One-pass Multi-view Clustering for Large-scale Data

Jiyuan Liu, Xinwang Liu*, Yuexiang Yang, Li Liu, Siqi Wang, Weixuan Liang and Jiangyong Shi
National University of Defense Technology
Changsha, Hunan, China. 410072.
{liujiyuan13, xinwangliu, yyx}@nudt.edu.cn

Abstract

Existing non-negative matrix factorization based multi-view clustering algorithms compute multiple coefficient matrices respect to different data views, and learn a common consensus concurrently. The final partition is always obtained from the consensus with classical clustering techniques, such as k-means. However, the non-negativity constraint prevents from obtaining a more discriminative embedding. Meanwhile, this two-step procedure fails to unify multi-view matrix factorization with partition generation closely, resulting in unpromising performance. Therefore, we propose an one-pass multi-view clustering algorithm by removing the non-negativity constraint and jointly optimize the aforementioned two steps. In this way, the generated partition can guide multi-view matrix factorization to produce more purposive coefficient matrix which, as a feedback, improves the quality of partition. To solve the resultant optimization problem, we design an alternate strategy which is guaranteed to be convergent theoretically. Moreover, the proposed algorithm is free of parameter and of linear complexity, making it practical in applications. In addition, the proposed algorithm is compared with recent advances in literature on benchmarks, demonstrating its effectiveness, superiority and efficiency.

1. Introduction

With the wide spread of multi-view data, multi-view clustering (MvC) algorithms are proposed to maximally integrate complementary information among views and reveal the underlying data structure for clustering [10, 18]. Most of them are developed on classical clustering methods, such as non-negative matrix factorization, k-means, spectral clustering, etc. [26, 21, 9, 17]. Therefore, MvC approaches can be roughly classified according to this criterion. In the paper, we concentrates on non-negative matrix factorization based ones.

*Corresponding author

Non-negative Matrix Factorization (NMF) [14] is one of the most fundamental clustering techniques in machine learning and data engineering tasks. It factorizes the input data into two parts, i.e. coefficient and base matrices [12]. Orthogonality [5] and low rank constraint [29] are widely explored in matrix factorization, achieving promising performance. In MvC setting, Gao et al. obtain the coefficient matrices via performing NMF on each data view, then push them towards a common consensus [6]. On the contrary, some researches assume that all views share an underlying consensus manifold, thus employ a single coefficient matrix to capture the intrinsic data structure [7]. Upon the two aforementioned frameworks, a large number of researchers [27, 30, 24, 7, 25] borrow the manifold regularization in [2] to further improve clustering performance. In specific, each view can be regarded as a manifold and manifold regularization is able to preserve the local geometry structure of data [30]. However, it requires building one or more similarity graphs, introducing higher computational and storage complexities, $\mathcal{O}(n^2)$ or even $\mathcal{O}(n^3)$ sometimes [27]. Nevertheless, Gao et al. impose orthogonality on the base matrix explicitly [7], while Zhang et al. do so implicitly [27], where the both are validated to be effective in experiment.

However, the aforementioned approaches limit the discriminative embedding learning by imposing non-negativity and fail to unify multi-view matrix factorization with partition generation closely, leading to unsatisfying performance. To address the issues, we propose an One-pass Multi-view Clustering (OPMC) algorithm. First, we remove the non-negativity constraint on both coefficient and base matrices. Instead of explicitly combining the objectives of matrix factorization and k-means in a unified formulation, we approximate the coefficient matrix with a consensus hard partition matrix and a view-specific centroid matrix, where no additional parameter is introduced. The overview of OPMC is presented in Fig. 1. It can be observed that the generated hard partition guides multi-view matrix factorization to produce more purposive coefficient matrix which, as a feedback, improves the quality of partition. In order to validate effectiveness of the proposal, we

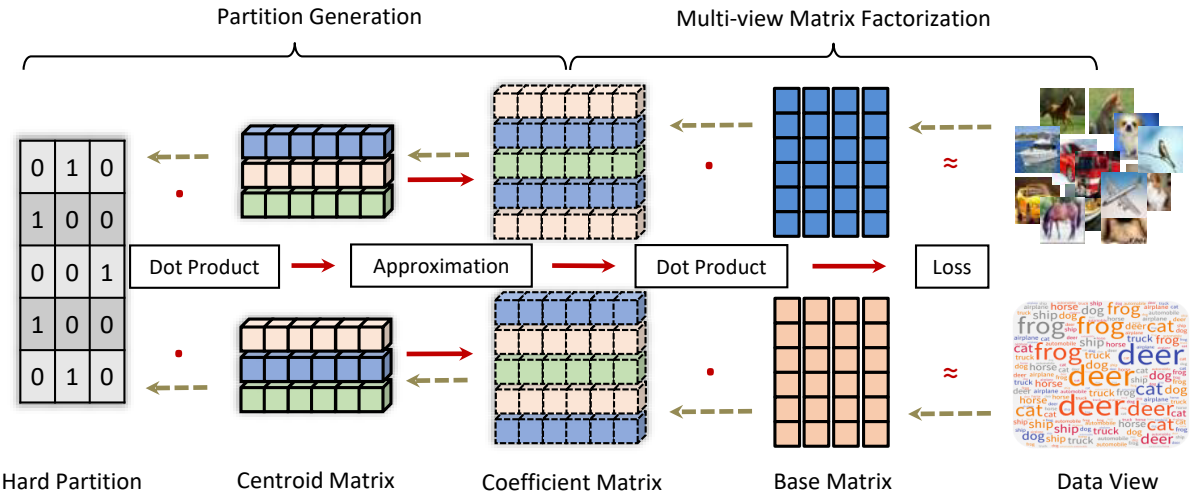

Figure 1. Overview of the proposed OPMC algorithm (*Taking the data of two views as an example*). Two semantic parts are concerned, including multi-view matrix factorization and partition generation. From left to right, the hard partition matrix passes through two view-specific transformations by multiply a centroid matrix respectively. Then, a coefficient matrix is obtained corresponding to each view. Note that, we use dotted line to represent coefficient matrix, for it does not explicitly given in our algorithm. By multiplying the coefficient matrices with base matrices, the data views can be reconstructed. From right to left, dotted arrows indicate that the clustering information flows from original data views to the consensus hard partition step by step.

design an ablation study by comparing single-view OPMC with ONMF [5]. Besides, extensive experiments are conducted and OPMC establishes state-of-the-art performance compared with recent advances on six benchmarks. Finally, the contributions are summarized as follows:

1) We find removing the non-negativity constraint and unifying matrix factorization with partition generation can improve the clustering performance, and validate their effectiveness with an ablation study.

2) We propose a non-parametric OPMC algorithm to address the multi-view data clustering problem. It achieves state-of-the-art performance on six benchmarks.

3) We design an alternate strategy to solve the resultant optimization problem. Its convergence and computational complexity ($\mathcal{O}(n)$) are analyzed theoretically and experimentally.

2. Related work

2.1. Single-view matrix factorization

Given n data observations $\mathbf{X} \in \mathbb{R}^{n\times d}$ drawn from k distributions, matrix factorization algorithms aims to decompose them into two parts, i.e. coefficient matrix $\mathbf{H} \in \mathbb{R}^{n\times k}$ and base matrix $\mathbf{W} \in \mathbb{R}^{k\times d}$. The most typical matrix factorization method is NMF [1] which regularizes the both matrix to be non-negative, as shown

$$\min_{\mathbf{H}\geq 0,\mathbf{W}\geq 0} f(\mathbf{X}, \mathbf{HW}) \tag{1}$$

where $f(\cdot)$ is the loss function. In most cases, l_2 and Kullback-Leibler divergence loss are adopted [14]. Furthermore, Ding et al. explore the benefits of orthogonality constraint on matrix factorization methods [5]. With adopting l_2 norm and regularizing the base matrix to be orthogonal, Eq. (1) can be formulated as

$$\min_{\mathbf{H}\geq 0,\mathbf{W}\geq 0} \|\mathbf{X} - \mathbf{HW}\|_F^2 \quad s.t.\ \mathbf{WW}^\top = \mathbf{I}_k. \tag{2}$$

The final partition is obtained by performing classical clustering algorithms, mostly k-means, on coefficient matrix $\mathbf{H}$.

2.2. Multi-view matrix factorization

Given the data from V views $\{\mathbf{X}\}_{v=1}^V$, in which $\mathbf{X}_v$ is drawn from $\mathbb{R}^{n\times d_v}$ and d_v is the feature dimension of v-th view, multi-view matrix factorization is to find an optimal $\mathbf{H}$ to reveal the consensus data structure of different views. Liu et al. formulate a joint matrix factorization process with the constraint that pushes coefficient matrix $\mathbf{H}_v$ of each view towards the common consensus $\mathbf{H}$ [6], as shown

$$\begin{aligned}\min_{\mathbf{H}\geq 0,\mathbf{H}_v\geq 0,\mathbf{W}\geq 0} &\sum_{v=1}^V \|\mathbf{X}_v - \mathbf{H}_v\mathbf{W}_v\|_F^2 + \lambda\sum_{v=1}^V g(\mathbf{H},\mathbf{H}_v)\\ &s.t.\ g(\mathbf{H},\mathbf{H}_v) = \gamma_v\|\mathbf{H} - \mathbf{H}_v\mathbf{Q}_v\|_F^2,\end{aligned} \tag{3}$$

where $\mathbf{Q}_v$ is a diagonal matrix for scalar matching. Meanwhile, Gao et al. propose to capture the underlying data structure with the consensus coefficient matrix $\mathbf{H}$ in all data

views [7]. The formulation is presented as

$$\begin{aligned} &\min_{\mathbf{H}\geq 0,\mathbf{W}\geq 0} \sum_{v=1}^{V} \|\mathbf{X}_v - \mathbf{H}\mathbf{W}_v\|_F^2 + \lambda h(\mathbf{H}) \\ &s.t.\ h(\mathbf{H}) = \mathrm{Tr}\left[\mathbf{H}^{*\top}(\sum_{v=1}^{V}\lambda_v \mathbf{L}_v)\mathbf{H}\right], \end{aligned} \tag{4}$$

in which $\mathbf{L}_v$ is the Laplacian matrix of v-th data view. With the obtained consensus coefficient matrix $\mathbf{H}$, standard k-means is adopted to compute the data partition.

3. The proposed method

It can be observed that both single-view and multi-view matrix factorization methods follow a two-step procedure of data clustering. As a result, the coefficient matrices are generated without sufficient guidance of clustering results, leading to unpromising performance. To address this issue, the proposed OPMC combines the two steps in a unified objective.

3.1. Objective

Taking the data $\mathbf{X}_v$ from v-th view, corresponding coefficient matrix can be obtained via

$$\min_{\mathbf{H}_v,\mathbf{W}_v} \|\mathbf{X}_v - \mathbf{H}_v\mathbf{W}_v\|_F^2 \quad s.t.\ \mathbf{W}_v\mathbf{W}_v^\top = \mathbf{I}_k \tag{5}$$

in which the orthogonality regularization is imposed on $\mathbf{W}_v$. Compared with Eq. (2), we remove the non-negativity constraint on $\mathbf{H}_v$ and $\mathbf{W}_v$. This encourages the model to learn a more discriminative embedding in a larger search region. As a by-product, it benefits the optimization process, for closed-form solutions can be obtained on them at each iteration. Furthermore, k-means algorithm aims to partition the data into k disjoint clusters with each characterized by its centroid, which can be formulated into

$$\begin{aligned} &\min_{\mathbf{Y}_v} \|\mathbf{H}_v - \mathbf{Y}_v\mathbf{C}_v\|_F^2 \\ &s.t.\ \mathbf{y}_v^{(i)} \in \{\mathbf{e}_1, \mathbf{e}_2, \cdots, \mathbf{e}_k\}, \end{aligned} \tag{6}$$

where $\mathbf{y}_v^{(i)}$ represents the i-th row of $\mathbf{Y}_v$. Meanwhile, $\mathbf{Y}_v \in \mathbb{R}^{n\times k}$ is the hard partition matrix with each row being an orthonormal basis of k-dimension space. Moreover, $\mathbf{C}_v \in \mathbb{R}^{k\times k}$ is a centroid matrix and its j-th row represents the j-th centroid of $\mathbf{H}_v$.

Rather than combining Eq. (5) and (6) in a unified formulation explicitly, OPMC firstly approximates $\mathbf{H}_v$ into $\mathbf{Y}_v$ and $\mathbf{C}_v$ by

$$\mathbf{H}_v \approx \mathbf{Y}_v\mathbf{C}_v. \tag{7}$$

With unifying Eq. (5) and (7), a 3-factor matrix factorization can be derived as

$$\begin{aligned} &\min_{\mathbf{Y}_v,\mathbf{C}_v,\mathbf{W}_v} \|\mathbf{X}_v - \mathbf{Y}_v\mathbf{C}_v\mathbf{W}_v\|_F^2 \\ &s.t.\ \mathbf{W}_v\mathbf{W}_v^\top = \mathbf{I}_k,\ \mathbf{y}_v^{(i)} \in \{\mathbf{e}_1, \mathbf{e}_2, \cdots, \mathbf{e}_k\}. \end{aligned} \tag{8}$$

Considering V data views, the objective of OPMC is formulated into

$$\begin{aligned} &\min_{\mathbf{Y},\{\mathbf{C}_v\}_{v=1}^V,\{\mathbf{W}_v\}_{v=1}^V} \frac{1}{V}\sum_{v=1}^{V} \|\mathbf{X}_v - \mathbf{Y}\mathbf{C}_v\mathbf{W}_v\|_F^2 \\ &s.t.\ \mathbf{W}_v\mathbf{W}_v^\top = \mathbf{I}_k,\ \mathbf{y}^{(i)} \in \{\mathbf{e}_1, \mathbf{e}_2, \cdots, \mathbf{e}_k\}. \end{aligned} \tag{9}$$

Note that the hard partition is unique in a specific clustering task, therefore, we employ a consensus $\mathbf{Y}$ across all views. At the same time, we set weights of all views to $1/V$. This is called *element-wise* objective, for the loss of every feature element is equally measured. Another widely adopted setting is called *view-wise* in which the weights are float on loss of each view. In experiments, we found *element-wise* OPMC outperforms the *view-wise* one consistently. Nevertheless, it can be observed that no hyper-parameters are required in Eq. (9), which is a great improvement over the recent advances in literature, since there is no validation set for parameter tuning in a clustering task.

3.2. Optimization

In order to optimize Eq. (9), we design an alternate strategy where each unknown variable is solved by fixing the others in each step.

3.2.1 $\mathbf{W}_v$ subproblem

It is obvious that $\{\mathbf{W}_v\}_{v=1}^V$ are independent from each other. Therefore, we fix $\{\mathbf{W}_p\}_{p=1,p\neq v}^V$, $\{\mathbf{C}_v\}_{v=1}^V$ and $\mathbf{Y}$, formulating the optimization respect to $\mathbf{W}_v$ into

$$\max_{\mathbf{W}_v} \mathrm{Tr}(\mathbf{W}_v\mathbf{B}) \quad s.t.\ \mathbf{W}_v\mathbf{W}_v^\top = \mathbf{I}_k, \tag{10}$$

where $\mathbf{B} = \mathbf{X}_v^\top\mathbf{Y}\mathbf{C}_v$. Eq. (11) can be efficiently optimized with singular value decomposition (SVD) technique, while the closed-form solution can be obtained via Theorem 1.

Theorem 1. *Defining economic rank-k singular value decomposition of matrix* $\mathbf{B}$ *as* $\mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$*, the closed-form solution of Eq. (10) should be*

$$\mathbf{W}_v^* = \mathbf{U}\mathbf{V}^\top. \tag{11}$$

Proof. Assuming $\mathbf{F} = \mathbf{V}^\top\mathbf{W}_v\mathbf{U}$, Eq. (10) can be rewrite as $\max_{\mathbf{F}} \mathrm{Tr}(\mathbf{F}\mathbf{\Sigma})$, in which $\mathbf{F}\mathbf{F}^\top = \mathbf{V}^\top\mathbf{W}_v\mathbf{U}\mathbf{U}^\top\mathbf{W}_v^\top\mathbf{V} = \mathbf{I}$. Since $\mathbf{F}$ is orthogonal, all of its elements are from -1 to 1. Meanwhile, $\mathbf{\Sigma}$ is a diagonal matrix and is composed of non-negative singular values $\{\sigma_j\}_{j=1}^k$. Therefore, $\mathrm{Tr}(\mathbf{F}\mathbf{\Sigma}) \leq \sum_{j=1}^k \sigma_j$. The equality holds when $\mathbf{F}$ is an identity matrix, leading to Eq. (11). □

Table 1. Complexity comparison between NMF and the proposed OPMC. NMFc is a naive setting of NMF on multi-view data (More details can be found in Section 4.1). Note that SVD decomposition in Eq. (11) takes $\mathcal{O}(a_1 d_v^2 k + a_2 k^3)$, where a_1 and a_2 are constants [8]. Meanwhile, $d = \sum_{v=1}^{V} d_v$.

Algorithm	Subproblem	Addition	Multiplication	Division	Overall
NMFc	$\mathbf{U}$	$(3k-1)dn - kd$	$3kdn + kd$	kd	$\mathcal{O}(kdn)$
	$\mathbf{V}$	$(3kd - k - d)n$	$k(3d+1)n$	nk	
OPMC	$\mathbf{W}_v$	$d_v n + k(2k-3)d_v + \mathcal{O}(a_1 d_v^2 k + a_2 k^3)$	$2k^2 d_v + \mathcal{O}(a_1 d_v^2 k + a_2 k^3)$	-	$\mathcal{O}(kdn)$
	$\mathbf{C}_v$	$d_v n + (k-1)kd_v - k^2$	$k^2 d_v + k^2$	k	
	$\mathbf{Y}$	$k(2d-1)n + (k-1)kd$	$k^2 d + kdn$	-	

3.2.2 $\mathbf{C}_v$ subproblem

Similar to $\mathbf{W}_v$ subproblem, we fix $\{\mathbf{C}_p\}_{p=1,p\neq v}^{V}$, $\{\mathbf{W}_v\}_{v=1}^{V}$ and $\mathbf{Y}$. As a result, Eq. (9) is reduced to

$$\min_{\mathbf{C}_v} \ \mathrm{Tr}(\mathbf{Y}^\top \mathbf{Y}\mathbf{C}_v\mathbf{C}_v^\top) - 2\mathrm{Tr}(\mathbf{W}_v\mathbf{X}_v^\top \mathbf{Y}\mathbf{C}_v). \tag{12}$$

With setting its derivation to zero, the minimum can be found when

$$\mathbf{C}_v = (\mathbf{Y}^\top\mathbf{Y})^{-1}\mathbf{Y}^\top\mathbf{X}_v\mathbf{W}_v^\top. \tag{13}$$

3.2.3 $\mathbf{Y}$ subproblem

It can be observed that the i-th row of hard partition $\mathbf{Y}$ satisfies 1-of-K encoding scheme. Therefore, we do an exhaustive search on k candidates, i.e. $\{\mathbf{e}_1, \mathbf{e}_2, \cdots, \mathbf{e}_k\}$, to find out the solution, which can be formally given as

$$\mathbf{y}_i = \{\mathbf{e}_j | j = \arg\min \sum_{v=1}^{V} \|\mathbf{x}_v^{(i)} - (\mathbf{C}_v\mathbf{W}_v)_j\|_F^2\}, \tag{14}$$

where subscript i, j denote the i, j-th row of corresponding matrix.

Additionally, the proposed alternate strategy is outlined in Algorithm 1.

Algorithm 1 One-pass Multi-view Clustering Algorithm

Input: Data $\{\mathbf{X}_v\}_{v=1}^{V}$ and number of cluster k
Output: Hard partition $\mathbf{Y}$

```
Initialize Y, {C}_{v=1}^V and {W}_{v=1}^V randomly.
while true do
    Compute the objective value obj(t) via Eq. (9).
    Update {W}_{v=1}^V via Eq. (11).
    Update {C}_{v=1}^V via Eq. (13).
    Update Y via Eq. (14).
    if (obj(t-1) - obj(t))/obj(t) < 1e^{-5} then
        break;
    end if
end while
return Y
```

3.3. Complexity and convergence

Computational complexity analysis of the proposed OPMC is provided in the following. From the definitions, we can find $\mathbf{X}_v \in \mathbb{R}^{n\times d_v}$, $\mathbf{C}_v \in \mathbb{R}^{k\times k}$ and $\mathbf{W}_v \in \mathbb{R}^{k\times d_v}$. Observing that $\mathbf{Y}$ keeps the cluster assignments, computation can be accelerated by performing *index* and *sum* operations rather than matrix multiplication when involving $\mathbf{Y}$. For instance, $\mathbf{A} = \mathbf{Y}\mathbf{C}_v$ requires $k^2 n$ element-wise multiplications. Alternatively, $\mathbf{A}(i,:) = \mathbf{C}_v(j,:)$ if $\mathbf{Y}(i,j) = 1$, which is obviously much faster than the former method. In specific, the complexity of OPMC is analyzed in Table 1. It can be observed that the proposed algorithm is linear to data number n and of $\mathcal{O}(kdn)$ in which $d = \sum_{v=1}^{V} d_v$. Moreover, it requires fewer operations, including addition, multiplication and division, compared with the classical NMF in each iteration. The two observations demonstrate OPMC's scalability on large-scale multi-view data.

Nevertheless, the proposed OPMC is theoretically guaranteed convergent to a local minimum. Similar to [19], we give out the convergence proof of the proposed OPMC in the following. For the ease of expression, we reformulate the objective in Eq. (9) into

$$\min_{\mathbf{Y},\{\mathbf{C}_v\}_{v=1}^{V},\{\mathbf{W}_v\}_{v=1}^{V}} \mathcal{J}(\mathbf{Y}, \{\mathbf{C}_v\}_{v=1}^{V}, \{\mathbf{W}_v\}_{v=1}^{V}) \tag{15}$$

Since the optimization strategy is a cyclical procedure, we use superscript t to represent the optimization round t. In $\mathbf{W}_v$ subproblem, with given $\mathbf{Y}^{(t)}$ and $\{\mathbf{C}_v\}_{v=1}^{V(t)}$, $\{\mathbf{W}_v\}_{v=1}^{V(t+1)}$ is obtained, resulting in

$$\begin{aligned} &\mathcal{J}(\mathbf{Y}^{(t)}, \{\mathbf{C}_v\}_{v=1}^{V(t)}, \{\mathbf{W}_v\}_{v=1}^{V(t+1)}) \\ &\leq \mathcal{J}(\mathbf{Y}^{(t)}, \{\mathbf{C}_v\}_{v=1}^{V(t)}, \{\mathbf{W}_v\}_{v=1}^{V(t)}). \end{aligned} \tag{16}$$

The similar inequality holds in $\mathbf{C}_v$ and $\mathbf{Y}$ subproblems. Therefore, we can get

$$\begin{aligned} &\mathcal{J}(\mathbf{Y}^{(t+1)}, \{\mathbf{C}_v\}_{v=1}^{V(t+1)}, \{\mathbf{W}_v\}_{v=1}^{V(t+1)}) \\ &\leq \mathcal{J}(\mathbf{Y}^{(t)}, \{\mathbf{C}_v\}_{v=1}^{V(t)}, \{\mathbf{W}_v\}_{v=1}^{V(t)}), \end{aligned} \tag{17}$$

which indicates the objective monotonically decreases along with iterations. Meanwhile, it is obvious that $\mathcal{J}$ is lower bounded by 0. Therefore, the proposed algorithm is guaranteed to be convergent theoretically.

4. Experiment

4.1. Experimental setting

In the following experiments, six multi-view datasets are chosen to evaluate the proposed algorithm, including

1) **HandWritten**[1] [23] collects 2000 digits, where six features are extracted, including 76-D fourier coefficient, 216-D profile correlation, 64-D Karhunen-Love coefficient, 240-D pixel average, 47-D Zernike moment and 6-D morphological features.

2) **Caltech101**[2] [15] contains 9144 pictures of objects belonging to 101 categories. Five features, including 48-D Gabor, 40-D Wavelet Moments, 254-D Cenhist, 512-D GIST and 928-D LBP, are adopted.

3) **SUNRGBD**[3] [22] consists of 10335 RGB and depth image pairs collected by researchers from Princeton University. We employ the deep neural network on the original images to extract features of two views.

4) **NUS-WIDE**[4] [4] is a web image dataset created by Lab for Media Search in National University of Singapore. Six types of features are concerned, including 4-D color histogram, 144-D color correlogram, 73-D edge direction histogram, 128-D wavelet texture, 225-D blockwise color moments and 500-D bag of words based on SIFT descriptions.

5) **AwA**[5] [13] contains 30475 images of 50 animals classes with six extracted features, including 2688-D color histogram, 2000-D local self-similarity, 252-D PHOG, 2000-D SIFT, 2000-D color SIFT and 2000-D SURF features.

6) **YtVideo**[6] [20] consists of 101499 Youtube videos. Five types of features are used, including 64-D audio volume, 512-D vision cuboids histogram, 64-D vision HIST, 647-D vision HOG, 838-D vision MISC features.

Their specifications are listed in Table 2. At the same time, the proposed algorithm is compared with seven comparative methods of linear complexity, including

1) **NMF** [14] (*baseline*). Two settings, i.e. NMFb and NMFc, are concerned. NMFb performs NMF on each view and the best result is reported, while NMFc on concatenated data of all views.

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

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

[3] http://rgbd.cs.princeton.edu/

[4] https://lms.comp.nus.edu.sg/wp-content/uploads/2019/research/nuswide/NUS-WIDE.html

[5] https://cvml.ist.ac.at/AwA/

[6] http://archive.ics.uci.edu/ml/datasets/YouTube+Multiview+Video+Games+Dataset

Table 2. Specifications of the chosen datasets.

Dataset	Number of		
	Samples	Views	Clusters
HandWritten	2000	6	10
Caltech101	9144	5	101
SUNRGBD	10335	2	45
NUS-WIDE	23953	5	31
AwA	30475	6	50
YtVideo	101499	5	31

2) **ONMF** [5] (*baseline*) imposes the orthogonality on NMF. Similarly, two settings including ONMFb and ONMFc are adopted.

3) **MNMF** [6] pushes the indicator matrices, generated from NMF on each view, towards a common consensus instead of fixing it directly.

4) **RMKMC** [3] extends the standard k-means into multi-view setting.

5) **LMSpC** [16] groups large-scale multi-view data by approximating the similarity graph in spectral clustering with bipartite graph.

6) **BMVC** [28] collaboratively encodes multi-view data into compact binary representations, then clusters them with binary matrix factorization.

7) **LMSuC** [11] employs anchor technique to extend subspace clustering algorithm on large-scale multi-view data.

Other matrix factorization based multi-view clustering algorithms, such as GCoNMF [27], MVMF [7], MDMF [30] and DiNMF [24], are not involved in the experiments, since they construct similarity graphs, leading to $\mathcal{O}(n^2)$ or higher complexity. Furthermore, we use the codes which are publicly available on authors' websites, perform grid-search on the parameters recommended in their papers and report the best. To eliminate the randomness, we run all comparative algorithms ten times, and the averages are presented. According to the discussion in section 5, we select the results corresponding to the minimum loss for the proposed OPMC, and also repeat ten times to report the averages. Additionally, the source code of OPMC is opened on Github[7].

4.2. Experiment results

In the following, we design multiple experiments to evaluate effectiveness, superiority, efficiency and convergence of the proposed algorithm.

[7] https://github.com/liujiyuan13/OPMC-code_release

Table 3. Ablation study on NMF and OPMC. Additionally, OPMCs refers to single-view OPMC by setting the view number V to 1.

Dataset	ACC		NMI		Purity		Time (s)	
	ONMFc	OPMCs	ONMFc	OPMCs	ONMFc	OPMCs	ONMFc	OPMCs
HandWritten	61.55	**90.65**	58.52	**83.30**	62.05	**90.65**	48.46	**1.11**
Caltech101	24.61	**25.59**	41.13	**46.08**	38.92	**44.82**	448.62	**31.08**
SUNRGBD	**18.98**	18.61	21.95	**25.32**	35.89	**39.36**	2742.80	**272.25**
NUS-WIDE	13.77	**16.28**	11.12	**15.10**	21.38	**26.75**	589.72	**35.29**
AwA	08.05	**09.09**	09.34	**11.59**	09.97	**11.31**	5535.58	**1090.61**
YtVideo	16.85	**22.52**	10.41	**20.53**	27.03	**31.02**	4023.83	**1029.63**

Table 4. Parameter number and performance comparison between the proposed OPMC and seven large-scale algorithms in literature. '-' indicates the algorithm fails on corresponding datasets due to memory limitation.

Dataset	NMFb	NMFc	ONMFb	ONMFc	MNMF	RMKMC	LMSpC	BMVC	LMSuC	OPMC
Param. num.	1	**0**	1	**0**	1	1	1	6	2	**0**
						ACC				
HandWritten	72.63	58.57	69.29	61.55	66.34	69.62	51.06	86.40	**92.10**	90.30
Caltech101	23.38	19.18	22.14	24.61	20.73	16.40	-	**27.71**	21.17	25.18
SUNRGBD	17.84	15.61	17.87	18.98	18.57	18.06	11.30	16.39	17.71	**19.44**
NUS-WIDE	13.37	11.82	12.35	13.77	12.91	15.40	-	15.30	12.46	**16.37**
AwA	08.31	06.13	08.48	08.05	06.74	08.89	-	**10.45**	08.18	09.49
YtVideo	17.55	03.82	16.59	16.85	10.17	12.38	-	19.41	17.25	**23.34**
						NMI				
HandWritten	65.76	49.98	65.98	58.52	60.33	69.21	47.60	84.03	**86.49**	82.73
Caltech101	44.20	40.51	40.28	41.13	41.53	27.47	-	45.33	43.49	**46.41**
SUNRGBD	22.36	21.75	21.42	21.95	23.29	23.86	07.20	19.22	20.71	**25.78**
NUS-WIDE	11.53	10.94	09.94	11.12	10.60	14.28	-	12.92	09.67	**15.32**
AwA	08.81	07.72	09.42	09.34	07.59	11.14	-	**12.30**	09.03	11.71
YtVideo	16.22	00.14	15.64	10.41	08.24	10.17	-	15.80	14.08	**20.74**
						Purity				
HandWritten	74.21	59.84	71.84	62.05	67.50	72.93	53.76	86.40	**92.10**	90.30
Caltech101	43.85	39.87	38.32	38.92	40.82	28.87	-	44.13	42.05	**44.59**
SUNRGBD	36.87	36.02	35.19	35.89	38.40	38.24	18.28	33.28	35.42	**40.46**
NUS-WIDE	24.69	22.65	20.64	21.38	23.40	26.15	-	25.04	21.02	**26.88**
AwA	10.50	08.21	10.75	09.97	08.57	11.02	-	**12.19**	10.03	11.23
YtVideo	28.19	26.62	27.74	27.03	26.68	26.87	-	30.78	**32.25**	31.78

4.2.1 Effectiveness

To demonstrate effectiveness of the proposal, we conduct an ablation study, where two algorithms are compared, i.e. ONMF and single-view OPMC with $V = 1$. In such setting, the only difference between them are whether the non-negativity constraint is imposed and the two steps are unified into a single objective. Nevertheless, we feed the concatenated data of all views to them and their performances and execution times are investigated in Table 3. It can be seen that OPMC outperforms ONMF in all metrics, including ACC, NMI and Purity, on *HandWritten*, *Caltech101*, *NUS-WIDE*, *AwA* and *YtVideo*. Although a little decrease in ACC, is observed on *SUNRGBD*, OPMC achieves consistent increases in NMI and Purity on all chosen datasets. Furthermore, OPMC is 5-10 times faster than ONMF. This is caused by the following two points:

1) OPMC approximates coefficient matrix $\mathbf{H}$ into a discrete label matrix $\mathbf{Y}$ and a centroid matrix $\mathbf{C} \in \mathbb{R}^{k \times k}$, which largely reduce the variable number and search region.

2) ONMF is optimized with gradient descent technique. On the contrary, OPMC removes the non-negativity constraint, therefore, adopts the alternate strategy where closed-form solutions are obtained in each step, requiring fewer iterations to converge.

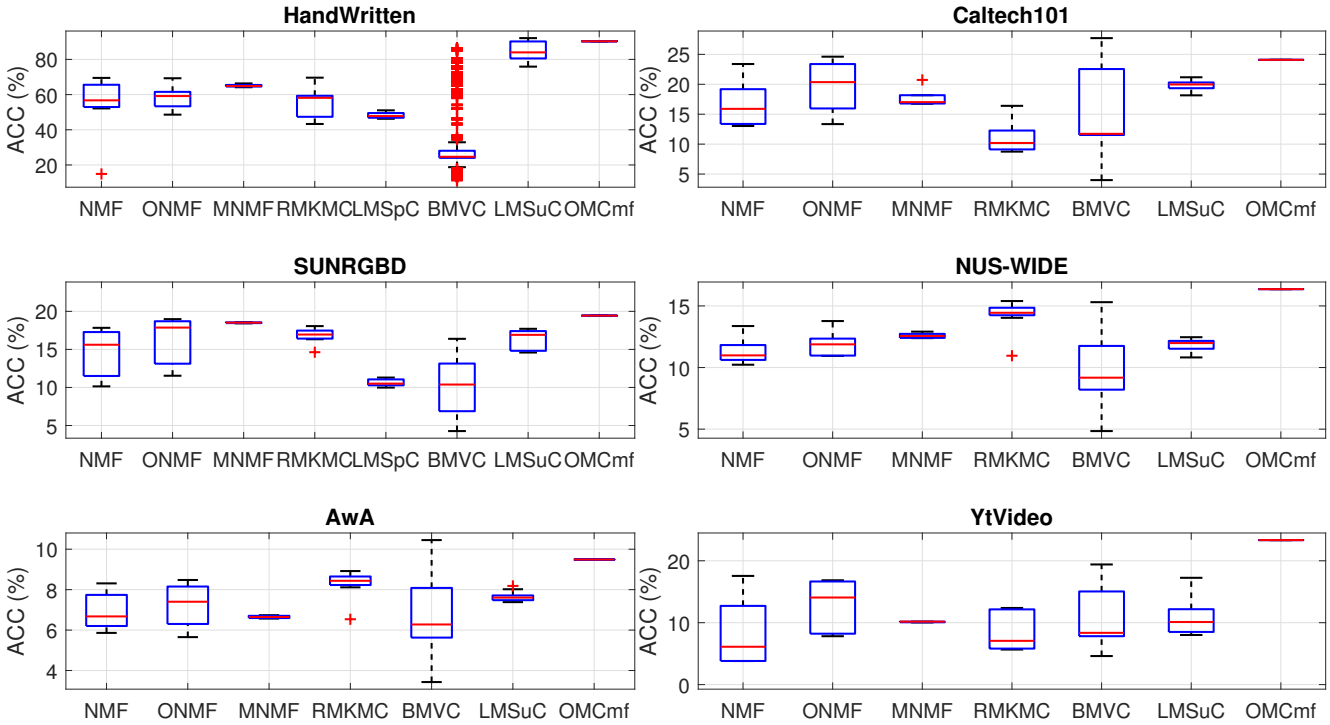

Figure 2. ACC with different parameter settings. LMSpC on *NUS-WIDE*, *AwA* and *YtVideo* are not shown due to memory overflow.

Table 5. Execution time comparison between the proposed OPMC and seven large-scale algorithms in literature.

Dataset	NMFb	NMFc	ONMFb	ONMFc	MNMF	RMKMC	LMSpC	BMVC	LMSuC	OPMC
HandWritten	**0.69**	1.80	36.89	48.46	32.04	11.18	21.91	3.74	13.37	0.83
Caltech101	71.52	89.76	193.42	448.62	1729.35	10722.52	-	**4.65**	661.88	46.04
SUNRGBD	186.07	29.26	2028.65	2742.80	5583.03	583.15	5727.64	**6.90**	774.87	204.24
NUS-WIDE	86.86	104.08	145.97	589.72	1600.64	3854.59	-	60.43	9372.25	**56.36**
AwA	163.91	593.39	1063.86	5535.58	15952.07	6245.74	-	**68.90**	2379.84	1474.19
YtVideo	488.79	555.58	2249.03	4023.83	9553.65	**59.69**	-	151.24	8291.37	671.92

Overall, jointly performing matrix factorization and partition generation while removing the non-negativity constraint can improve the performance and efficiency.

4.2.2 Superiority

In order to validate the superiority of the proposed algorithms, we conduct extensive experiments on comparative methods in literature. Their performances are collected in Table 4. Three observations can be obtained as follows:

1) The proposed OPMC outperforms the best of baselines including NMFb, NMFc, ONMFb and ONMFc, over all datasets. Meanwhile, some baselines achieves better performances than several comparative methods on a part of datasets. For instance, ONMFc shows to be the second best , i.e. 18.98% in ACC, on *SUNRGBD*, which conversely demonstrate the superiority of OPMC.

2) It can be observed that OPMC consistently exceeds the recent advances in literature on *Caltech101*, *SUNRGBD*, *NUS-WIDE* and *YtVideo*. Although LMSuC and BMVC achieves better performances on *HandWritten* and *AwA*, respectively, they perform grid search on multiple parameters. OPMC outperforms them in most parameter settings, as shown in Fig. 2.

3) LMSpC fails on *Caltech101*, *NUS-WIDE*, *AwA* and *YtVideo* due to memory overflow. This also illustrates OPMC's scalability on large-scale datasets.

Furthermore, the proposed OPMC is compared with the comparative methods on six datasets with different parameters. We grid search their parameters ten times and compute the averages in each parameter setting. The obtained averages are presented in Fig. 2. Note that, although the two baselines do not require parameters, they need to choose which data view to handle with, therefore, their perfor-

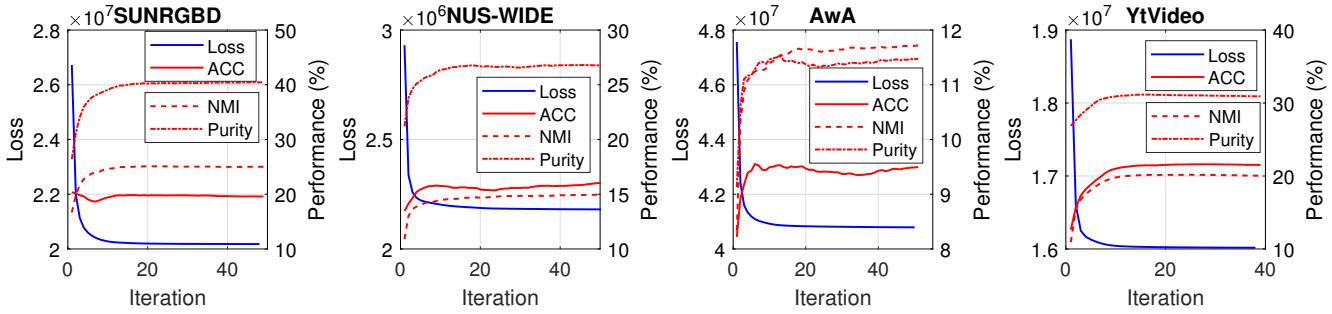

Figure 3. Loss and performance along with iteration on four large-scale datasets, including *SUNRGBD*, *NUS-WIDE*, *AwA* and *YtVideo*.

mances are not constants. It can be seen that performances of the most comparative methods, including NMF, ONMF, RMKMC, BMVC and LMSuC, are highly dependent on parameter choice. Though MNMF obtains stable results in different parameter settings, the proposed OPMC outperforms it by large margins. Overall, with summarizing Table 4 and Fig. 2, we can conclude that OPMC achieves state-of-the-art performance and is much more feasible in real-world applications due to its non-parametric property.

4.2.3 Efficiency and convergence

To demonstrate the linear computational complexity of OPMC, we ran all algorithms ten times on the chosen datasets and collect their averages in Table 5. For fair comparison, they are executed in parallel where one *Intel(R) Core(TM) i9-10900X CPU @ 3.70GHz* is allocated each time. It can be observed that the proposed OPMC shows comparable results with classical NMF, which is also consistent with the complexity analysis in Table 1. Meanwhile, the loss of OPMC on *SUNRGBD*, *NUS-WIDE*, *AwA* and *YtVideo* along with iteration is presented in Fig. 3. Results on relatively smaller datasets, i.e. *HandWritten* and *Caltech101*, are presented in Appendix due to space limit. It can be found that the objective value monotonically decreases to a minimum, validating the converge analysis in Section 3.3 experimentally. We also observe that OPMC's performances increase with the decrease of loss in Fig. 3, verifying its rationality and effectiveness .

5. Discussion

The proposed OPMC is guaranteed convergent to a local minimum instead of a global one. As a result, its initialization will inevitably introduce randomness to the final partition, which is especially obvious in small datasets, such as *HandWritten*. To obtain a better performance, we recommend to repeat OPMC multiple times and select the results corresponding to the smallest loss. Fig. 4 shows OPMC's performances, including ACC and NMI, respect to the objective loss of 100 times. Purity presents too many overlaps with ACC, therefore, is shown in Appendix for clarity. It can be found that OPMC's performances are negatively correlated with loss. This observation well validates effectiveness of the objective design and our recommendation.

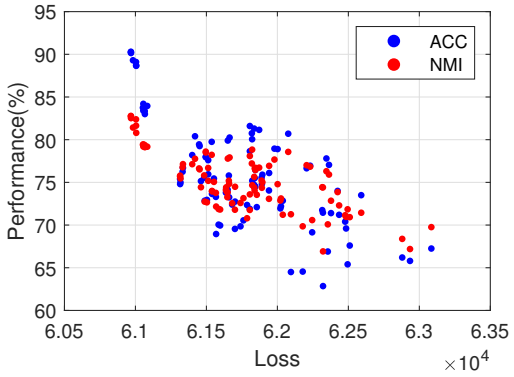

Figure 4. ACC and NMI variation respect to objective loss on *HandWritten*.

6. Conclusion

Current non-negative matrix factorization based multi-view clustering algorithms decompose the data view and generate the clustering results separately. However, the non-negativity constraint over-limits the discriminative embedding learning. Meanwhile, they fail to unify matrix factorization and partition generation closely, resulting in unsatisfying performance. Therefore, we propose an one-pass multi-view clustering algorithm, which is non-parametric and of linear complexity. Its effectiveness, superiority and efficiency are validated by comparing with recent advances.

Acknowledgments

The work is supported by National Key R&D Program of China (No. 2020AAA0107100), National Natural Science Foundation of China (No. 61922088, 61773392, 61872377, 61976196 and 62006236), Education Ministry-China Mobile Research Funding (No. MCM20170404), Hunan Provincial Natural Science Foundation (No. 2020JJ5673) and NUDT Research Project (No. ZK20-10).

References

[1] Michael W. Berry, Murray Browne, Amy Nicole Langville, V. Paul Pauca, and Robert J. Plemmons. Algorithms and applications for approximate nonnegative matrix factorization. *Comput. Stat. Data Anal.*, 52(1):155–173, 2007.

[2] Deng Cai, Xiaofei He, Jiawei Han, and Thomas S. Huang. Graph regularized nonnegative matrix factorization for data representation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 33(8):1548–1560, 2011.

[3] Xiao Cai, Feiping Nie, and Heng Huang. Multi-view k-means clustering on big data. In *IJCAI 2013, Proceedings of the 23rd International Joint Conference on Artificial Intelligence, Beijing, China, August 3-9, 2013*, pages 2598–2604. IJCAI/AAAI, 2013.

[4] Tat-Seng Chua, Jinhui Tang, Richang Hong, Haojie Li, Zhiping Luo, and Yantao Zheng. NUS-WIDE: a real-world web image database from national university of singapore. In Stéphane Marchand-Maillet and Yiannis Kompatsiaris, editors, *Proceedings of the 8th ACM International Conference on Image and Video Retrieval, CIVR 2009, Santorini Island, Greece, July 8-10, 2009*. ACM, 2009.

[5] Chris H. Q. Ding, Tao Li, Wei Peng, and Haesun Park. Orthogonal nonnegative matrix tri-factorizations for clustering. In *Proceedings of the Twelfth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Philadelphia, PA, USA, August 20-23, 2006*, pages 126–135. ACM, 2006.

[6] Jing Gao, Jiawei Han, Jialu Liu, and Chi Wang. Multi-view clustering via joint nonnegative matrix factorization. In *Proceedings of the 13th SIAM International Conference on Data Mining, May 2-4, 2013. Austin, Texas, USA*, pages 252–260. SIAM, 2013.

[7] Shengxiang Gao, Zhengtao Yu, Taisong Jin, and Ming Yin. Multi-view low-rank matrix factorization using multiple manifold regularization. *Neurocomputing*, 335:143–152, 2019.

[8] Judith D. Gardiner. Fundamentals of matrix computations (david s. watkins). *SIAM Rev.*, 35(3):520–521, 1993.

[9] Zhenyu Huang, Joey Tianyi Zhou, Xi Peng, Changqing Zhang, Hongyuan Zhu, and Jiancheng Lv. Multi-view spectral clustering network. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, pages 2563–2569. ijcai.org, 2019.

[10] Zhao Kang, Xinjia Zhao, Chong Peng, Hongyuan Zhu, Joey Tianyi Zhou, Xi Peng, Wenyu Chen, and Zenglin Xu. Partition level multiview subspace clustering. *Neural Networks*, 122:279–288, 2020.

[11] Zhao Kang, Wangtao Zhou, Zhitong Zhao, Junming Shao, Meng Han, and Zenglin Xu. Large-scale multi-view subspace clustering in linear time. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 4412–4419. AAAI Press, 2020.

[12] Yehuda Koren, Robert M. Bell, and Chris Volinsky. Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37, 2009.

[13] Christoph H. Lampert, Hannes Nickisch, and Stefan Harmeling. Attribute-based classification for zero-shot visual object categorization. *IEEE Trans. Pattern Anal. Mach. Intell.*, 36(3):453–465, 2014.

[14] Daniel D. Lee and H. Sebastian Seung. Algorithms for non-negative matrix factorization. In *Advances in Neural Information Processing Systems 13, (NIPS) 2000, Denver, CO, USA*, pages 556–562. MIT Press, 2000.

[15] Fei-Fei Li, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2004, Washington, DC, USA, June 27 - July 2, 2004*, page 178. IEEE Computer Society, 2004.

[16] Yeqing Li, Feiping Nie, Heng Huang, and Junzhou Huang. Large-scale multi-view spectral clustering via bipartite graph. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, January 25-30, 2015, Austin, Texas, USA*, pages 2750–2756. AAAI Press, 2015.

[17] Jiyuan Liu, Xinwang Liu, Jian Xiong, Qing Liao, Sihang Zhou, Siwei Wang, and Yuexiang Yang. Optimal neighborhood multiple kernel clustering with adaptive local kernels. *IEEE Transactions on Knowledge and Data Engineering*, 2020.

[18] Jiyuan Liu, Xinwang Liu, Yuexiang Yang, Xifeng Guo, Marius Kloft, and Liangzhong He. Multiview subspace clustering via co-training robust data representation. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–13, 2021.

[19] Jiyuan Liu, Xinwang Liu, Yuexiang Yang, Siwei Wang, and Sihang Zhou. Hierarchical multiple kernel clustering. In *Proceedings of the Thirty-fifth AAAI Conference on Artificial Intelligence, (AAAI-21), Virtually, February 2-9, 2021*, 2021.

[20] Omid Madani, Manfred Georg, and David A. Ross. On using nearly-independent feature families for high precision and confidence. *Mach. Learn.*, 92(2-3):457–477, 2013.

[21] Xi Peng, Zhenyu Huang, Jiancheng Lv, Hongyuan Zhu, and Joey Tianyi Zhou. COMIC: multi-view clustering without parameter selection. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97, pages 5092–5101. PMLR, 2019.

[22] Shuran Song, Samuel P. Lichtenberg, and Jianxiong Xiao. SUN RGB-D: A RGB-D scene understanding benchmark suite. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 567–576. IEEE Computer Society, 2015.

[23] Martijn van Breukelen, Robert P. W. Duin, David M. J. Tax, and J. E. den Hartog. Handwritten digit recognition by combined classifiers. *Kybernetika*, 34(4):381–386, 1998.

[24] Jing Wang, Feng Tian, Hongchuan Yu, Chang Hong Liu, Kun Zhan, and Xiao Wang. Diverse non-negative matrix factorization for multiview data representation. *IEEE Trans. Cybern.*, 48(9):2620–2632, 2018.

[25] Jie Wen, Zheng Zhang, Yong Xu, and Zuofeng Zhong. Incomplete multi-view clustering via graph regularized matrix

factorization. In *Computer Vision - ECCV 2018 Workshops - Munich, Germany, September 8-14, 2018, Proceedings, Part IV*, volume 11132 of *Lecture Notes in Computer Science*, pages 593–608. Springer, 2018.

[26] Changqing Zhang, Yajie Cui, Zongbo Han, Joey Tianyi Zhou, Huazhu Fu, and Qinghua Hu. Deep partial multi-view learning. *CoRR*, abs/2011.06170, 2020.

[27] Xinyu Zhang, Hongbo Gao, Guopeng Li, Jianhui Zhao, Jianghao Huo, Jialun Yin, Yuchao Liu, and Li Zheng. Multi-view clustering based on graph-regularized nonnegative matrix factorization for object recognition. *Inf. Sci.*, 432:463–478, 2018.

[28] Zheng Zhang, Li Liu, Fumin Shen, Heng Tao Shen, and Ling Shao. Binary multi-view clustering. *IEEE Trans. Pattern Anal. Mach. Intell.*, 41(7):1774–1782, 2019.

[29] Zhenyue Zhang and Keke Zhao. Low-rank matrix approximation with manifold regularization. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(7):1717–1729, 2013.

[30] Handong Zhao, Zhengming Ding, and Yun Fu. Multi-view clustering via deep matrix factorization. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, pages 2921–2927. AAAI Press, 2017.