Neurocomputing 403 (2020) 53–62

Contents lists available at ScienceDirect

Neurocomputing

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

Multi-view image clustering based on sparse coding and manifold consensus

Xiaofei Zhu[a,*], Jiafeng Guo[b], Wolfgang Nejdl[c], Xiangwen Liao[d], Stefan Dietze[e]

[a] College of Computer Science and Engineering, Chongqing University of Technology, Chongqing 400054, China
[b] Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China
[c] L3S Research Center, Leibniz Universität Hannover, Hannover 30167, Germany
[d] College of Mathematics and Computer Science, Fuzhou University, Fuzhou 350108, China
[e] Knowledge Technologies for the Social Sciences, Leibniz Institute for the Social Sciences, Cologne 50667, Germany

ARTICLE INFO

Article history:
Received 4 January 2019
Revised 3 December 2019
Accepted 13 March 2020
Available online 26 March 2020

Communicated by Dr. Min Xu

Keywords:
Multi-view clustering
Sparse coding
Manifold consensus

ABSTRACT

Multi-view clustering has received an increasing attention in many applications, where different views of objects can provide complementary information to each other. Existing approaches on multi-view clustering mainly focus on extending Non-negative Matrix Factorization (NMF) by enforcing the constraint over the coefficient matrices from different views in order to preserve their consensus. In this paper, we argue that it is more reasonable to utilize the high-level manifold consensus rather than the low-level coefficient matrix consensus (as conducted in state-of-the-art approaches) to better capture the underlying clustering structure of the data. For this purpose, we propose MMRSC (Multiple Manifold Regularized Sparse Coding), which aims to preserve the consensus over multiple manifold structures from different views. Experimental results on two publicly available real-world image datasets demonstrate that our proposed approach can significantly outperform the state-of-the-art approaches for the multi-view image clustering task. Moreover, we also conduct computational complexity analysis and the result shows that MMRSC can effective handle the multi-view clustering problem without increasing the computational cost as compared to GraphSC.

1. Introduction

In the past decade, people have witnessed an increasing amount of available image data which are comprised of multiple views (or representations). For example, on Flickr, an image can be represented by its visual contents, annotated tags, social comments, and so on. These different image views usually provide complementary information to each other, and a fundamental problem is how to integrate the multiple image views effectively in order to obtain a better representation.

Many research efforts have been devoted to clustering objects based on their multiple representations, referred to as multi-view clustering. A straightforward solution is to convert these multiple representations of objects into a unified long feature vector by simply concatenating all representations. However, in most real-world applications, multiple views would have different properties, thus this simple solution may not work effectively. Recently, a number of research work has appeared in the literature for addressing the multi-view clustering problem, and the state-of-the-art approaches are these Non-negative Matrix Factorization (NMF) based models [1–3]. The hypothesis of these methods is that different views should reflect the same underlying clustering structure of the data. These studies resort to extending the NMF algorithm to handle multi-view clustering by enforcing different consensus constraints over the coefficient matrices learnt from different views. For example, Akata et al. [1] adopt a hard-consensus constraint by utilizing a shared common coefficient matrix for different views. Some researchers [2,3] also attempt to employ a soft-consensus constraint by enforcing the coefficient matrices learnt from different views either towards a common consensus [2] or to be similar to each other [3]. Although these NMF-based approaches have achieved promising results in several studies, it would be not optimal to apply these techniques to the domain of multi-view image clustering because: (1) the properties of different views of image may vary greatly (e.g., varying from visual features to textual features), and make it improper to learn a representation by using the low-level coefficient matrix consensus over different views. (2) the noise issue, which has been reported as one of the frequently faced problems in image processing [4], has been largely ignored in existing NMF-based approaches.

* Corresponding author.
E-mail address: zxf@cqut.edu.cn (X. Zhu).

https://doi.org/10.1016/j.neucom.2020.03.052
0925-2312/

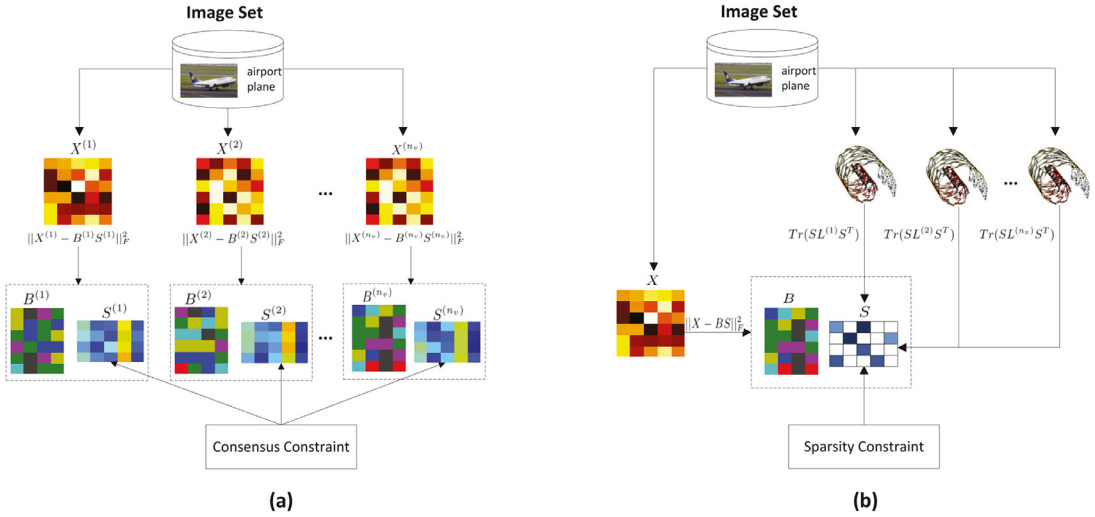

Fig. 1. Comparison of the frameworks of (a) previous NMF-based approaches (e.g., CollNMF, MultiNMF and CoNMF) and (b) our proposed approach MMRSC. Previous NMF-based approaches seek to utilize the low-level coefficient matrix consensus constraint over different views by enforcing either a hard-consensus constraint (e.g., CollNMF) or a soft-consensus constraint (e.g., MultiNMF and CoNMF). In contrast, our MMRSC resorts to adopting the high-level manifold consensus over different views, in which the learnt representation can reflect the intrinsic clustering structure of different views. In addition, MMRSC also adopts the sparse coding framework instead of NMF-based framework in order to make the learnt representation more robust to noise.

In this paper, we propose to leverage the high-level manifold consensus instead of the low-level coefficient matrix consensus to handle the multi-view image clustering. To this end, we extend the GraphSC (Graph regularized Sparse Coding) in order to have a capability to enforce the learnt representation to be consistent with the underlying manifold structures from different views. It is worth noting that GraphSC is specifically designed for dealing with a single-view scenario, and it is impractical to directly apply GraphSC to handle the multi-view image clustering problem, which has been verified in our experiments. By utilizing the high-level manifold consensus, MMRSC can effectively deal with the limitation of existing state-of-the-art approaches which enforce consensus constraint over the low-level coefficient matrices. In MMRSC, we construct a set of graph Laplacians to represent the underlying manifold structures of different views, and incorporate them into our optimization process. Through this process, the learnt representation can better reflect the intrinsic clustering structures of different views and preserve the high-level manifold consensus. In addition, since MMRSC is based on a sparse coding framework, which leads to the learnt image representation to be more robust to noise. To the best of our knowledge, this is the first work on utilizing the high-level manifold consensus as well as sparse coding for multi-view image clustering problem. Fig. 1 shows a comparison between these NMF-based Multi-View Clustering methods (i.e., CollNMF [1], MultiNMF [2] and CoNMF [3]) and our proposed MMRSC.

The major contributions of this paper are summarized as follows: (1) exploiting the intrinsic manifold structures from different views to capture their underlying clustering structures, and addressing the multi-view image clustering problem by incorporating the high-level manifold consensus constraint rather than the low-level coefficient matrix consensus as adopted in state-of-the-art approaches. (2) Proposing the MMRSC by extending the GraphSC to have a substantial capability for addressing the multi-view image clustering problem. (3)Providing the computational complexity analysis of MMRSC. (4) Evaluating the performance of our approach on two real-world image datasets, which have quite different properties, and experimental results show that our method consistently outperforms the state-of-the-art multi-view clustering approaches.

The rest of the paper is organized as follows. Section 2 describes the related work. In Section 3, we discuss the details of our proposed approach. The experimental results are reported in Section 4. In the end, Section 5 concludes the work and discusses some future work.

2. Related work

In this section, we will discuss previous works which are most related to our proposed approach. These works can be generally grouped into two categories, namely, multi-view clustering and sparse coding.

2.1. Multi-view clustering

Recently, multi-view clustering [5,6] has received a lot of attentions. Previous approaches of multi-view clustering can be briefly grouped into three categories: early integration, intermediate integration, and late integration approaches. Early integration methods attempt to first integrate multiple views into a unified view, and then apply existing clustering algorithm on the integrated view. For example, Chaudhuri et al. [7] learn a low-dimensional subspace by applying Canonical Correlation Analysis (CCA) over multi-view data, and employ k-means on the subspace to obtain the clustering. Intermediate integration algorithms [8,9] integrate multiple views during the clustering process, e.g., Ramage et al. [9] propose to extend LDA by assuming topics from different views sharing a common underlying distribution. At last, late integration methods [10,11] conduct clustering over each view individually, and then fuse these results towards a common consensus. For example, Greene et al. [11] construct an intermediate matrix representation of these clustering results from each view, and then apply a factorization procedure over the new representation for clustering.

The works most closely related to our work are the NMF-based multi-view clustering approaches. Akata et al. [1] propose to learn a shared coefficient matrix across different views through a joint non-negative matrix factorization process, referred to as CollNMF. However, enforcing such a hard-consensus constraint often leads to poor performance due to the fact that distinct views would carry

different properties. It also has been identified that when no normalization is carried out, this method is equivalent to conducting NMF on a concatenated representation of all views. To address this issue, Liu et al. [2] propose an extended joint matrix factorization method, named MultiNMF. Instead of learning a common shared coefficient matrix, MultiNMF introduces a soft-consensus regularization into the matrix factorization process in order to enforce the learnt coefficient matrices from different views towards a common consensus matrix. He et al. [3] further relaxes the constraint of MultiNMF by employing co-regularization in the factorization process and propose a novel multi-view clustering algorithm, called CoNMF. In CoNMF, they introduce two instantiations (i.e., pair-wise CoNMF and cluster-wise CoNMF) to co-regularize on each pair of views. Xu et al. [12] employs the $l_{2,1}$-norm to constrain the canonical loadings, measure the canonical correlation loss term, as well as tackle noise issue to some degree.

Although these algorithms have demonstrated promising performance, they will inevitably suffer from the limitation that enforcing the consensus constraint via the low-level coefficient matrix. It will restrict its capability to effectively represent multiple image views, in which the clustering quality of different views varies considerably. In our approach, we seek to address the multi-view image clustering problem by introducing consensus constraint through a high-level manifold consensus. It is worth noting that Gao et al. [13] also consider the manifold information in their model, but this model relies on class label information which is not available in the multi-view clustering scenario. While our method is learned in an unsupervised mode, which does not rely on any class label information. Another difference between [13] and our work is that in [13] it aims at making use of hypergraph to deal with noise issue while we leverage sparse coding instead. Although hypergraph is considered more clean than traditional nearest graph, constructing a hypergraph is computational expensive as it needs to solve a linear regression problem, i.e., a sparse regression problem or an elastic-net regularized linear regression problem.

2.2. Sparse coding

Sparse coding (SC) [14,15] has been successfully employed in many applications, such as face recognition [16,17], image clustering [18], image classification [19], etc. The limitation of sparse coding is that it needs to solve a nondifferentiable ℓ_1-norm problem, while the computational cost is very expensive. Several research work has been proposed to solve this optimization problem effectively. For example, Lee et al. [15] accelerate the optimization process by adopting a feature-sign search method, which can reduce the nondifferentiable problem to an unconstrained quadratic programming (QP). To improve the quality of sparse representations, recently, many variants of sparse coding methods have been proposed via imposing additional constraints into the objective function. For example, Liu et al. [20] extend sparse coding by adding a nonnegative constraint. Kavukcuoglu et al. [21] enforce a spatial consistent constraint to learn local invariant sparse representations. Zheng et al. [18] incorporate the manifold structure into the sparse coding and propose graph regularized method, called GraphSC.

It is worth noting that our MMRSC is an extension of GraphSC by incorporating multiple graph Laplacian regularizers into the objective function, and aims to preserve the high-level manifold consensus across multiple views. Although MMRSC takes a similar form to that of GraphSC, the key difference comes from the substantial capability of MMRSC on handling multi-view clustering problem. To the best of our knowledge, this is the first work that employing sparse coding, instead of NMF-based framework, to cope with the multi-view clustering issue. By further introducing the sparsity constraint into the objective function, it makes the learnt coefficient matrix of MMRSC have a nice sparsity property, which can be computationally more efficient and more robust to noise. Although there are some extensions of GraphSC, which are substantially different to ours. For example, in [22], Long et al. extend GraphSC for transfer learning through taking into account the minimization of distribution divergence between labeled and unlabeled data.

3. Our proposed approach

In this section, we first briefly introduce of the concept of Sparse Coding, and then we formally illustrate our proposed approach.

3.1. Sparse coding(SC)

Given a set of data points $X = [\mathrm{x}_1, \ldots, \mathrm{x}_n] \in \mathbb{R}^{d\times n}$, where n is the number of data points and d is the dimensionality of the feature space, let $B = [\mathrm{b}_1, \ldots, \mathrm{b}_m] \in \mathbb{R}^{d\times m}$ be the dictionary (or basis) matrix, where each column b_i represents a basis vector in the dictionary, and $S = [\mathrm{s}_1, \ldots, \mathrm{s}_n] \in \mathbb{R}^{m\times n}$ be the coefficient (or coding) matrix, where each column s_i is a sparse representation for a data point x_i. SC aims to represent the input dataset as a sparse linear combination of the basis vectors in the dictionary, where each data point x_i is well represented by only a small number of non-zero coefficients, i.e., $\mathrm{x}_i \approx \sum_{j=1}^{m} \mathrm{b}_j s_i^{(j)} = B\mathrm{s}_i$. To this end, SC needs to find a good representation which can minimize the reconstruction error, as well as selecting only a few basis vectors to linearly reconstruct the original feature vectors [18,23,24]. The objective function of SC can be formulated as follows:

$$\min_{B,S} ||X - BS||_F^2 + \beta \sum_{i=1}^{n} ||s_i||_1$$
$$s.t. ||b_i||^2 \le c, i = 1, \ldots, m \tag{1}$$

where $\|\cdot\|_F$ denotes the matrix Frobenius norm, β is a tunable parameter for controlling the trade-off between reconstruction error and sparsity, and constant c is used to impose a norm constraint for the basis vectors.

3.2. Multi-manifold regularized sparse coding

In our approach, we aim at learning a new representation which can better capture the underlying clustering structure of each view. The basic assumption is that the learnt representation should vary smoothly along the manifolds of different views, i.e., if two data points x_i and x_j are close in more view geometries, their corresponding coefficients s_i and s_j should be more close to each other with respect to the new basis B.

3.2.1. Objective function

To this end, we propose to exploit the manifold structure embedded in each view and incorporate them as set of graph Laplacian constraints into the sparse coding framework.

Formally, let $X^{(1)}, X^{(2)}, \ldots, X^{(n_v)}$ denote the n_v views. Here for the vth view, we build a k-nearest neighbor graph, denoted as $G^{(v)}$, to encode its manifold information. Let $W^{(v)}$ be the weight matrix corresponding to $G^{(v)}$, where $w_{ij}^{(v)} = 1$ if x_i and x_j are among the k-nearst neighbors of each other with respect to the vth view, otherwise $w_{ij}^{(v)} = 0$. We then define the Laplacian matrix as $L^{(v)} = W^{(v)} - D^{(v)}$, where $D^{(v)}$ is a diagonal matrix with (i, i)-element equal to the sum of the ith row of $W^{(v)}$.

In order to preserve the manifold structures of multiple views, we represent these manifold structures as a set of graph Laplacian constraints, which can be easily formalized as $\frac{1}{2}\sum_{i,j=1}^{n}||s_i - s_j||^2 W_{ij}^{(v)} = Tr(SL^{(v)}S^T), v = 1,\ldots,n_v$, and incorporate these constraints into the objective function in Eq. (1). Therefore, the objective function of MMRSC can be formalized as:

$$\min_{B,S} ||X - BS||_F^2 + \sum_{v=1}^{n_v}\alpha_v Tr(SL^{(v)}S^T) + \beta\sum_{i=1}^{n}||s_i||_1$$
$$s.t. ||b_i||^2 \le c, i = 1,\ldots,m \tag{2}$$

where X is the original data representation[1], n_v is the number of graph Laplacian constraints, and $\alpha_v \ge 0$ is the graph regularization parameter of the vth manifold. When we increase α_v in Eq. (2), the influence of the vth manifold regularizer increases, and the corresponding effect is that s_i and s_j become more similar to each other if they are close in the vth view. On the other hand, when we decrease α_v, the influence of the vth manifold regularizer will decrease as well. In an extreme case, if we set all $\alpha_v = 0$, $v = 1,\ldots,n_v$, our approach will regress to the standard sparse coding. In Section 4.6, we will discuss in detail the impact of parameters on the performance of our method.

The objective function in (2) is convex either in B or in S, while it is not convex in both of them simultaneously. For learning S and B, we resort to an iteratively optimization method as proposed in [15]. The optimization contains two steps: (1) fix the dictionary B while learning coefficients S; then (2) fix the coefficients S while learning the dictionary B. We iteratively execute these two steps until convergence, or until a pre-specified iteration number is reached.

3.2.2. Learning sparse coefficient matrix **S**

In this section, we consider how to learn the sparse coefficient matrix S by fixing the dictionary B. For this purpose, the optimization problem (2) becomes:

$$\min_{S} ||X - BS||_F^2 + \sum_{v=1}^{n_v}\alpha_v Tr(SL^{(v)}S^T) + \beta\sum_{i=1}^{n}||s_i||_1 \tag{3}$$

In order to facilitate manipulations in vector form, we rewrite the problem (3) as:

$$\min_{\{s_i\}} \sum_{i=1}^{n}||x_i - Bs_i||^2 + \sum_{i,j=1}^{n}\left(\sum_{v=1}^{n_v}\alpha_v L_{ij}^{(v)}\right)s_i^T s_j + \beta\sum_{i=1}^{n}||s_i||_1 \tag{4}$$

Regarding the regularization terms $\sum_{i,j=1}^{n}(\sum_{v=1}^{n_v}\alpha_v L_{ij}^{(v)})\, s_i^T s_j$ in the problem (4), each s_i is coupled with other coefficient vectors $\{s_j\}_{j\neq i}$. In order to solve this problem, we optimize over each s_i individually by keeping other coefficient vectors fixed, and get the following optimization problem for each s_i:

$$\min_{s_i} f(s_i) = ||x_i - Bs_i||^2 + \left(\sum_{v=1}^{n_v}\alpha_v L_{ii}^{(v)}\right)s_i^T s_i + s_i^T h_i + \beta\sum_{j=1}^{m}|s_i^{(j)}| \tag{5}$$

where $h_i = 2\sum_{j\neq i}(\sum_{v=1}^{n_v}\alpha_v L_{ij}^{(v)})s_j$, and $s_i^{(j)}$ is the jth coefficient of s_i.

Since problem (5) with ℓ_1-regularization is non-differentiable when s_i has values of 0, we cannot adopt the standard unconstrained optimization methods to solve this problem. Several approaches are available for solving this problem [15,25,26]. In this paper, we follow an efficient solution proposed in [15], and use the feature-sign search algorithm to solve the problem (5).

[1] In this paper, we leverage the concatenated representation as X. Note that other representations can also be considered as X.

To simplify notation, we define $g(s_i) = ||x_i - Bs_i||^2 + (\sum_{v=1}^{n_v}\alpha_v L_{ii}^{(v)})s_i^T s_i + s_i^T h_i$, then $f(s_i) = g(s_i) + \beta\sum_{j=1}^{m}|s_i^{(j)}|$. Let $\nabla_i^{(j)}|s_i|$ be the sub-differentiable value of the jth coefficient of s_i. If $|s_i^{(j)}| > 0$, then the absolute value function $|s_i^{(j)}|$ is differentiable, thus $\nabla_i^{(j)}|s_i| = sign(s_i^{(j)})$, and the corresponding optimality condition is $\nabla_i^{(j)}g(s_i) + \beta sign(s_i^{(j)}) = 0$. If $|s_i^{(j)}| = 0$, the absolute value function $|s_i^{(j)}|$ becomes non-differentiable. We then set $\nabla_i^{(j)}|s_i|$ to $[-1, 1]$, and the corresponding optimality condition is $|\nabla_i^{(j)}g(s_i)| \leqslant \beta$.

Here, we consider the case when the optimality condition is violated, i.e., $|\nabla_i^{(j)}g(s_i)| > \beta$, when $|s_i^{(j)}| = 0$. Without loss of generality, we suppose $\nabla_i^{(j)}g(s_i) > \beta$, and it means $\nabla_i^{(j)}f(s_i) > 0$ regardless of the sign of $s_i^{(j)}$. In order to reduce $f(s_i)$, we should decrease $s_i^{(j)}$. Thus we take $sign(s_i^{(j)}) = -1$ as $s_i^{(j)}$ starts at zero. Similarity, if $\nabla_i^{(j)}g(s_i) < -\beta$, then we can take $sign(s_i^{(j)}) = 1$.

Assuming we have known the $sign(s_i^{(j)})$ at the optimal value, then each item $|s_i^{(j)}|$ of the ℓ_1-form in the problem (5) can be replaced by

$$\begin{cases} s_i^{(j)} & \text{if } s_i^{(j)} > 0 \\ -s_i^{(j)} & \text{if } s_i^{(j)} < 0 \\ 0 & \text{if } s_i^{(j)} = 0 \end{cases} \tag{6}$$

Therefore, the problem (5) can be reduced to a standard, unconstrained quadratic optimization problem (QP), which can be solved analytically and efficiently. To sum up, the algorithmic procedure can be described as:

- for each s_i, we search for $\{sign(s_i^{(j)})\}_{j=1,\ldots,m}$;
- get the optimal coefficient s_i^* by solving the reduced QP problem;
- return the optimal coefficients matrix $S^* = [s_1^*,\ldots,s_n^*]$.

The feature-sign search algorithm maintains an active set $A \triangleq \{j|s_i^{(j)} = 0, |\nabla_i^{(j)}| > \beta\}$ of potentially nonzero coefficients and their corresponding signs $\boldsymbol{\theta} = [\theta_1,\ldots,\theta_m]$ while updating each s_i, and it then systematically searches for the optimal active set and coefficient signs. In particular, this algorithm consists of a series of "feature-sign steps", and for each step:

- given a current setting for the active set and the signs, it computes the analytical solution $\hat{s}_i^{new}$ to the resulting unconstrained QP;
- then it updates the solution, the active set and the signs using an efficient discrete line search between the current solution and $\hat{s}_i^{new}$.

Each feature-sign step strictly reduces the objective $f(s_i)$, and the overall algorithm will converge to a global optimum in a finite number of steps [15].

3.2.3. Learning dictionary **B**

For solving the optimization problem in (2) over the dictionary B, we fix the coefficients S and the problem reduces to a least squares problem with quadratic constraints:

$$\min_{B} ||X - BS||_F^2$$
$$\text{s.t.} ||b_i||^2 \le c, i = 1,\ldots,m. \tag{7}$$

There are several methods can be used for solving this optimization problem, in this paper, we choose the more efficient Lagrange dual method to solve the optimization problem [15]. Due to the limitations of space, here we only give the optimal solution for B as follows:

$$B = XS^T \cdot (SS^T + \Lambda)^{-1} \tag{8}$$

where $\Lambda = diag(\vec{\lambda})$, $\vec{\lambda} = [\lambda_1, \ldots, \lambda_m]^T$, and each $\lambda_i \geq 0$ is a dual variable. We refer the reader to Ref. [15] for more details.

3.2.4. Computational complexity analysis

The computational complexity of MMRSC consists of three parts. First, it takes $O(\sum_{v=1}^{n_v} d^{(v)} \log(k) n \log(n))$ to build multiple k-nearest neighbor graphs[2] to encode the manifold information from n_v views, where $d^{(v)}$ is the dimensionality of the vth view. Second, MMRSC adopts an iterative way to learn B and S. Specifically, in each iteration, MMRSC involves computing a ℓ_1-regularized least squares problem (S), and a least squares problem with quadratic constraints (B). For learning S, MMRSC uses the state-of-the-art solution, i.e., the feature-sign search algorithm [15], and the computational cost is $O(nmK)$, where K is the number of non-zero entries of S. For learning B, the computational cost is $O(nmd)$. Thus the overall computational complexity of the second part is $O(t_1 m(K + d)n)$, where t_1 is the number of iterations and $d = \sum_{v=1}^{n_v} d^{(v)}$. At last, we adopt the k-means for clustering on the learnt sparse representation, and the cost is $O(t_2 cmn)$, where t_2 is the number of iterations in k-means, and c is the number of clusters. Putting everything together, the computational complexity of MMRSC[3] is

$$O\left(\sum_{v=1}^{n_v} d^{(v)} \log(k) n \log(n) + t_1 m(K + d)n + t_2 cmn\right)$$

Note that, the computational complexity of MMRSC is equivalent to that of GraphSC. Without increase the computational complexity, MMRSC can outperform GraphSC by capturing structure information over the view-level rather than over the mixed space as in GraphSC.

3.3. Image clustering

For clustering image data, similar to Ref. [7], we first utilize the proposed approach to project multi-view image data into a sparse and lower dimensional representation, then apply any clustering algorithm such as k-means to conduct clustering. It is worth noting that, in this work, we leverage the concatenated representations of all different views to construct X, which is simple but effective. Note that we can also employ other ways to construct X. As it is not the main concern of this work, we may explore alternatives as part of future work.

As the main concern of this work is to learn powerful image representations for image clustering, fair comparison of the performance of different methods can only be ensured when employing the same clustering method. Otherwise, differences in performance may be due to the different learnt representations or due to the clustering method, resulting in inconclusive experiments. For this reason, we employ k-means as the same clustering method, which is a simple, established and widely used clustering method. It is also interesting to employ additional clustering methods, such as k-means++ [27] and x-means [28]. As it is out of the scope of this paper, we will explore them in the future work.

It is worth noting that we can also learn a new representation with the same dimension as the number of ground-truth clusters where each dimension represents a cluster membership, and then select the maximal dimension as the final cluster label. Due to the limitation of space, we only report the results of applying k-means on the learnt representation since it achieves better performance in our experiments.

[2] The ball tree structure was utilized to construct the k-nearest neighbor graph.

[3] In practice, we could apply PCA to reduce the dimension of the data in order to accelerate the computation. For clarity, here we only report the general computational complexity.

4. Experiments

In this section, we empirically evaluate the proposed multi-view image clustering algorithm on two real-world image datasets. The experimental results demonstrate the effectiveness of our proposed algorithm.

4.1. Datasets

The datasets used in our experiments are the Handwritten Digits dataset and the MirFlickr dataset, where the former is homogeneous and the latter is heterogeneous. Here, the homogeneous and heterogeneous datasets reflect their degree of difference among different views. Homogeneity means that the views are from fewer different representation spaces, such as the Handwritten Digits dataset, where all views are visual features. In contrast, heterogeneity indicates that the difference among views are large, as is the case with the MirFlickr dataset, where views vary from textual descriptions to visual descriptions. Through these different settings, we would like to verify whether our approach works well in different scenarios. It is worth noting that, although our method can be naturally applied to more views, here we only consider two views since it is more straightforward to compare the performance of different approaches in the two scenarios. The statistics of the two datasets used in our experiments are summarized in Table 1.

Handwritten digits. This dataset is from the UCI repository[4], and consists of 2000 images of digits with 200 images for each of the ten digit classes. For this dataset, similar to Ref. [2], we use two types of feature sets (i.e., fourier and pixel) as the two views in our experiments. The fourier view consists of 76 fourier coefficients of the character shape, which are rotation invariant features (e.g., samples from class '6' and class '9' can not be distinguished based on these features). The pixel view consists of 240 features which were extracted by splitting the image of 30 $\times$ 48 pixels into 240 tiles of 2 $\times$ 3 windows.

MirFlickr. This dataset [29] comprises 25,000 images from the Flickr[5] and the associated tag annotations contributed by users. We clean the raw tag data by removing stop words, converting letters into lower case, and ignoring non-English tags. Moreover, we further disregard tags with a frequency less than 3 and images with less than 2 tags in order to reduce the noise. As in MirFlickr an image may belong to multiple categories, we select 10 categories, which are less correlated to each other (i.e., *tree, night, clouds, flower, food, dog, car, bird, baby, lake*), as ground-truth clusters and retain only images which belong to a single category, in order to facilitate a hard clustering evaluation. Finally, 7425 images are obtained and distributed in 10 categories. The tags are weighted by using the TF-IDF weighting scheme [30]. For the visual features in this dataset, we use Lire [31] to extract 305-D global features, including the 192-D Fuzzy Color and Texture Histogram [32], 33-D MPEG-7 Color layout [33], and 80-D MPEG-7 Edge Histogram [33]. Finally, we have two views of MirFlickr dataset, one is the 8,740 dimensional tag view and the other is the 305 dimensional visual view.

4.2. Baseline

To demonstrate the performance of our proposed model, we compare it with 7 baseline approaches. These approaches can be grouped into two categories: traditional single-view methods (Kmeans, NMF, SC, GraphSC) and state-of-the-art multi-view methods (CollNMF, MultiNMF, CoNMF). In order to make a fair comparison for traditional single-view approaches, in the multi-view

[4] https://archive.ics.uci.edu/ml/datasets.html.

[5] https://www.flickr.com/.

scenario, we concatenate different views together, and then apply Kmeans, NMF, SC, and GraphSC on the concatenated data representation. For clarity, we call Kmeans, NMF, SC, and GraphSC on the concatenated data representation as ConcatKmeans, ConcatNMF, ConcatSC, Concat GraphSC, respectively. The details of these baseline approaches are given as follows:

- ConcatKmeans: This method is a naive solution, which directly runs the k-means clustering algorithm over the concatenated data representation without conducting any factorization.
- ConcatNMF: This method applies non-negative matrix factorization (NMF) over the concatenated data representation in order to learn a new representation of the data. In contrast to k-means, this method first factorizes the original data representation into a basis matrix and a coefficient matrix, and then applies k-means over the learnt coefficient matrix for clustering.
- ConcatSC: This method employs sparse coding over the concatenated data representation as to learn a sparse representation, and then carries out k-means algorithm over this sparse representation for clustering. Compared to ConcatNMF, ConcatSC incorporates the sparsity constraint into the objective function to enforce the learnt coefficient matrix to be sparse.
- ConcatGraphSC [18]: This method is an extension of ConcatSC, which further incorporates the local manifold structure as an additional constraint in the objective function.
- CollNMF [1]: This method employs NMF for multi-view clustering by applying a joint matrix factorization process to learn the new presentation. In particular, it enforces different views to share the same coefficient matrices.
- MultiNMF [2]: This method is an extension of CollNMF, which relaxes the constraint of requiring the learnt coefficient matrices to be identical, by restricting the learnt coefficient matrices across different views towards a common consensus matrix. In MultiNMF, the regularization parameter is set to be 0.01, as suggested by the authors.
- CoNMF [3]: This method further relaxes the consensus constraint used in MultiNMF. Specifically, instead of enforcing different learned coefficient matrices towards a common consensus matrix, CoNMF employs co-regularization in the factorization process through two paradigms (i.e., pair-wise CoNMF and cluster-wise CoNMF). In the following experiments, we only report the results conducted by pair-wise CoNMF as it performs much better than cluster-wise CoNMF in all setting of our experiments. As mentioned before, the learned coefficient matrix of each view can be used for clustering and there are no guidelines for selecting the best performing coefficient matrix, we report the results of both the best and the worst performing coefficient matrices, which are referred to as CoNMF-B and CoNMF-W, respectively. The regularization parameters are set to be 1 as suggested by the authors.

For implementation, we first apply all methods except ConcatKmeans to learn a new representation with the same dimension (e.g., a 64-dimensional vector) for the data, and then apply k-means algorithm on the new representation for clustering. We carry out the experiments by conducting 20 test runs with different initializations [7]. To illustrate the improvements of MMRSC over all baseline methods, we also conducted the statistical significance test t-test at p-value < 0.05 (5% significance level). In MMRSC, the parameters β and k are empirically set as 0.1 and 3, respectively, and the parameters $\alpha_v (v = 1, \ldots, n_v)$ are uniformly set as 1. For simplicity, we use α instead of α_v $(v = 1, \ldots, n_v)$ for all views. In Section 4.6, we give a detailed discussion about the sensitivity of MMRSC with respect to these parameters.

4.3. Evaluation metrics

For evaluation, two standard clustering metrics, the accuracy (AC) and the normalized mutual information (NMI), are used to measure the performance.

Given an image x_i, let r_i and l_i be the predicted cluster and the ground-truth cluster provided by the dataset, respectively. The AC is defined as follows:

$$AC = \frac{\sum_{i=1}^{n} \delta(l_i, map(r_i))}{n} \tag{9}$$

where n indicates the total number of images and $\delta(x, y)$ is the delta function that equals 1 if $x = y$, and equals 0 otherwise, and $map(r_i)$ is the mapping function that maps each predicted cluster r_i to the best ground-truth cluster. The Kuhn–Munkres algorithm [34] is used to find the best mapping.

For two clusters C and C', their mutual information metric $MI(C, C')$ is defined as follows:

$$\mathrm{MI}(C, C') = \sum_{c_i \in C, c'_j \in C'} p(c_i, c'_j) \cdot \log_2 \frac{p(c_i, c'_j)}{p(c_i)p(c'_j)} \tag{10}$$

where $p(c_i)$ and $p(c'_j)$ are the probabilities that a sample arbitrarily selected from the dataset belongs to the clusters c_i as well as c'_j at the same time. We define the normalized mutual information NMI as follows:

$$\mathrm{NMI}(C, C') = \frac{\mathrm{MI}(C, C')}{\max(H(C), H(C'))} \tag{11}$$

where $H(C)$ and $H(C')$ are the entropies of C and C', respectively. NMI takes values between 0 and 1, with higher values indicating a closer match to the true clustering.

4.4. Single-view clustering

In this subsection, we compare the performance of three different single view clustering algorithms: k-means, NMF and MMRSC, which allows us to get a better understanding of properties of these methods on a single view. It is worth noting that, for single-view clustering, these NMF-based multi-view clustering approaches (e.g., CollNMF, MultiNMF, and CoNMF) will regress to NMF.

From Fig. 2(a), we observe that the views of Handwritten Digits are relatively homogeneous and of high quality. For example, the accuracies of conducting k-means on fourier view and pixel view are 0.728 and 0.786, respectively. And the pixel view is better than fourier view with an improvement of accuracy around 8%. The accuracy of NMF on each view of Handwritten Digits is worse than that of k-means. This is because that mapping the original representation into a latent low-rank space would cause the loss of information. In contrast, the accuracy of MMRSC on each view is comparable to that of k-means, and the reason is that the benefit of incorporating the sparsity and structure constraints can compensate for the loss of information during the factorization process.

On the MirFlickr dataset (see Fig. 2(b)), the views are heterogeneous and of low quality. For instance, the accuracies of k-means on tag view and visual view are only 0.244 and 0.295, respectively. The tag view is superior to the visual view with an improvement of accuracy up to 21%. Fig. 3 shows some sample images from the two datasets, and we observe that the images from the MirFlickr dataset are more complex than that of the Handwritten Digits dataset. For example, images from the same class (e.g., class 'bird') in the MirFlickr dataset often have different backgrounds while images from the same class (e.g., class '2') in the Handwritten Digits dataset share an identical background. Besides, the views of Handwritten Digits (e.g., fourier view and pixel view) are both

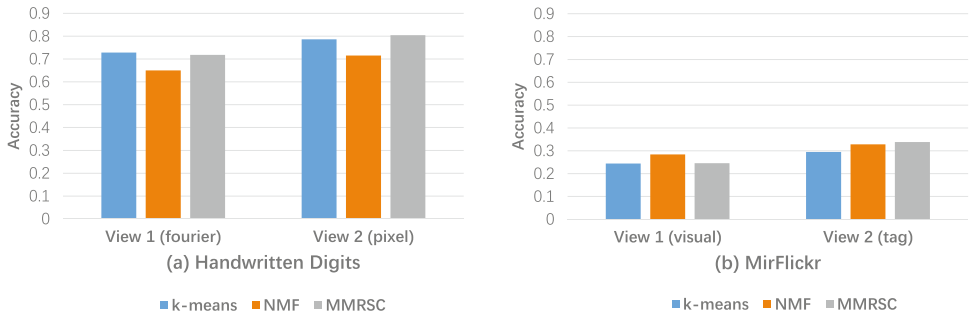

Fig. 2. Clustering performance of three methods (k-means, NMF and MMRSC) on each single view for two datasets.

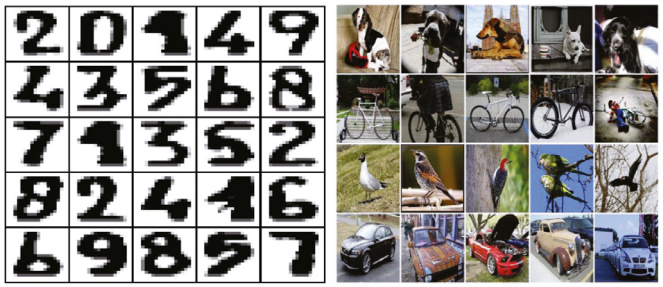

Fig. 3. Sample images from two datasets. (a) Handwritten digits and (b) MirFlickr.

Table 1
Statistics of the two datasets: Handwritten Digits and MirFlickr.

Dataset	#Image	View		#Cluster
		Name	#Feature	
Digits	2000	fourier	76	10
		pixel	240	
MirFlickr	7425	visual	305	10
		tag	8740	

extracted from visual contents, while the views of MirFlickr (e.g, visual view and tag view) are much more heterogeneous, which are extracted from visual and textual contents, respectively. On the MirFlickr dataset, the accuracy of NMF is better than that of k-means, because MirFlickr contains relatively more noise, and representing the original view by a compact representation learnt by NMF can help to alleviate the noise problem. Similar to the results on Handwritten Digits, MMRSC can achieve a significant better performance (paired t-test with p-value < 0.05) as compared to that of k-means on each view of the MirFlickr dataset.

In summary, on both datasets, MMRSC can achieve significant better or competitive performance as compare with k-means on both data set, while NMF demonstrates an unstable performance.

4.5. Multi-view clustering

Comparison on Handwritten Digit. Table 2 summarizes the clustering results of all methods on the Handwritten Digits dataset. As mentioned above, the top 4 baseline methods (i.e., ConcatK-means, ConcatNMF, ConcatSC, and ConcatGraphSC) in Table 2 are applied directly on a concatenated feature representation. It is interesting to see that the performance of ConcatNMF is even worse than that of ConcatKmeans which performs the k-means clustering algorithm directly on the concatenated representation without any factorization. This result shows that the representation learnt by applying ConcatNMF over the combined view cannot effectively deal with the multi-view clustering problem.

Table 2
Clustering performance (mean ± standard deviation) on the Handwritten Digits dataset. Performance metrics Accuracy and Normalized Mutual Information (NMI) are shown. Paired t-tests are performed and the symbol † indicates that MMRSC is significant better than the corresponding algorithm at p-value < 0.05. The best performance is indicated in bold.

Dataset	Handwritten Digits	
Method	Accuracy (%)	NMI (%)
ConcatKmeans	77.3 ± 5.2 †	74.4 ± 2.3 †
ConcatNMF	71.2 ± 6.5 †	68.8 ± 4.2 †
ConcatSC	81.8 ± 5.2 †	76.3 ± 2.9 †
ConcatGraphSC [18]	83.2 ± 6.5 †	77.4 ± 3.4 †
CollNMF [1]	70.1 ± 7.0 †	63.7 ± 4.1 †
MultiNMF [2]	87.5 ± 0.8 †	79.4 ± 0.7 †
CONMF-W [3]	84.2 ± 4.8 †	77.3 ± 2.6 †
CONMF-B [3]	84.3 ± 5.9 †	78.7 ± 3.1 †
MMRSC	**90.5 ± 4.9**	**86.3 ± 2.6**

Both ConcatSC and ConcatGraphSC are better than ConcatK-means, with an improvement of accuracy up to 5.8% and 7.6%, respectively. The reason is that SC enforces the learnt coefficient matrix to be sparse. This sparsity property makes the new representation more robust and has been shown to achieve encouraging performance in many applications, like image denoising [4]. ConcatGraphSC further improves the performance by incorporating the local manifold information of the concatenated representation. It is worth noting that the Handwritten Digits dataset is a homogeneous dataset, where the qualities of each view are nearly equivalent, therefore incorporating the manifold information of the concatenated representation can benefit the performance. As we will discuss later on the MirFlickr dataset, which is a heterogeneous dataset, this type of manifold constraint would hurt the performance if directly incorporated into the objective function of sparse coding.

It is surprising to see that the performance of CollNMF is even worse than that of ConcatKmeans. This is due to the fact that the constraint adopted by CollNMF is too strong, since it enforces a shared coefficient matrix across all views. MultiNMF achieves the best performance among all baseline methods, because it relaxes the constraint of CollNMF by enforcing the coefficient matrices of different views towards a common consensus matrix. This result shows that MultiNMF can have a promising performance when the dataset is homogeneous. CoNMF is another extension of CollNMF and it relaxes the constraint by employing the co-regularization in the factorization process. Although both CoNMF-B and CoNMF-W perform worse than MultiNMF, they still obtain the second best performance among all baseline approaches. Our proposed method, MMRSC, consistently significantly (paired t-test was conducted at p-value < 0.05) outperforms all the competing methods, with an improvement of accuracy up to 3.4% over the best performing baseline method MultiNMF.

Table 3
Clustering performance (mean ± standard deviation) on the MirFlickr dataset. Performance metrics Accuracy and Normalized Mutual Information (NMI) are shown. Paired *t*-tests are performed and the symbol † indicates that MMRSC is significant better than the corresponding algorithm at *p*-value < 0.05. The best performance is indicated in bold.

Dataset	MirFlickr	
Method	Accuracy (%)	NMI (%)
ConcatKmeans	28.5 ± 3.2 †	13.3 ± 4.8†
ConcatNMF	31.4 ± 3.7 †	16.4 ± 4.5†
ConcatSC	35.7 ± 2.5 †	22.2 ± 3.3†
ConcatGraphSC [18]	33.4 ± 2.5 †	18.7 ± 2.5 †
CollNMF [1]	31.5 ± 2.0 †	17.1 ± 2.1 †
MultiNMF [2]	24.0 ± 0.9 †	12.0 ± 2.3 †
CONMF-W [3]	21.0 ± 1.3 †	6.6 ± 0.6 †
CONMF-B [3]	36.6 ± 3.6	21.5 ± 3.1 †
MMRSC	**37.9 ± 1.9**	**23.2 ± 1.3**

This is because MMRSC introduces a set of high-level structure constraints, which can preserve the manifold structures of different views. Moreover, it employs sparse coding rather than NMF to learn a new representation, and can further take advantage of sparsity property to better deal with the noise issue. It suggests that employing the high-level structure constraints is more effective than imposing the consensus constraint directly over the low-level coefficient matrices as conducted in these NMF-based approaches.

Comparison on MirFlickr. As can be seen from Table 3, on the MirFlickr dataset, we find that the performance of ConcatNMF is better than that of ConcatKmeans. This shows that when the dataset is heterogeneous, directly applying the k-means clustering algorithm over a concatenated representation may not work effectively. Unsurprisingly, both ConcatSC and ConcatGraphSC are better than ConcatNMF and ConcatKmeans, due to the incorporation of the sparsity property. One interesting result is that ConcatGraphSC is worse than ConcatSC on MirFlickr, this is because the manifold structure based on the combined view is unreliable. Similar to the results on Handwritten Digits, the performance of CollNMF is comparable to that of ConcatNMF. This is also consistent with the analysis that CollNMF is equivalent to conducting NMF on a combined view [2]. The performance of MultiNMF is worse than ConcatKmeans because MultiNMF can perform well only when the dataset is homogeneous. Regarding the CoNMF method, it is interesting to see that the performance of CoNMF-W and CoNMF-B vary greatly. CoNMF-B outperforms all other baseline methods, reach an accuracy of 0.366, while CoNMF-W underperforms all other baseline methods, with an accuracy of 0.21. As we mentioned before, the drawback of CoNMF is that it is impractical to select the best performing coefficient matrix, thus limits its application. MMRSC significantly outperforms CoNMF-B for the NMI metric (paired *t*-test, *p*-value < 0.05), and also has a better performance than CoNMF-B for the Accuracy metric with a *p*-value = 0.08. It shows that on the heterogeneous dataset Mirflickr, MMRSC can still achieve a better performance.

It is worth noting that the clustering performance of the proposed method over the MirFlickr dataset is not very high, as it is a quite challenging task due to the heterogeneous multi-modalities, noise, and issues with incompleteness. The motivation to utilize this dataset are two-fold: (1) we want to verify the performance of our algorithm on some very challenging clustering tasks; (2) automatically organizing social media data, like the MirFlickr dataset, is critical for helping users to easily understand the complex data pattern and make decisions.

4.6. Impact of parameters

In this section, we conduct experiments to analyze the sensitivity of our methods with respect to the parameters: α, β, and k.

4.6.1. Parameter α

The parameter α controls the influence of the multiple graph Laplacian constraints. We vary α as {0, 0.01, 0.05, 0.1, 0.2, 0.5, 1, 2, 5, 10, 20, 50}, while fixing the other parameters to the following values: $\beta = 0.1$ and $k = 3$. Fig. 4(a) shows the impact of α to MMRSC on the Handwritten Digits dataset. We can observe that the performance increases as we increase α, and reaches the best performance when α varies from 0.05 to 1, which is followed by a considerable drop of performance. This is because when $\alpha = 0$, MMRSC regresses to the standard sparse coding model *SC* where no manifold constraint is involved. When we increase α, the multiple manifold structures from different views are preserved. The results show that these manifold structures are beneficial for MMRSC when they are incorporated as a set of graph Laplacian regularizers in the objective function. When α is too large, like $\alpha > 5$, the performance will drop significantly. Fig. 4(d) demonstrates the impact of α to MMRSC on the MirFlickr dataset, and similar results are observed.

4.6.2. Parameter β

The parameter β reflects the influence of the sparsity constraint. We vary β as {0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0, 2.0, 5.0} while fixing other parameters (i.e., $\alpha = 1$ and $k = 3$). The results on Handwritten Digits are shown in Fig. 4(b), and we observe an increase of performance when introducing the sparsity constraint (e.g., $\beta > 0$). Moreover, the performance of MMRSC is not very sensitive to β when we vary β from 0.1 to 2, and it tends to drop if we further increase the value of β (e.g., when $\beta = 5$). While on the MirFlickr dataset as shown in Fig. 4(e), we observe that MMRSC is more sensitive to β and demonstrates an early drop in performance as compared to the results on Handwritten Digits. In particular, the performance of MMRSC continues to rise and reaches a peak when $\beta = 0.3$. Then it begins to drop gradually with the increase of β, and encounters a dramatically decrease after $\beta > 0.9$. This is because the dimension of MirFlickr is much higher than that of Handwritten Digits, and a strong sparsity constraint (i.e., a large value of β) will limit the complexity of MMRSC to model this high-dimensional data representation. In general, incorporating sparsity constraint can improve the performance, and keeping a relatively small β (e.g., $0.1 \leq \beta \leq 0.3$) is necessary to ensure a better trade-off between sparsity and model complexity.

4.6.3. Parameter k

The parameter k is the number of nearest neighbors for the construction of manifolds, which are then incorporated as a set of graph Laplacian constraints in the objective function (see Eq. (2)) of MMRSC. We vary k as {1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 20, 50} while keeping other parameters fixed: $\alpha = 1$, and $\beta = 0.1$. The results on Handwritten Digits and MirFlickr are shown in Fig. 4(c) and (f), respectively. On the Handwritten Digits dataset, the performance of MMRSC first increases quickly until $k = 3$ and then becomes stable. After k exceeds 10, it starts to decrease. The reason is straightforward: when $k > 10$, the possibility that noisy relationships are added into the manifold, which will potentially affect the performance of our method.

The performance on the MirFlickr dataset demonstrates similar trend, with the exception that there is just a slight increase when k varies from 1 to 3. The reason is that the MirFlickr dataset are collected from various scenarios (see Fig. 3(b)) and are supposed to be more diverse as compared to the Handwritten Digits dataset, which would lead to more noisy manifolds (e.g., the manifold from visual view) when k is small and degrade the effect of the manifold constraint.

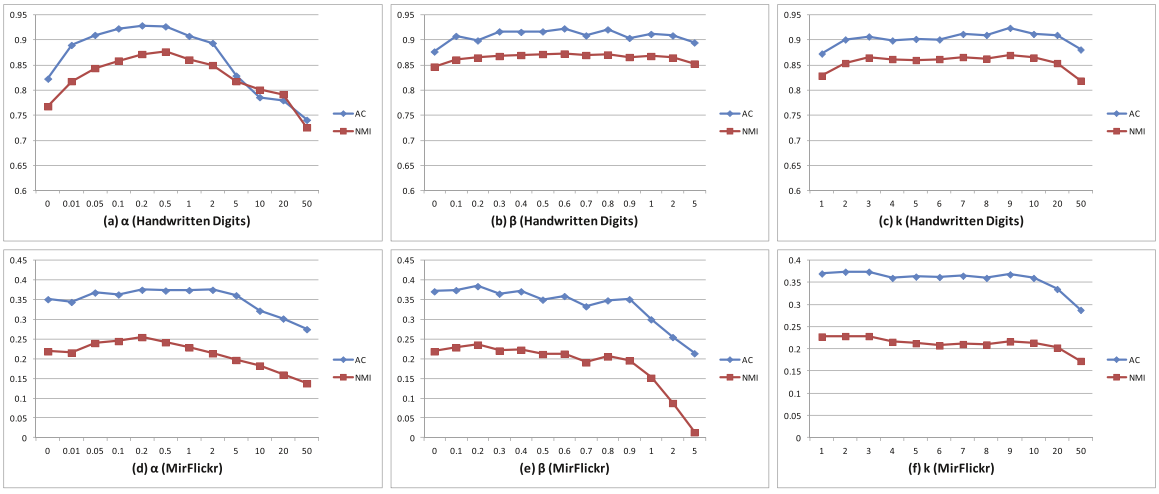

Fig. 4. The impact of parameters (α, β, and k) to MMRSC on Handwritten Digits (above) and MirFlickr (below).

5. Conclusions

In this paper, we have studied the problem of multi-view image clustering. We propose MMRSC, a novel framework that incorporates the high-level manifold consensus constraint in order to capture the underlying clustering structures of different views. In addition, we also take the sparsity issue into account and resort to exploiting the sparse coding framework, instead of utilizing the NMF framework, to deal with the multi-view clustering problem. Experimental results show that the proposed approach consistently outperforms the baseline methods in terms of both accuracy and normalized mutual information for the multi-view image clustering task. For future work, we intend to apply our MMRSC model in multi-view classification, where the label information could be introduced as an additional regularization term to generate proper representation. For example, we can modify our MMRSC model by enforcing a class consensus assumption, i.e., for the learnt representation, the distance of data points from the same class should be more closer than the distance of data points from different classes.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

CRediT authorship contribution statement

Xiaofei Zhu: Conceptualization, Formal analysis, Writing - original draft, Supervision. **Jiafeng Guo:** Writing - review & editing, Formal analysis, Supervision. **Wolfgang Nejdl:** Writing - review & editing, Formal analysis, Supervision. **Xiangwen Liao:** Writing - review & editing, Formal analysis, Supervision. **Stefan Dietze:** Writing - review & editing, Formal analysis, Supervision.

Acknowledgments

The work was partially supported by the National Natural Science Foundation of China (No. 61722211, 61976054), the Federal Ministry of Education and Research (No. 01LE1806A), and the Chongqing Research Program of Basic Research and Frontier Technology (No. cstc2017jcyjBX0059, cstc2017jcyjAX0339).

References

[1] Z. Akata, C. Thurau, C. Bauckhagee, Non-negative matrix factorization in multimodality data for segmentation and label prediction, in: Proceedings of the Sixteenth Computer Vision Winter Workshop, 2011.
[2] J. Liu, C. Wang, J. Gao, J. Han, Multi-view clustering via joint nonnegative matrix factorization, in: Proceedings of theSIAM International Conference on Data Mining (SDM), 2013, pp. 252–260.
[3] X. He, M.-Y. Kan, P. Xie, X. Chen, Comment-based multi-view clustering of web 2.0 items, in: Proceedings of the Twenty-third International Conference on World Wide Web (WWW), 2014, pp. 771–782.
[4] M. Elad, M. Aharon, Image denoising via sparse and redundant representations over learned dictionaries, Trans. Image Proc. 15 (12) (2006) 3736–3745.
[5] J. Zhao, X. Xie, X. Xu, S. Sun, Multi-view learning overview: recent progress and new challenges, Inf. Fusion 38 (2017) 43–54.
[6] L. Feng, L. Cai, Y. Liu, S. Liu, Multi-view spectral clustering via robust local subspace learning, Soft Comput. 21 (8) (2017) 1937–1948.
[7] K. Chaudhuri, S.M. Kakade, K. Livescu, K. Sridharan, Multi-view clustering via canonical correlation analysis, in: Proceedings of the ICML, 2009, pp. 129–136. New York, NY, USA
[8] A. Kumar, P. Rai, H. Daumé III, Co-regularized multi-view spectral clustering, in: Proceedings of the NIPS, 2011. Granada, Spain
[9] D. Ramage, P. Heymann, C.D. Manning, H. Garcia-Molina, Clustering the tagged web, in: Proceedings of the Second ACM International Conference on Web Search and Data Mining, in: WSDM '09, ACM, New York, NY, USA, 2009, pp. 54–63.
[10] E. Bruno, S. Marchand-Maillet, Multiview clustering: a late fusion approach using latent models, in: Proceedings of the Thirty-second International ACM SIGIR Conference on Research and Development in Information Retrieval, in: SIGIR '09, ACM, New York, NY, USA, 2009, pp. 736–737.
[11] D. Greene, P. Cunningham, A matrix factorization approach for integrating multiple data views, in: Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases: Part I, in: ECML PKDD '09, Springer-Verlag, Berlin, Heidelberg, 2009, pp. 423–438.
[12] M. Xu, Z. Zhu, X. Zhang, Y. Zhao, X. Li, Canonical correlation analysis with $l_{2,1}$-norm for multiview data representation, IEEE Trans. Cybern. 99 (2019) 1–11.
[13] S. Gao, Z. Yu, T. Jin, M. Yin, Multi-view low-rank matrix factorization using multiple manifold regularization, Neurocomputing 335 (2019) 143–152.
[14] I.E. Ohiorhenuan, F. Mechler, K.P. Purpura, A.M. Schmid, Q. Hu, J.D. Victor, Sparse coding and high-order correlations in fine-scale cortical networks, Nature 466 (2010) 617–621.
[15] H. Lee, A. Battle, R. Raina, A.Y. Ng, Efficient sparse coding algorithms, in: Proceedings of the Conference on Neural Information Processing Systems (NIPS), 2007, pp. 801–808.
[16] J. Sun, Q. Zhuo, C. Ma, W. Wang, Sparse image coding with clustering property and its application to face recognition, Pattern Recogn. 34 (9) (2001) 1883–1884.
[17] J. Wright, A.Y. Yang, A. Ganesh, S.S. Sastry, Y. Ma, Robust face recognition via sparse representation, IEEE Trans. Pattern Anal. Mach. Intell. 31 (2) (2009) 210–227.
[18] M. Zheng, J. Bu, C. Chen, C. Wang, L. Zhang, G. Qiu, D. Cai, Graph regularized sparse coding for image representation, IEEE Trans. Image Process. 20 (5) (2011) 1327–1336.

[19] J. Yang, K. Yu, Y. Gong, T.S. Huang, Linear spatial pyramid matching using sparse coding for image classification, in: Proceedings of the CVPR, 2009, pp. 1794–1801.
[20] Y. Liu, F. Wu, Z. Zhang, Y. Zhuang, S. Yan, Sparse representation using nonnegative curds and whey, in: Proceedings of the CVPR, 2010, pp. 3578–3585.
[21] K. Kavukcuoglu, M. Ranzato, R. Fergus, Y. LeCun, Learning invariant features through topographic filter maps, in: Proceedings of the CVPR, 2009, pp. 1605–1612.
[22] M. Long, G. Ding, J.W. 0001, J. Sun, Y. Guo, P.S. Yu, Transfer sparse coding for robust image representation, in: Proceedings of the CVPR, 2013, pp. 407–414.
[23] J.J. Wang, H. Bensmail, N. Yao, X. Gao, Discriminative sparse coding on multi–manifolds, Knowl.-Based Syst. 54 (2013) 199–206.
[24] B.A. Olshausen, D.J. Field, Emergence of simple-cell receptive field properties by learning a sparse code for natural images, Nature 381 (1996) 607–609.
[25] G. Andrew, J. Gao, Scalable training of l1-regularized log-linear models, in: Proceedings of the Twenty-forth International Conference on Machine Learning, 2007, pp. 33–40.
[26] M. Schmidt, G. Fung, R. Rosales, Fast optimization methods for l1 regularization: a comparative study and two new approaches, in: Proceedings of the Eighteenth European Conference on Machine Learning, in: ECML '07, Springer-Verlag, Berlin, Heidelberg, 2007, pp. 286–297.
[27] D. Arthur, S. Vassilvitskii, K-means++: the advantages of careful seeding, in: Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms, 2007.
[28] D. Pelleg, A.W. Moore, X-means: extending k-means with efficient estimation of the number of clusters, in: Proceedings of the Seventeenth International Conference on Machine Learning (ICML), 2000, pp. 727–734.
[29] M.J. Huiskes, M.S. Lew, The MIR Flickr retrieval evaluation, in: Proceedings of the MIR, ACM, New York, NY, USA, 2008, pp. 39–43.
[30] R.A. Baeza-Yates, B.A. Ribeiro-Neto, Modern Information Retrieval – The Concepts and Technology Behind Search, Second edition, Pearson Education Ltd., Harlow, England, 2011.
[31] M. Lux, S.A. Chatzichristofis, Lire: lucene image retrieval: an extensible java cbir library, in: Proceedings of the Sixteenth ACM international conference on Multimedia, in: MM '08, ACM, New York, NY, USA, 2008, pp. 1085–1088.
[32] S.A. Chatzichristofis, Y.S. Boutalis, Fcth: fuzzy color and texture histogram – a low level feature for accurate image retrieval, in: Proceedings of the 2008 Ninth International Workshop on Image Analysis for Multimedia Interactive Services, 2008, pp. 191–196. Washington, DC, USA
[33] S.F. Chang, T. Sikora, A. Puri, Overview of the MPEG-7 standard, IEEE Trans. Circuits Syst. Video Technol. 11 (6) (2001) 688–695.
[34] L. Lovász, M. Plummer, Matching Theory, Amsterdam, The Netherlands: North-Holland, 1986.

Xiaofei Zhu received his Ph.D. degree at the Institute of Computing Technology, Chinese Academy of Science (ICT-CAS) in 2012. Then he spent four years as a Postdoctoral Research Fellow at the L3S Research Center, Leibniz University Hannover. Currently, he is a full professor at Chongqing University of Technology. His research interests include web search, data mining and machine learning, and he has published more than 30 papers in international conferences and journals, including the top conferences like SIGIR, WWW, CIKM, TKDE, etc. He has won the Best Paper Awards of CIKM (2011). He serves on the program committees and editorial board of several international conferences and journals, including SIGIR, AAAI, CIKM, etc.

Jiafeng Guo is currently a Professor in Institute of Computing Technology, Chinese Academy of Sciences, and University of Chinese Academy of Sciences. He received his Ph.D. in Computer Software and Theory from the University of Chinese Academy of Sciences, Beijing, China, in 2009. He has worked on a number of topics related to web search and data mining, including query representation and understanding, learning to rank, and text modeling. His current research focuses on representation learning and neural models for information retrieval and filtering. He has published more than 80 papers in several top conferences/journals such as SIGIR, WWW, CIKM, IJCAI, and TKDE. His work on information retrieval has received the Best Paper Award in ACM CIKM (2011), Best Student Paper Award in ACM SIGIR (2012) and Best Full Paper Runner-up Award in ACM CIKM (2017). Moreover, he has served as the PC member for the prestigious conferences including SIGIR, WWW, KDD, WSDM, and ACL, and the associate editor of TOIS.

Wolfgang Nejdl has been full professor of computer science at the University of Hannover since 1995. He received his M.Sc. (1984) and Ph.D. degree (1988) at the Technical University of Vienna, and has been associate professor at the RWTH Aachen from 1992 to 1995. He worked as visiting researcher / professor at Xerox PARC, Stanford University, UIUC, EPFL Lausanne, and at PUC Rio. Prof. Nejdl heads the L3S Research Center (http://www.L3S.de/) as well as the Distributed Systems Institute/Knowledge Based Systems, and does research in the areas of search and information retrieval, information systems, semantic web technologies, peer-to-peer infrastructures, databases, technology-enhanced learning and artificial intelligence. Wolfgang Nejdl published more than 230 scientific articles, as listed at DBLP, and has been program chair, program committee and editorial board member of numerous international conferences and journals, see also http://www.kbs.uni-hannover.de/nejdl/.

Xiangwen Liao is currently a professor at Fuzhou University. He received his Ph.D. degree at the Institute of Computer Sciences, Chinese Academy of Sciences (ICT-CAS) in 2009. His research interests include natural language processing, information retrieval and extraction, data mining, and social media analysis. He has published more than 30 papers in international conferences and journals, including FCS, Physica A, NLPCC, IJCAI, etc. He serves on the program committees and editorial board of several international conferences and journals, including ACL, EMNLP, NLPCC etc.

Stefan Dietze is a professor at the University of Dõsseldorf and the scientific director of Knowledge Technologies for the Social Sciences (WTS), GESIS Leibniz Institute for the Social Sciences. He received his Ph.D. from the Institute for Computer Science of the University of Potsdam (Ph.D./Dr. rer. nat. in Applied Computer Science), Germany, in 2004. His professional career led him to the Fraunhofer Institute for Software and Systems Engineering (ISST) in Berlin. Then, he spent five years as a Postdoctoral Research Fellow at the Knowledge Media Institute (KMi) of The Open University in Milton Keynes, UK. Before joining GESIS, he led a research group at the L3S research center at Leibniz University Hannover. His research interests include semantic web technologies, semantic search, and information retrieval. He has published numerous papers in prestigious journals and international conferences, including SIGIR, WWW, CIKM, ESWC, CHIIR and so on. He has been general/poster/track chair, program committee and editorial board member of numerous international conferences and journals.