noise levels were relatively low [0%, 2%]. However, the accuracy of Co-FKM dropped from 0.7243 to 0.1469 as the noise levels increased from 0% to 10%, and its performance was getting close to its improved version WV-Co-FCM. The overall trends of the performance for top three methods were similar in terms of RI measurement as shown in Fig. 7.

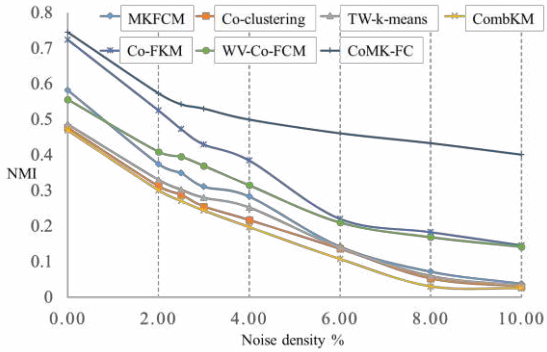


Fig. 6. NMI of each algorithm over noisy Brodatz texture images

We further analyzed the performance of MKFCM, WV-Co-FCM and CoMK-FC by comparing their weights of each view for the datasets with 0% and 2% SPN. The clustering results of CoMK-FC are given in Fig. 8(d)-(e). As shown in Fig. 8(f)-(g), view-1 was assigned the largest weight while the weights of view-3 and view-4 are close to zeros by WV-Co-FCM. For the multiview features used in our experiments, all views are supposed to make equal contributions. This is because that the characteristics of the

features by Gabor filter for each view are the same. However, as WV-Co-FCM calculated Euclidean distance for each view, the higher the feature dimensions, the larger the distance between samples and cluster prototypes, and hence the smaller the corresponding weights.

In CoMK-FC, the weights of all the views contributed equally for clustering, which complied with the same characteristics of Gabor features along the four views. When all the views have equal importance, the partition information of each view can mutually complement effectively during the clustering process. And thereby, the satisfactory clustering results were achieved.

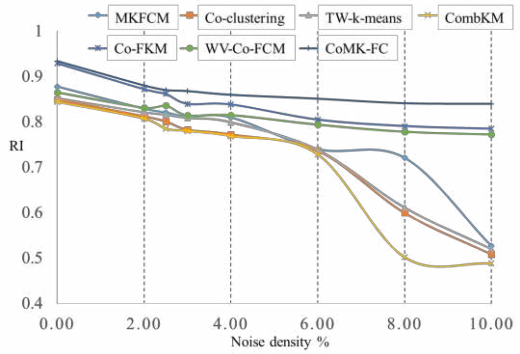


Fig. 7. RI of each algorithm over noisy Brodatz texture images

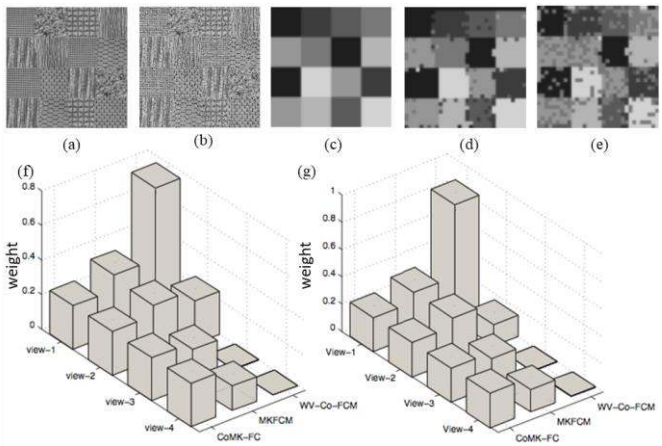


Fig. 8. Analysis on Brodatz texture images. (a) Original texture image, (b) Original texture image with 2% SPN, (c) Ideal segmentation results, (d)-(e) The results of CoMK-FC over (a) and (b) respectively, (f)-(g) The weights of each view for COMK-Fc, MKFCM, WV-Co-FCM over (a) and (b) respectively

E. Experimental results on UCI Benchmarking Datasets

The proposed method was also evaluated over two real multiview datasets obtained from the UCI repository [38] including the multiple features (MF) dataset and image segmentation (IS) dataset. The details of the available feature vectors are listed in Table IV. The other six algorithms were also performed over the two datasets for comparison. Note that our method was executed with only one basis kernel for each view the same as MKFCM in this experiment (linear kernel for view-1 and view-4, Gaussian kernel for view-2, view-3, view-4 and view-6). The parameters settings of these methods are given in Table II.

The clustering results of NMI and RI are shown in Table V. And it can be seen that the proposed CoMK-FC algorithm achieved the best results. Since the multi-task algorithms were not able to effectively combine the clustering results from different subsets, the performance of CombKM was inferior compared to other algorithms.

MKFCM, WV-Co-FCM and CoMK-FC achieved more stable clustering performance than the others as shown in Table V. This was mainly because that these methods were able to adaptively identify the importance of each view, which in turn demonstrated that the view weighting mechanism contributed to enhancing the stability of these algorithms.

TABLE IV
DESCRIPTIONS OF MF AND IS DATASETS AND THE COMPOSITION OF EACH VIEW

Data	View	Composition of Each View	Dimension	Cluster	Size
MF	Mfeat-fou view	76 Fourier coefficients of the character shapes	76	10	2000
	Mfeat-fac view	216 profile correlations	216		
	Mfeat-kar view	64 Karhunen-Love correlations	64		
	Mfeat-pix view	240 pixel averages in 2 x 3 windows	240		
	Mfeat-zer view	47 Zernike moments	47		
	Mfeat-mor view	6 morphological variables	6		
			649		
IS	Shape view	9 features for the shape information of 7 images	9	7	2310
	RGB view	10 features for the RGB values of 7 images	10		
			19		

In addition, although our method was executed with only one kernel function (the same as MKFCM) for each view, our results were better than that of MKFCM. This indicated that collaborative learning was able to effectively combine the clustering results from each view to enhance the performance of the clustering algorithm.

F. Experimental results on Corel3400 Datasets

To further evaluate the performance of the proposed CoMK-FC algorithm, our method was compared with the other six related algorithms over the Corel3400 image database which is a subset of COREL [39]. Corel3400 contains images from 34 categories where each category contains 100 images. Although the foreground object is salient in the images, the distance and angle of the object, color, lighting, and background composition have great variations within each class, which makes it difficult to achieve good clustering results by unsupervised clustering algorithms.

A number of five-class subsets were extracted and six representative ones were adopted in this paper including Lions, Leopards & Jaguars, Buses, Museum Duck Decoys, Roses as Corel-1; Elephants, British Motor Collection, Merchant Ships, Prehistoric World, Owls as Corel-2; Rhinos and Hippos, British Motor Collection, Buses, Dinosaur Illustrations, Planes of War as Corel-3; British Motor Collection, Cruise Ships, Buses, Prehistoric World, Hawks & Falcons as Corel-4; Cruise Ships, Buses, African Antelope, Roses, wildlife as Corel-5; Prehistoric World, Hawks & Falcons, Owls, African Antelope, Roses as Corel-6. Feature vectors which represent the images in terms of

seven views, three color-related views and four texture-related views are detailed in Table VI.

TABLE V
CLUSTERING PERFORMANCE (NMI AND RI (mean±sd) OF DIFFERENT ALGORITHMS ON THE MF AND IS DATASET

Datasets / Algorithm	MF		IS	
	NMI	RI	NMI	RI
MKFCM	0.8671 (±0.0148)	0.9753 (±0.0020)	0.6198 (±0.0106)	0.8670 (±0.0050)
Co-clustering	0.7578 (±0.0230)	0.9362 (±0.0171)	0.5525 (±0.0211)	0.8393 (±0.0199)
TW-*k*-means	0.8321 (±0.0381)	0.9535 (±0.0194)	0.5903 (±0.0415)	0.8551 (±0.0226)
CombKM	0.7211 (±0.0399)	0.9126 (±0.0152)	0.5567 (±0.0288)	0.8366 (±0.0232)
Co-FKM	0.8423 (±0.0275)	0.9614 (±0.0123)	0.5927 (±0.0360)	0.8661 (±0.0243)
WV-Co-FCM	0.8879 (±0.0102)	0.9763 (±0.0071)	0.6212 (±0.0085)	0.8693 (±0.0011)
CoMK-FC	**0.9205** (±0.0085)	**0.9859** (±0.0047)	**0.6417** (±0.0098)	**0.8810** (±0.0093)

TABLE VI
DESCRIPTIONS OF COREL3400 DATASETS AND THE COMPOSITION OF EACH VIEW

Features	View	Composition of Each View	Dimension		Cluster	Size
Color Features	Color Histogram	Color Hsv Histogram64	64	338	34	3400
	Color Moment	Color Luv Moment123	9			
	Color Coherence	Color Hsv Coherence64	128			
Texture Features	Tamura Texture1	Coarseness Vector	10			
	Tamura Texture2	Directionality	8			
	Wavelet Texture	Wavelet Texture	104			
	MASAR Texture	MRSAR	15			

For the implementation, our method was executed with two linear kernel functions, two polynomial kernel functions and 10 Gaussian kernel functions for each view. The settings of parameters η, m, and λ are shown in Table II. We ran CoMK-FC to produce 20 clustering results with different initial partitions. And the best results achieved by the algorithm with respect to NMI and RI were displayed.

In order to compare the clustering performance with each other over the six subsets of Corel3400 intuitively, Fig. 9 and Fig. 10 show the stacked values of NMI and RI of each algorithm respectively. The experimental results demonstrated that CoMK-FC was superior to other algorithms. As each feature vector was normalized to the same length as 1/7, the Euclidian distance was suitable for evaluating image distances. Thus, the multiview learning clustering algorithms, i.e. WV-Co-FCM and Co-FKM, both obtained good clustering results. WV-Co-FCM achieved better results than Co-FKM due to the incorporation of collaborative learning mechanism and weighted view. Without collaborative learning mechanisms, CombKM ranked the last in terms of clustering accuracy and the results of MKFCM was slightly better than TW-k-means.

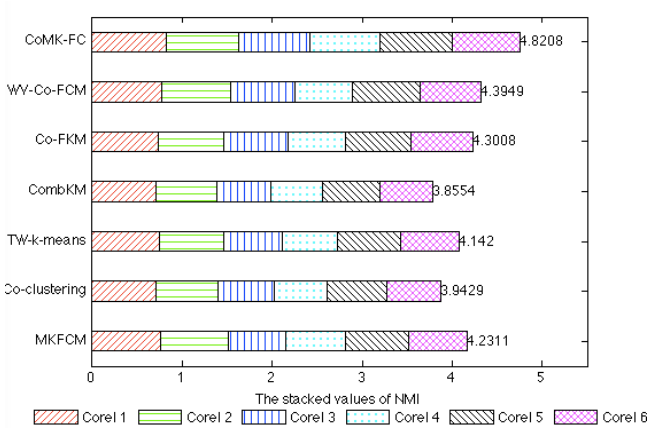


Fig. 9. The stacked values of NMI of each algorithm

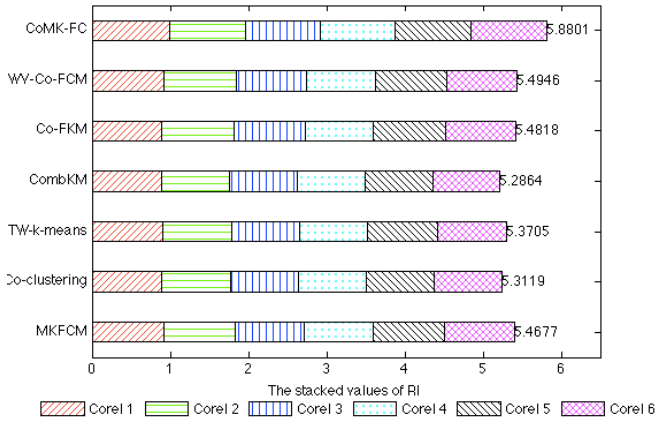


Fig. 10. The stacked values of RI of each algorithm

G. *Analysis of Parameters Setting and initialization*

There are three parameters in CoMK-FC, the fuzzification degree m and two parameters η and λ to adjust the penalty corresponding to the partition disagreement and the weight of each view. In the experiments on the Corel-1 dataset, the clustering performance was evaluated using different parameters m, η and λ.

We set the parameter η as {0.3, 0.6}, m as {1.03, 1.05, 1.08, 1.1, 1.25, 1.5, 1.8, 2.0, 2.5, 3}, and λ as {50, 200, 400, 600, 800, 1000, 1500, 2000, 3000, 5000} for the Corel-1 dataset. When fixing η as 0.3 or 0.6, for each pair of m and λ, we ran our method 20 times with different initial partitions, and the best values of NMI and RI are shown in Fig. 11. Fig. 11 shows that an appropriate value for m should be less than 1.25 and λ should be more than 400. When m was greater than 1.25, the values of NMI and RI became unstable and decreased rapidly. From Fig. 11(c) and (f), we can see that as increases, the weight of each view becomes flatter.

Further, when there was no noticeable difference between the weights of each view, as η changed, the results of CoMK-FC stayed stable. As shown in Fig. 11, although the results were hardly affected by η, overall, the results of $\eta = 0.3$ were slightly better than the results of $\eta = 0.6$. It was noted that when m was set to 2, the clustering accuracy of our method dropped

abruptly.

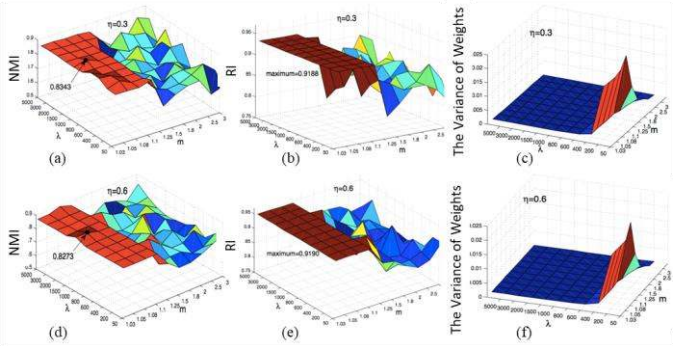


Fig. 11. Analysis of the affection on results on Corel-1 dataset. (a)–(c) The values of NMI, RI and the variance of weights against m and λ ($\eta = 0.3$); (d)-(f) The values of NMI, RI and the variance of weights when $\eta = 0.6$.

Without loss of generality, we summarize the following mechanisms to control the performance of CoMK-FC by setting different values of m, η and λ: (1) Large λ allows more views to contribute to the clustering, while small λ makes only important views contribute to the clustering; (2) When the weight of each view tends to be equal, the results would be hardly affected by η; (3) The values of m generally should be less than 1.25.

H. Statistical analysis

In order to determine the significant difference among the competing methods, we used two nonparametric methods, Wilcoxon Signed Rank Test and Quade Test from the suggested paper [45], over all the datasets for all seven algorithms in comparison. The statistical analysis results are listed in Table VII.

1) Wilcoxon Signed Rank Test

In Wilcoxon signed rank test [45], we calculated the performance differences between CoMK-FC and each of the other six algorithms on the experimental datasets. Let d_i be the performance difference between the two algorithms on the i th experiment out of N experiments, and the differences d_i were ranked according to their absolute values; the mean ranking was assigned in case of ties.

(a) Calculate the sum of ranks: Let R^+ (R^-) be the sum of ranks as the first algorithm is better (worse) than the second algorithm. The calculation formula is:

$$R^+ = \sum_{d_i>0} rank(d_i) + \frac{1}{2}\sum_{d_i=0} rank(d_i) \tag{48}$$

$$R^- = \sum_{d_i<0} rank(d_i) + \frac{1}{2}\sum_{d_i=0} rank(d_i) \tag{49}$$

where the sum of R^+ and R^- is $0.5N(N+1)$.

(b) Two-tailed test: In the condition of zero assumption, for two-tailed test, there is little diversity between R^+ and R^- meaning that the performance of two algorithms has no obvious diversity. If one of both is very small or more specifically $T = \min(R^+, R^-)$ is very small, we doubt the assumption and consider that there is obvious diversity between optimization performances of the two algorithms.

TABLE VII
STATISTICAL ANALYSIS: CoMK-FC COMPARED WITH OTHER ALGORITHMS.

Datasets / Algorithm	Wilcoxon Signed Rank Test				Quade Test		
	R^+	R^-	T	Z	C_i	$C_i - H_1$	$C_i - H_2$
MKFCM	2	208	2	-3.8453	491	296.8	297.0
Co-clustering	0	210	0	-3.9199	1401	846.8	847.0
TW-*k*-means	0	210	0	-3.9199	731	536.8	537.0
CombKM	0	210	0	-3.9199	1206	1011.8	1012.0
Co-FKM	2	208	2	-3.8453	427	232.8	233.0
WV-Co-FCM	22	188	22	-3.0986	318	123.8	124.0

(c) Calculate Z value as

$$Z = (T - \frac{1}{4}N(N+1)) \Big/ \sqrt{\frac{1}{24}N(N+1)(2N+1)} \tag{50}$$

2) Quade test

Quade test is a classical nonparametric procedure for performing multiple statistical comparisons for more than two algorithms. We set the null hypothesis H_0 indicating that there was no significant difference between the proposed CoMK-FC algorithm and the other six algorithms.

Assumed that there are k algorithms being compared with each other on N experiments, for i th experiment, the results from different algorithms will be sorted from 1 to k, represented by symbol $r_{i,j}(1 \le j \le k)$. Then we would get a rank matrix $R \in R^{N \times K}$.

(a) Calculate range within a group: Range within a group is the difference between the optimal value and the worst value from optimization results of different optimization algorithms, that is:

$$Y_{i,j} = \max_j r_{i,j} - \min_j r_{i,j} \tag{51}$$

(b) Calculate the relative value S_{ij} of each observed value within a group: Let S_j be the sum of each S_{ij} belonging to one group:

$$S_j = \sum_{i=1}^{N} S_{ij}, \quad S_{ij} = Q_i[r_{i,j} - \frac{k+1}{2}], \quad j=1,2,...,k \tag{52}$$

where Q_i is the array of range's ranks after sorting in ascending order and $(k+1)/2$ is the average rank order within a group.

(c) Compare CoMK-FC algorithm with other algorithms: According to the degrees of freedom $(k-1)$ and $(N-1)(k-1)$, and significant level α, we can obtain F boundary values by querying the F distribution table. If F_Q is greater than F boundary value, we reject H_0:

$$C_i = |S_i - S_7| > t_\alpha \left[\frac{2N(A-B)}{(N-1)(k-1)}\right]^{\frac{1}{2}} \tag{53}$$

where S_i represents the results of the i th algorithm over all experiments where our proposed CoMK-FC is the 7th algorithm, C_i is the absolute difference between CoMK-FC and ith algorithm ($i \in [1,...,k-1]$), $A = \sum_{i=1}^{N}\sum_{j=1}^{k} S_{ij}^2$ $B = \frac{1}{N}\sum_{j=1}^{k} S_j^2$.

In both Wilcoxon Signed Rank Test and Quade test, N is 20,

α is 0.05. According to literature [46], we can know the value of T should be equal to or less than 52, and Z should be smaller than -1.96 in Wilcoxon Signed Rank Test. In Quade test, since t_α is between 1.982 and 1.980 by checking "T Distribution Table", the right of Equation (53) will be between H_1 =194.1629 and H_2 =193.9670. As the results shown in Table VII, all the values of T are less than 52, all the values of Z are less than -1.96, and all the values of C_i are greater than $\max\{H_1, H_2\}$. Thus, the hypothesis H_0 is rejected and there is significant difference between the proposed CoMK-FC algorithm and the other six algorithms.

TABLE VIII
CLUSTERING PERFORMANCE (NMI AND RI) OF ICOMK-FC WITH DIFFERENT COMBINED KERNELS ON THE MF AND IS DATASET

Datasets / Combined Kernels	Parameters	MF			IS		
		NMI	RI	RunTime	NMI	RI	RunTime
1×Gaussian (Gau)	$\sigma = \log(0.005)$	0.8799 ±0.0083	0.9697 ±0.0095	133.32	0.6178 ±0.0073	0.8596 ±0.0108	78.10
1×Hermite (Her)	$n = 4$	0.8307 ±0.0075	0.9485 ±0.0077	105.33	0.5874 ±0.0024	0.8531 ±0.0086	35.32
1×Gau + 1×Her	$\sigma = \log(0.005), n = 2$	0.9185 ±0.0063	0.9807 ±0.0086	235.62	0.6315 ±0.0033	0.8798 ±0.0054	103.12
1×Gau + 2×Her	$\sigma = \log(0.005), n = 2,3$	0.9206 ±0.0078	0.9861 ±0.0053	331.24	0.6420 ±0.0028	0.8800 ±0.0062	134.28
1×Gau + 3×Her	$\sigma = \log(0.005), n = 2,3,4$	0.9202 ±0.0070	0.9857 ±0.0047	425.83	0.6412 ±0.0052	0.8791 ±0.0068	160.79
1×Gau + 4×Her	$\sigma = \log(0.005), n = 2,3,4,5$	0.9203 ±0.0037	0.9854 ±0.0062	520.92	0.6421 ±0.0027	0.8802 ±0.0038	188.93
2×Gau + 4×Her	$\sigma = \log(0.01, 0.005), n = 2,3,4,5$	0.9210 ±0.0052	0.9863 ±0.0034	644.37	0.6420 ±0.0058	0.8813 ±0.0021	261.24
3×Gau+ 4×Her	$\sigma = \log(0.1, 0.01, 0.005), n = 2,3,4,5$	0.9201 ±0.0041	0.9860 ±0.0020	760.22	0.6413 ±0.0079	0.8802 ±0.0087	338.65

I. Scalability Analysis

In this section, the scalability of CoMK-FC with respect to the number of kernels was investigated. In the experiments, CoMK-FC with different numbers of combined kernels was evaluated over the multiple features (MF) and image segmentation (IS) datasets. Because the Hermite orthogonal polynomial kernel has only one parameter chosen from natural numbers, the parameter optimization was facilitated greatly. Gaussian kernel and Hermite kernel were the basis kernels for constructing kernel bank, and CoMK-FC was tested on the basis of incrementing one more kernel at each run of our algorithm. The details on kernel combination, the clustering results of NMI, RI, and run-time are listed in Table VIII.

The experimental results demonstrated that the addition of an effective kernel could improve the performance of the algorithm. For instance, the performance of combining one Hermite kernel with one Gaussian kernel was better than the performance of solely one Hermite kernel or one Gaussian kernel. This is mainly because that Hermite orthogonal polynomial kernel has advantage in representing global information while Gaussian kernel depicts local information.

The running time (in seconds) of CoMK-FC increased linearly with the increment of the kernel number. However, the performance of CoMK-FC kept steady although the number of combined kernels increased. This is mainly because that adding an ineffective kernel affects little on clustering performance.

J. Discussion

In our proposed CoMK-FC algorithm, we introduce a common multi-kernel space to more effectively reflect the partition information of an individual view of multiview data; and we innovatively fuse the local partitions from collaborative learning with the global clustering guidance from the composite kernel space into a single objective function. The experimental results from synthetic and public datasets demonstrated that the proposed CoMK-FC algorithm outperformed the algorithms in comparison in terms of clustering accuracy.

The first finding of CoMK-FC is that the common multi-kernel space contributes to improving robustness to kernel selection. This finding has been validated by the experiments on synthetic dataset 1. As expected, for the case with non-spherical data as the two encircling ring-shapes in Fig. 4, it is not separable in Euclidean space. Our proposed CoMK-FC method maps the sample points from Euclidean space onto different kernel spaces. Since better separation could be obtained from some of the kernel spaces, e.g. Gaussian kernel for view 1 and polynomial kernel for view 2, our proposed CoMK-FC method constructs a common multi-kernel space for each view through a linear combination of multiple kernel spaces. To achieve a better partition through this common multi-kernel space, with automatic weights adjustment, our method assigns more effective kernels with higher weights. The common multi-kernel space also contributes to the scalability of our method. As shown in Table

VIII, this common space is a scalable platform in terms of the capacity for kernels. Adding an ineffective kernel affects little on clustering performance, which further justifies that CoMK-FC is immune to ineffective kernels. The immunity to ineffective kernels also enables our method to tackle the difficulties on kernel selections.

Another finding of our method is that unifying global guidance with local partitions as one single objective function couples the merits of collaborative learning and multiple-kernel learning. Mathematical derivation has proved the convergence of our object function. The contribution of the unified clustering function has been justified by the experiments on synthetic dataset 2 and UCI benchmarking MF and IS datasets. Without the global guidance, all the other clustering methods in comparison failed to cluster the dataset into the expected groups due to their tendency to cluster the samples equally. The less satisfactory results from WV-Co-KFCM, which is our model without the global term, further justify the contribution of global clustering guidance. Without the global guidance, WV-Co-KFCM tends to wrongly classify some samples points from the larger classes into the smaller classes in order to obtain an equally divided classification. In the experiments of MF and IS datasets, with attribute to the capability of adaptively identifying the importance of each view, MKFCM, WV-Co-FCM and our proposed CoMK-FC achieved more stable clustering performance than the others. By comparison with MKFCM, our CoMK-FC method achieved better results due to the effective combination of clustering results from each view through collaborative learning.

V. Conclusions and Future Work

In this paper, we proposed a collaborative multi-kernel fuzzy clustering algorithm which defined an objective function to unify the local partition and the global clustering as guidance. We firstly constructed a common multi-kernel space for multiview data to provide a more effective description of partition information from each individual view. This common multi-kernel space provided a mechanism for adaptive distance measurement in the clustering of complex multiview data. During the clustering process, our method not only considered the collaborative learning between each view, but also took into account the global clustering to guide the local partition. This global guidance in the collaborative learning contributed to the improvement of both the robustness and accuracy of our method. The experimental evaluation results on synthetic and public datasets demonstrated that our method outperformed the state-of-the-art multitask and multiview clustering algorithms in terms of clustering accuracy. In our future research, we will extend our method to multimodality biomedical image segmentation and will take the domain knowledge into consideration when choosing kernels.

References

[1] T. Kanungo, D. M. Mount, N. S. Netanyahu, C. D. Piatko, R. Silverman, and A. Y. Wu, "An efficient k-means clustering algorithm: analysis and implementation," *IEEE Trans. Pattern Analysis and Machine intelligence*, vol. 24, no. 7, pp. 881-92, 2002.

[2] J. C. Bezdek, Pattern Recognition with Fuzzy Objective Function Algorithms, New York, NY, USA: Plenum Press, 1981.

[3] A.W.C. Liew, H. Yan and N.F. Law, "Image Segmentation Based on Adaptive Cluster Prototype Estimation," *IEEE Trans. Fuzzy Systems*, vol.13, no.4, pp. 444-453, Aug. 2005.

[4] U. Von Luxburg, "A tutorial on spectral clustering. Statistics and Computing," *Hans Journal of Data Mining*," vol. 17, no. 4, pp. 395-416, 2007.

[5] W. Shitong, K. F. L. Chung, D. Zhaohong, H. Dewen, and W. Xisheng, "Robust maximum entropy clustering algorithm with its labeling for outliers," *Soft Computing*, vol. 10, no. 7, pp. 555-63, 2006.

[6] N. R. Pal, K. Pal, J. M. Keller, J. C. Bezdek, "A possibilistic fuzzy c-means clustering algorithm," *IEEE Trans. Fuzzy Systems*, vol. 13, no. 4, pp. 517-30, 2005.

[7] Z. Ji, Y. Xia, Q. Sun, G. Cao, "Interval-valued possibilistic fuzzy C-means clustering algorithm," *Fuzzy Sets and Systems*, vol. 253, pp. 138-156 2014.

[8] S. Bickel and T. Scheffer, "Multi-view clustering," in Proc. 4th IEEE Int. Conf. Data Mining, Washington, DC, USA, 2004, pp. 19-26.

[9] S. Bickel and T. Scheffer, "Estimation of mixture models using Co-EM," in Proc. Eur. Conf. Mach. Learn. (ECML), Porto, Portugal, 2005, pp. 35-46.

[10] V. R. de Sa, "Spectral clustering with two views," in Proc. 22nd Int. Conf. Mach. Learn. Workshop Learn. Multiple Views, Bonn, Germany, 2005, pp. 20-27.

[11] D. Zhou and C. J. Burges, "Spectral clustering and transductive learning with multiple views," in Proc. 24th Int. Conf. Mach. Learn., 2007, pp. 1159-1166.

[12] X. Chen, X. Xu, J. Z. Huang, and Y. Ye, "TW-k-means: Automated two-level variable weighting clustering algorithm for multiview data," *IEEE Trans. Knowl. Data Eng*., vol. 25, no. 4, pp. 932-944, Apr. 2013.

[13] X. Liu, S. Ji, W. Glänzel, and B. D. Moor, "Multiview partitioning via tensor methods," *IEEE Trans. Knowl. Data Eng*., vol. 25, no. 5, pp. 1056-1069, May 2013.

[14] 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.

[15] T. Swarnkar And P. Mitra, "Graph-based unsupervised feature selection and multiview clustering for microarray data," *Journal of Biosciences*, vol. 40, no. 4, pp. 755-67, 2015.

[16] S. Sun, "A survey of multi-view machine learning," *Neural Comput. Appl.*, vol. 23, no. 7-8, pp. 2031-2038, 2013.

[17] C. Xu, D. Tao and C. Xu, "Multi-View Learning With Incomplete Views," in *IEEE Transactions on Image Processing*, vol. 24, no. 12, pp. 5812-5825, Dec. 2015.

[18] M. Kan, S. Shan, H. Zhang, S. Lao, and X. Chen, "Multi-View Discriminant Analysis," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 38, no. 1, pp. 188-194, Jan. 2016.

[19] C. Xu, D. Tao, and C. Xu, "Multi-View Intact Space Learning," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 37, no. 12, pp. 2531-2544, Dec. 2015.

[20] S. F. Hussain and S. Bashir, "Co-clustering of multi-view datasets," *Knowledge and Information Systems*, vol. 47, no. 3, pp. 545-570, Jun. 2016.

[21] W. Pedrycz, "Collaborative fuzzy clustering," *Pattern Recognit. Lett.*, vol. 23, no. 14, pp. 1675-1686, 2002.

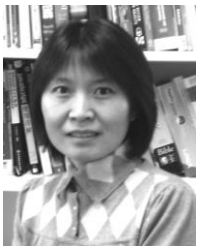
[22] S. Mitra, H. Banka, and W. Pedrycz, "Rough-fuzzy collaborative clustering systems," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 36, no. 4, pp. 795-805, Aug. 2006.

[23] L. F. S. Coletta, L. Vendramin, E. R. Hruschka, R. J. G. B. Campello, and W. Pedrycz, "Collaborative fuzzy clustering algorithms: Some refinements and design guidelines," *IEEE Trans. Fuzzy Syst.*, vol. 20, no. 3, pp. 444-462, Jun. 2012.

[24] G. Cleuziou, M. Exbrayat, L. Martin, and J.-H. Sublemontier, "CoFKM: A centralized method for multiple-view clustering," in Proc. 9th IEEE Int. Conf. Data Mining (ICDM), Miami, FL, USA, 2009, pp. 752-757.

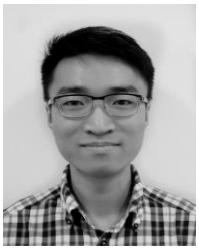
[25] Y. Jiang, F. L. Chung, and S. Wang, "Collaborative fuzzy clustering from multiple weighted views," *IEEE Trans on Cybernetics*, vol. 45, no. 4, pp. 688-701, 2015.
[26] V. R. de Sa, P. W. Gallagher, J. M. Lewis, and V. L. Malave, " Multi-view kernel construction," *Mach. Learn.*, vol. 79, pp. 47-71, 2010.
[27] D. Nguyen, L. Ngo, L. Pham, W. Pedrycz, "Towards hybrid clustering approach to data classification: Multiple kernels based interval-valued Fuzzy C-Means algorithm," *Fuzzy Sets and Systems*, vol. 279, pp. 17-39, 2015.
[28] H. C. Huang, Y. Y. Chuang, and C. S. Chen, "Multiple kernel fuzzy clustering," *IEEE Trans. Fuzzy Systems*, vol. 20, no. 1, pp. 120-134, 2012.
[29] L. Chen, C. L. P. Chen, and M. Lu, "A Multiple-Kernel Fuzzy C-Means Algorithm for Image Segmentation," *IEEE Trans. Systems, Man, and Cybernetics*, Part B (Cybernetics), vol. 41, no. 5, pp. 1263-1274, 2011.
[30] Naouel Baili and Hichem Frigui, "Fuzzy Clustering with Multiple Kernels," in 2011 IEEE International Conference on Fuzzy Systems, Taipei, Taiwan, pp.490-496, Jun. 2011.
[31] G. Tzortzis, A. Likas, "Kernel-based weighted multi-view clustering," in Proc of the 12th IEEE international conference on Data Mining, Piscataway: IEEE, pp. 675-684, 2012.
[32] D. Dovžan, and I. Skrjanc, "Recursive fuzzy c-means clustering for recursive fuzzy identification of time-varying processes," *ISA Transactions*, vol. 50, no.2, pp. 159-169, 2011.
[33] P. Angelov, D. Filev and N. Kasabov, "An Extended Version of the Gustafson-Kessel Algorithm for Evolving Data Stream Clustering," in book Evolving Intelligent Systems: Methodology and Applications, pp. 273-299, 2010.
[34] D. Dovžan, and I. Skrjanc, "Recursive clustering based on a Gustafson–Kessel algorithm," *Evolving Systems*, vol. 2, no. 1. pp. 15-24, 2011.
[35] Q. Gu and J. Zhou, "Learning the shared subspace for multitask clustering and transductive transfer classification," in Proc of the 9th IEEE Int Conf on Data Mining, Miami FL: IEEE, pp.159-168, 2009.
[36] S. Xie, H. Lu, and Y. He, "Multi-task co-clustering via nonnegative matrix factorization," in Proc. 21st International Conference on Pattern Recognition (ICPR2012), pp. 2954-2958, 2012.
[37] Trygve Randen, Brodatz Textures [EB/OL]. (2007). [Online]. Available: http://www.ux.uis.no/~tranden/brodatz.html
[38] K. Bache and M. Lichman, UCI Machine Learning Repository. Irvine, CA, USA: Univ. California, School of Information and Computer Science. (2013). [Online]. Available: http://archive.ics.uci.edu/ml
[39] http://www.cs.virginia.edu/~xj3a/research/CBIR/Download.htm
[40] A. Strehl and J. Ghosh, " Cluster ensembles — A knowledge reuse framework for combining multiple partitions," *J. Mach. Learn. Res.*, vol. 3, pp. 583-617, Mar. 2003.
[41] L. Hubert and P. Arabie, " Comparing partitions, " *Journal of classification*, vol. 2, no. 1, pp. 193-218, 1985.
[42] V. Hooshmand Moghaddam and J. Hamidzadeh, "New Hermite orthogonal polynomial kernel and combined kernels in Support Vector Machine classifier," *Pattern Recognition*, Vol. 60, PP. 921–935, 2016.
[43] L. Szymanski, B. McCane, "Visualising Kernel Spaces", Image and Vision Computing New Zealand (IVCNZ), pp. 449-452, 2011.
[44] C. Liu and H. Wechsler, "Gabor feature based classification using the enhanced fisher linear discriminant model for face recognition," *IEEE Trans. Image Processing*, vol. 11, no. 4, pp. 467-476, 2002.
[45] García S., Fernández A., Luengo J., Herrera F., "Advanced nonparametric tests for multiple comparisons in the design of experiments in computational intelligence and data mining: Experimental analysis of power", *Information Sciences*, vol. 180, pp. 2044–2064, 2010.
[46] J. Sheskin, Handbook of Parametric and Nonparametric Statistical Procedures, Chapman & Hall/CRC, 2000.

Shan Zeng received the B.E. degree in mechanical engineering in 2003 and the M.E. degree in mechatronics engineering in 2009 from Wuhan Polytechnic University. He received the Ph.D. degree in pattern recognition and intelligent systems in 2012 from Huazhong University of Science and Technology. In 2003, he joined the College of Mathematics and Computer Science, Wuhan Polytechnic University, Wuhan, China, where he is currently an Associate Professor. During 2015–2016, he was a visiting scholar at Macau University and The University of Sydney. His research interests include pattern recognition, machine learning, image processing, and hyperspectral imaging with their applications in nondestructive testing of food quality.



Xiuying Wang has a Ph.D. in computer science and currently is a Senior Lecturer, and Associate Director, Multimedia Lab, School of Information Technologies, The University of Sydney, Australia. Her research interests include biomedical data computing and visual analytics, biomedical image registration, identification, clustering and segmentation.

Hui Cui received the B.E degree from Harbin Institute of Technology, China in 2011 and Ph.D. degree in computer science from The University of Sydney, Australia in 2016. She is a postdoctoral research associate in BMIT research group, the University of Sydney. Her research interests include biomedical image processing and data analytics.



Chaojie Zheng received the B.E. degree in software engineering from the University of Sydney in 2010 and is currently pursuing a Ph.D. degree in the University of Sydney, Australia. His research interests include multi-modality data registration and fusion, image segmentation and image processing for biomedical images.

(David) Dagan Feng is currently Director (Research), Institute of Biomedical Engineering & Technology, and Academic Director, USYD-SJTU Joint Research Alliance. He has been Head, School of Information Technologies, Faculty of Engineering and Information Technologies and Associate Dean, Faculty of Science, University of Sydney. He has been the Chair Professor, Advisory Professor, Guest Professor, Adjunct Professor or Chief Scientist in different world-known universities and institutes. He is the Founder and Director of the Biomedical & Multimedia Information Technology Research Group at the University of Sydney. He has served as Chairs or Editors of different committees and key journals in the area. Professor Feng has been elected as Fellow of ACS (Australia), HKIE (Hong Kong), IET (UK), IEEE (USA), and Australian Academy of Technological Sciences and Engineering. He received his PhD in Computer Science from the University of California, Los Angeles in 1988.