Incomplete Multi-Modal Visual Data Grouping

Handong Zhao[†], Hongfu Liu[†] and Yun Fu[†‡]
[†]Department of Electrical and Computer Engineering, Northeastern University, Boston, USA, 02115
[‡]College of Computer and Information Science, Northeastern University, Boston, USA, 02115
{hdzhao, hfliu, yunfu}@ece.neu.edu

Abstract

Nowadays multi-modal visual data are much easier to access as the technology develops. Nevertheless, there is an underlying problem hidden behind the emerging multi-modality techniques: What if one/more modal data fail? Motivated by this question, we propose an unsupervised method which well handles the incomplete multi-modal data by transforming the original and incomplete data to a new and complete representation in a latent space. Different from the existing efforts that simply project data from each modality into a common subspace, a novel graph Laplacian term with a good probabilistic interpretation is proposed to couple the incomplete multi-modal samples. In such a way, a compact global structure over the entire heterogeneous data is well preserved, leading to a strong grouping discriminability. As a non-trivial contribution, we provide the optimization solution to the proposed model. In experiments, we extensively test our method and competitors on one synthetic data, two RGB-D video datasets and two image datasets. The superior results validate the benefits of the proposed method, especially when multi-modal data suffer from large incompleteness.

1 Introduction

In recent years, a large volume of techniques emerge in artificial intelligence field thanks to the easy accessibility of multi-modal data captured from multiple sensors [Cai *et al.*, 2013; Zhao and Fu, 2015; Zhang *et al.*, 2015; Liu *et al.*, 2016]. Working in an unsupervised manner, multi-modal grouping (or clustering) offers a general view of the heterogeneous data grouping structure, which has been drawing extensive attention [Bickel and Scheffer, 2004; Ding and Fu, 2014; Blaschko and Lampert, 2008; Chaudhuri *et al.*, 2009; Fred and Jain, 2005; Singh and Gordon, 2008; Cao *et al.*, 2015]. While beneath the prosperous studies of the multi-modal data grouping problem, there is an underlying problem, i.e., when the data from one modality/more modalities are inaccessible because of sensor failure or other reasons, most methods mentioned above would inevitably degenerate or even fail.

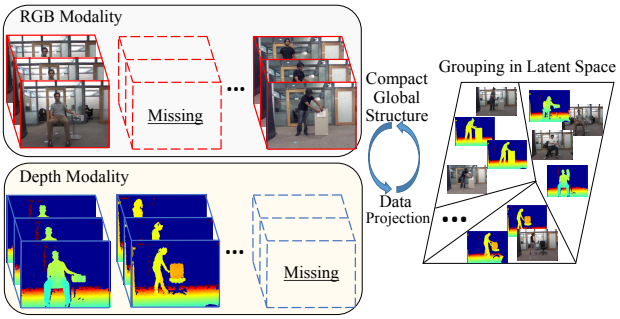


Figure 1: Framework of the proposed method. Take RGB-D video sequence as an example, to solve the IMG problem, we project the incomplete RGB-D data into a latent space as well as preserve the compact global structure simultaneously.

In this paper, we focus on this challenging, i.e., Incomplete multi-modality Grouping (IMG) problem.

To solve IMG problem, a natural thought is to reuse the existing techniques by remedying the incomplete data. In [Li *et al.*, 2014], two strategies are applied to facilitate to fit IMG problem, i.e., remove samples suffering from missing information or preprocess the incomplete samples to fill in the missing information. Obviously, the first strategy changes the number of samples, which essentially disobeys the goal of the original problem. The second strategy has been experimentally tested to be not good enough [Shao *et al.*, 2015].

Most recently, there are few attempts proposed to solve IMG problem. [Li *et al.*, 2014] proposed a pioneer work to handle two-modal incomplete data case, by projecting the partial data into a common latent subspace via nonnegative matrix factorization (NMF) and ℓ_1 sparse regularizer. Following this line, a similar idea of weighted NMF and $\ell_{2,1}$ regularizer was proposed in [Shao *et al.*, 2015]. However, both methods [Li *et al.*, 2014; Shao *et al.*, 2015] overlook the global structure over the entire data samples.

Inspired by this, we propose a novel method integrating the latent subspace generation and the compact global structure into a unified framework as shown in Figure 1. More specifically, a novel graph Laplacian term coupling the complete visual data samples is introduced in latent space, where the similar samples are more likely to be grouped together.

Compared with the existing approaches, the contributions of our method are three folds:

- We propose a novel method to deal with IMG problem for visual data with the consideration of the compact global structure in the low-dimensional latent space. The practice is achieved through a Laplacian graph on complete data instances, bridging the complete-modal samples and partial-modal samples.
- Nontrivially, we provide the optimization solution to our proposed objective, where three auxiliary variables are introduced to make the optimization of the proposed graph Laplacian term happen under the incomplete multi-modality setting.
- The superior results on six datasets, i.e., one synthetic and four visual datasets, validate the effectiveness of the proposed method. Specifically, when data suffer from large incompleteness, we raise the NMI performance bar by more than 30% and 10% for the synthetic and real-world visual data, respectively.

2 Method

We start with the introduction of some basic operator notations used in this paper. $\mathrm{tr}(\cdot)$ is the operator to calculate the trace of matrix. $\langle A, B\rangle$ is the inner product of two matrixes calculated as $\mathrm{tr}(A^{\mathrm{T}}B)$. $\|\cdot\|_{\mathrm{F}}^2$ denotes the Frobenius norm. Operator $(A)_+$ works as $\max(0, A)$ to make the matrix (or vector) non-negative. Other variable and parameter notations are introduced later in the manuscript.

2.1 Incomplete multi-modality Grouping

For the ease of discussion, we use two-modal case for illustration. Given a set of data samples $X = [x_1, \ldots, x_i, \ldots, x_N],\ i = 1, \ldots, N$, where N is the total number of samples. Each sample has two modalities, i.e., $x_i = [x_i^{(1)},\ x_i^{(2)}]$. For IMG problem, we follow the setting in [Li *et al.*, 2014], the input data is separated as an incomplete modal sample set $\hat{X} = \{\hat{X}^{(1,2)}, \hat{X}^{(1)}, \hat{X}^{(2)}\}$ instead of the complete multi-modal data X, where $\hat{X}^{(1,2)}$, $\hat{X}^{(1)}$, and $\hat{X}^{(2)}$ denote the data samples presented in both modalities, modal-1 and modal-2, respectively. The feature dimensions of modal-1 and modal-2 data are d_1 and d_2, and the numbers of shared samples and unique samples in modal-1 and modal-2 are c, m and n, respectively. Accordingly, we have $\hat{X}^{(1,2)} \in \mathbb{R}^{c\times(d_1+d_2)}$, $\hat{X}^{(1)} \in \mathbb{R}^{m\times d_1}$, $\hat{X}^{(2)} \in \mathbb{R}^{n\times d_2}$, $N = c + m + n$. Same as traditional multi-view clustering, the goal of IMG is to group the samples into their corresponding clusters.

Previous methods, like MultiNMF [Liu *et al.*, 2013] and PVC [Li *et al.*, 2014], pursuit a common latent space via nonnegative matrix factorization (NMF) where samples from different views can be well grouped. In this work, we follow this line to find a latent common subspace for heterogenous multi-modal visual data. However differently, we get rid of the non-negative constraint to make the optimization much easier. Besides, the major contribution of this paper is to demonstrate the correctness and effectiveness of the proposed global constraint in IMG problem.

Given the latent dimension of projective subspace k, we denote $P_c^{(1)} \in \mathbb{R}^{c\times k}$ and $P_c^{(2)} \in \mathbb{R}^{c\times k}$ as the latent representations of $\hat{X}^{(1,2)} = [X_c^{(1)}; X_c^{(2)}]$ from two different modalities. Note that $X_c^{(1)}$ and $X_c^{(2)}$ are the samples existing in both modalities, thus $P_c^{(1)}$ and $P_c^{(2)}$ are expected to be close, i.e., $P_c^{(1)} \to P_c \leftarrow P_c^{(2)}$. Consequently, we have the basic incomplete multi-modality grouping formulation as

$$\min_{\substack{P_c,\hat{P}^{(1)},\hat{P}^{(2)}\\ U^{(1)},U^{(2)}}} \left\| \begin{bmatrix} X_c^{(1)} \\ \hat{X}^{(1)} \end{bmatrix} - \begin{bmatrix} P_c \\ \hat{P}^{(1)} \end{bmatrix} U^{(1)} \right\|_{\mathrm{F}}^2 + \left\| \begin{bmatrix} X_c^{(2)} \\ \hat{X}^{(2)} \end{bmatrix} - \begin{bmatrix} P_c \\ \hat{P}^{(2)} \end{bmatrix} U^{(2)} \right\|_{\mathrm{F}}^2 + \lambda(\|U^{(1)}\|_{\mathrm{F}}^2 + \|U^{(2)}\|_{\mathrm{F}}^2), \tag{1}$$

where λ is a trade-off parameter, P_c is the shared latent representation from $X_c^{(1)}$ and $X_c^{(2)}$, $U^{(1)} \in \mathbb{R}^{k\times d_1}$, $U^{(2)} \in \mathbb{R}^{k\times d_2}$ are known as the bases in matrix decomposition, $\hat{P}^{(1)} \in \mathbb{R}^{m\times k}$, $\hat{P}^{(2)} \in \mathbb{R}^{n\times k}$ are the latent low-dimensional coefficients for missing modal samples corresponding to $\hat{X}^{(1)}$ and $\hat{X}^{(2)}$. Regularizers $\|U^{(1)}\|_{\mathrm{F}}^2$ and $\|U^{(2)}\|_{\mathrm{F}}^2$ are used to prevent the trivial solution.

2.2 Complete Graph Laplacian

By concatenating the projective coefficients in latent subspace $P = [P_c; \hat{P}^{(1)}; \hat{P}^{(2)}] \in \mathbb{R}^{N\times k}$, we can directly apply clustering method on P to get the cluttering result. However, it is deserved to know that the learned coefficients P is without global property which is crucial in subspace clustering. For the traditional multi-modality learning problem, global constraints are easy to be incorporated because of the complete modality setting, such as low-rank constraint in [Ding and Fu, 2014]. While in IMG problem, this cannot be easily achieved. To tackle this, we propose to learn a Laplacian graph incorporating all the samples in the latent space.

Integrating the idea of graph Laplacian and Eq. (1) into the unified objective function, we have our formulation with the complete graph Laplacian term $\mathcal{G}$ as

$$\begin{aligned}\min_{\substack{P_c,\hat{P}^{(1)},\hat{P}^{(2)}\\ U^{(1)},U^{(2)},A}} & \left\| \begin{bmatrix} X_c^{(1)} \\ \hat{X}^{(1)} \end{bmatrix} - \begin{bmatrix} P_c \\ \hat{P}^{(1)} \end{bmatrix} U^{(1)} \right\|_{\mathrm{F}}^2 + \left\| \begin{bmatrix} X_c^{(2)} \\ \hat{X}^{(2)} \end{bmatrix} - \begin{bmatrix} P_c \\ \hat{P}^{(2)} \end{bmatrix} U^{(2)} \right\|_{\mathrm{F}}^2 + \mathcal{G}(P, A) + \mathcal{R}(U, A). \\ \text{s.t. } & \forall i\ A_i^T\mathbf{1} = 1,\ A_i \succeq 0.\end{aligned} \tag{2}$$

Here,

$$\mathcal{G}(P, A) = \beta\mathrm{tr}(P^{\mathrm{T}}L_A P), \tag{3}$$

$$\mathcal{R}(U, A) = \lambda(\|U^{(1)}\|_{\mathrm{F}}^2 + \|U^{(2)}\|_{\mathrm{F}}^2) + \gamma\|A\|_{\mathrm{F}}^2, \tag{4}$$

where $L_A \in \mathbb{R}^{N\times N}$ is the Laplacian matrix of similarity matrix $A \in \mathbb{R}^{N\times N}$, defined by $L_A = D - A$, in which the degree matrix is the diagonal matrix with $D_{ii} = \sum_{j=1}^{N} A_{ij}$. Several remarks are made here.

Remark 1: Thanks to the graph Laplacian term L_A, we bridge the sample connection between complete modal samples and partial modal samples. In such a way, the global constraint on the complete set of data samples is integrated into the objective, which in turn influences the projected coefficients in low-dimensional space with the global structure. The practice of adding a graph term on the complete set of data gives the name of "*complete graph Laplacian*".

Remark 2: A is the affinity graph, with each element denoting the similarity between two data samples in latent subspace. We normalize each column as the summation equals to 1 as well as all the elements are nonnegative, making A have a probability interpretation. This naturally provides us with an opportunity to do spectral clustering on the optimized A, which none of the existing partial multi-view methods have.

Remark 3: As shown in Eq. (4), the regularizers we add are all Frobenius norms for simplicity. According to [Lu *et al.*, 2012], other regularizers such as ℓ_1-norm or trace (nuclear) norm are also good choices for preserving the global structure that benefits clustering performance.

2.3 Optimization

As seen in Eq.(2), in order to learn a meaningful affinity matrix in a unified framework, our proposed objective includes several matrix factorization terms, regularizers and constraints. It is obviously not jointly convex w.r.t. all the variables. Instead, we plan to update each variable at a time via augmented Lagrange Multiplier (ALM) with alternating direction minimizing strategy [Lin *et al.*, 2011].

However, it is noted that P_c, $\hat{P}^{(1)}$ and $\hat{P}^{(2)}$ are difficult to be optimized because of the following reasons: (1) Laplacian graph L_A is the graph measuring the affinities among all sample points, that is, we have to update them as a whole; (2) There is no way to directly combine P_c, $\hat{P}^{(1)}$ and $\hat{P}^{(2)}$ for optimization since $\hat{P}^{(1)}$ and $\hat{P}^{(2)}$ do not share the same basis, and even the size of input data in modal-1 is not the same as that in modal-2 ($[X_c^{(1)}; \hat{X}^{(1)}] \in \mathbb{R}^{(c+m)\times d_1}$ for modal-1, $[X_c^{(2)}; \hat{X}^{(2)}] \in \mathbb{R}^{(c+n)\times d_2}$ for modal-2). This dilemma makes the variables P_c, $\hat{P}^{(1)}$ and $\hat{P}^{(2)}$ impossible to be optimized individually nor together as P. To solve this challenge, we propose to introduce three auxiliary variables $Q_c \in \mathbb{R}^{c\times k}$, $\hat{Q}^{(1)} \in \mathbb{R}^{m\times k}$ and $\hat{Q}^{(2)} \in \mathbb{R}^{n\times k}$ for P_c, $\hat{P}^{(1)}$ and $\hat{P}^{(2)}$ respectively. In this way, we separately update the affinity matrix A (Laplacian L_A) and matrix factorization with the bridges of $P_c = Q_c$, $\hat{P}^{(1)} = \hat{Q}^{(1)}$ and $\hat{P}^{(2)} = \hat{Q}^{(2)}$.

Correspondingly, the augmented Lagrangian function of Eq. (2) with three auxiliary variables is written as

$$\begin{aligned}\mathcal{C}_{(\forall i\ A_i^T\mathbf{1}=1; A_i\geq 0)} = &\left\|\begin{bmatrix} X_c^{(1)} \\ \hat{X}^{(1)} \end{bmatrix} - \begin{bmatrix} P_c \\ \hat{P}^{(1)} \end{bmatrix} U^{(1)}\right\|_{\mathrm{F}}^2 + \\ &\left\|\begin{bmatrix} X_c^{(2)} \\ \hat{X}^{(2)} \end{bmatrix} - \begin{bmatrix} P_c \\ \hat{P}^{(2)} \end{bmatrix} U^{(2)}\right\|_{\mathrm{F}}^2 + \lambda(\|U^{(1)}\|_{\mathrm{F}}^2 + \|U^{(2)}\|_{\mathrm{F}}^2) \\ &+ \beta\mathrm{tr}(Q^{\mathrm{T}}L_AQ) + \gamma\|A\|_{\mathrm{F}}^2 + \langle Y, P-Q\rangle + \frac{\mu}{2}\|P-Q\|_{\mathrm{F}}^2,\end{aligned} \tag{5}$$

where $Y = [Y_c; \hat{Y}^{(1)}; \hat{Y}^{(2)}]$ is the lagrangian multiplier, and $\mu > 0$ is a penalty parameter. Specifically, variables P, Q, $U^{(1)}$, $U^{(2)}$, A in the $\tau+1$ iteration are updated as follows:

Update $\mathbf{U^{(1)}}$&$\mathbf{U^{(2)}}$. Fixing P, Q and A, the Lagrangian function w.r.t. $U_{(\tau+1)}^{(1)}$ is written as:

$$\mathcal{C}(U^{(1)}) = \|X_c^{(1)} - P_cU^{(1)}\|_{\mathrm{F}}^2 + \lambda\|U^{(1)}\|_{\mathrm{F}}^2. \tag{6}$$

This is a standard least square problem with regularization, with its solution as

$$U_{(\tau+1)}^{(1)} = ({P_{c(\tau)}}^{\mathrm{T}}P_{c(\tau)} + \lambda I_k)^{-1}{P_{c(\tau)}}^{\mathrm{T}}X_c^{(1)}. \tag{7}$$

Here I_k is the identity matrix with k-dimension. Similarly, we have the following function to update $U_{(\tau+1)}^{(2)}$:

$$U_{(\tau+1)}^{(2)} = ({P_{c(\tau)}}^{\mathrm{T}}P_{c(\tau)} + \lambda I_k)^{-1}{P_{c(\tau)}}^{\mathrm{T}}X_c^{(2)}. \tag{8}$$

Update P. This part includes three subproblems, i.e., update $P_{c(\tau+1)}$, $\hat{P}_{(\tau+1)}^{(1)}$ and $\hat{P}_{(\tau+1)}^{(2)}$. For $P_{c(\tau+1)}$, by fixing other variables, the corresponding Lagrangian function is

$$\begin{aligned}\mathcal{C}(P_c) = &\|X_c^{(1)} - P_cU^{(1)}\|_{\mathrm{F}}^2 + \|X_c^{(2)} - P_cU^{(2)}\|_{\mathrm{F}}^2 \\ &+ \langle Y_c, P_c - Q_c\rangle + \frac{\mu}{2}\|P_c - Q_c\|_{\mathrm{F}}^2.\end{aligned} \tag{9}$$

With the help of KKT condition that $\partial(\mathcal{C}(P_c))/\partial(P_c) = 0$, we have the following solver for $P_{c(\tau+1)} =:$

$$\left(2X_c^{(1)}{U_{(\tau+1)}^{(1)}}^{\mathrm{T}} + 2X_c^{(2)}{U_{(\tau+1)}^{(2)}}^{\mathrm{T}} - Y_{c(\tau)} + \mu Q_{c(\tau)}\right)R_{(\tau+1)}^{-1}, \tag{10}$$

where $R_{(\tau+1)} = 2U_{(\tau+1)}^{(1)}{U_{(\tau+1)}^{(1)}}^{\mathrm{T}} + 2U_{(\tau+1)}^{(2)}{U_{(\tau+1)}^{(2)}}^{\mathrm{T}} + \mu I_k$. Similarly, we obtain the solutions for $\hat{P}_{(\tau+1)}^{(1)}$ as

$$(2X_c^{(1)}{U_{(\tau+1)}^{(1)}}^{\mathrm{T}} - Y_{1(\tau)} + \mu\hat{Q}_{(\tau)}^{(1)})(2U_{(\tau+1)}^{(1)}{U_{(\tau+1)}^{(1)}}^{\mathrm{T}} + \mu I_k)^{-1}, \tag{11}$$

and $\hat{P}_{(\tau+1)}^{(2)}$ as

$$(2X_c^{(2)}{U_{(\tau+1)}^{(2)}}^{\mathrm{T}} - Y_{2(\tau)} + \mu\hat{Q}_{(\tau)}^{(2)})(2U_{(\tau+1)}^{(2)}{U_{(\tau+1)}^{(2)}}^{\mathrm{T}} + \mu I_k)^{-1}. \tag{12}$$

Update Q. Recall that the motivation of introducing auxiliary $Q = [Q_c; \hat{Q}^{(1)}; \hat{Q}^{(2)}]$ is to bridge the gap of global representation of all the data samples in different modalities. Therefore, instead of individually updating $Q_{c(\tau+1)}$, $\hat{Q}_{(\tau+1)}^{(1)}$ and $\hat{Q}_{(\tau+1)}^{(2)}$. We update $Q_{(\tau+1)}$ as a whole, with the Lagrangian function written as

$$\mathcal{C}(Q) = \beta\mathrm{tr}(Q^{\mathrm{T}}L_AQ) + \langle Y, P-Q\rangle + \frac{\mu}{2}\|P-Q\|_{\mathrm{F}}^2. \tag{13}$$

Correspondingly, the solver of $Q_{(\tau+1)}$ via KKT condition is

$$\left(\beta(L_{A_{(\tau)}}^{\mathrm{T}} + L_{A(\tau)}) + \mu I_N\right)^{-1}(Y_{(\tau)} + \mu P_{(\tau+1)}), \tag{14}$$

where I_N is the identity matrix with N-dimension.

Update A ($\mathbf{L_A}$). Fixing other variables, the graph A-problem is in the following form

$$\begin{aligned} &\min_{A} \beta \mathrm{tr}(Q^{\mathrm{T}} L_A Q) + \gamma \|A\|_{\mathrm{F}}^2 \\ &\text{s.t. } \forall i \; A_i^T \mathbf{1} = 1; \; A_i \succeq 0. \end{aligned} \tag{15}$$

As discussed in **Remark 2**, A has the probability interpretation with each element considered as the similarity probability between two data samples. Therefore, we divide problem (15) into a set of subproblems $A^i_{(\tau+1)}$ according to sample index i as

$$A^i_{(\tau+1)} = \underset{A^i \in \{\alpha | \alpha^T \mathbf{1} = 1; \alpha \succeq 0\}}{\operatorname{argmin}} \|A^i + S^i_{(\tau+1)}\|_{\mathrm{F}}^2, \tag{16}$$

where $S^i_{(\tau+1)}$ is a column vector with its element j defined as $S^{ij}_{(\tau+1)} = \dfrac{\beta \|Q^i_{(\tau+1)} - Q^j_{(\tau+1)}\|_{\mathrm{F}}^2}{4\gamma}$. A detailed deduction can be found in [Guo, 2015].

To sum up, for the complete algorithm, we initialize the variables and parameters (in the iteration #0, denoted as "$*_{(0)}$") in ALM as follows: penalty parameter $\mu_{(0)} = 10^{-3}$, $\rho = 1.1$, the max penalty parameter $\mu_{\max} = 10^6$, stopping threshold $\epsilon = 10^{-6}$, $P_{(0)} = Q_{(0)} = Y_{(0)} = \mathbf{0} \in \mathbb{R}^{N \times k}$, $P_{c(0)} = Q_{c(0)} = Y_{c(0)} = \mathbf{0} \in \mathbb{R}^{c \times k}$, $\hat{P}^{(1)}_{(0)} = \hat{Q}^{(1)}_{(0)} = \hat{Y}^{(1)}_{(0)} = \mathbf{0} \in \mathbb{R}^{m \times k}$, $\hat{P}^{(2)}_{(0)} = \hat{Q}^{(2)}_{(0)} = \hat{Y}^{(2)}_{(0)} = \mathbf{0} \in \mathbb{R}^{n \times k}$, $U^{(1)}_{(0)} = \mathbf{0} \in \mathbb{R}^{k \times d_1}$, $U^{(2)}_{(0)} = \mathbf{0} \in \mathbb{R}^{k \times d_2}$, $A_{(0)} = L_{A(0)} = \mathbf{0} \in \mathbb{R}^{N \times N}$. Then we update each variable one by one as discussed above until convergence.

2.4 Complexity Analysis

Note that with different partial example ratios, the dimensions of $\hat{X}^{(1,2)}$, $\hat{X}^{(1)}$ and $\hat{X}^{(2)}$ are different, i.e., $c,\ m,\ n$ vary. For simplicity, we consider the extreme case that no incomplete data exist, that is, complete (traditional) multi-view clustering case. The feature dimensions of different modalities are all d.

In **Algorithm 1**, the most time-consuming parts are the matrix multiplication and inverse operations when updating $U^{(1)},\ U^{(2)},\ P,\ Q,\ A$. For each iteration, the inverse operations in Eqs. (7)(8)(10)(11)(12) cost $\mathcal{O}(k^3)$ due to the $k \times k$ size matrix. While the inverse on graph in Eq. (14) takes time of $\mathcal{O}(N^3)$. Usually $k \ll N$, then the asymptotic upper-bound for inverse operation can be expressed as $\mathcal{O}(N^3)$. The multiplication operations take $\mathcal{O}(dkN)$ when updating $U^{(1)},\ U^{(2)},\ P,\ Q$. It costs $\mathcal{O}(N^3)$ when updating A. However, it is noted that the number of operations of $O(N^3)$ in each iteration is only 2, the major computations are of order $\mathcal{O}(dkN)$. Suppose M is the number of operations consuming $O(dkN)$, L is the iteration time. In sum, the time complexity of our algorithm is $\mathcal{O}(MLdkN + 2LN^3)$.

3 Experiment

- **Synthetic data** are comprised of two modalities. We first choose the cluster c_i each sample belongs to, and then generate each of the modalities $x_i^{(1)}$ and $x_i^{(2)}$ from a two-component Gaussian mixture model. Two modalities are combined to form the sample $(x_i^{(1)}, x_i^{(2)}, c_i)$. We sample 100 points from each modality. The cluster means in modal-1 are $\mu_1^{(1)} = (1\ \ 1)$, $\mu_2^{(1)} = (3\ \ 4)$, in modal-2 are $\mu_1^{(2)} = (1\ \ 2)$, $\mu_2^{(2)} = (2\ \ 2)$. The covariance for modal-1 are

$$\Sigma_1^{(1)} = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1.5 \end{pmatrix}, \Sigma_2^{(1)} = \begin{pmatrix} 0.3 & 0.2 \\ 0.2 & 0.6 \end{pmatrix}$$

$$\Sigma_1^{(2)} = \begin{pmatrix} 0.01 & 0 \\ 0 & 0.01 \end{pmatrix}, \Sigma_2^{(2)} = \begin{pmatrix} 0.6 & 0.1 \\ 0.1 & 0.5 \end{pmatrix}$$

- **Real-world visual data:** (a) ***MSR Action Pairs dataset*** [Oreifej and Liu, 2013] is a RGB-D action dataset containing 12 types of activities performed by 10 subjects. Each actor provides 360 videos for each modality. (b) ***MSR Daily Activity dataset*** [Wang *et al.*, 2012] contains 16 types of activities performed by 10 subjects. Each actor repeats an action twice, providing 320 videos for each of the RGB and depth channels. For the above two RGBD video sequences, we temporally normalize each video clip to 10 frames with spatial resolution of 120×160. Histograms of gradient oriented feature is extracted from both depth and RGB videos with patch size 8×8. Thus, a total of 3000 patches are extracted from each video, with the feature dimensionality of 31. We will clarify this in our final version. (c) ***BUAA NirVis*** [Huang *et al.*, 2012] contains two types of data, i.e., visual spectral (VIS) and near infrared (NIR) data. The first 10 subjects with 180 images are used. To fasten the computation, we resize the images to 10×10, and vectorize them. (d) ***UCI handwritten digit***[1] consists of 0-9 handwritten digits data from UCI repository. It includes 2000 examples, with one modality being the 76 Fourier coefficients and modal-2 being the 240 pixel averages in 2×3 windows.

For the compared methods, we consider the following algorithms as the baselines. (1) **BSV** (Best Single View): Due to the missing samples in each modality, we cannot directly perform k-means clustering on each modality data. Following [Shao *et al.*, 2015], we firstly fill in all the missing data with the average features for each modality, and then perform clustering on each modality, and report the best result. (2) **Concat**: Feature concatenation is a straightforward way to deal with multi-modal data, which serves as our second baseline. Same as BSV, we firstly fill in all the missing data with the average features for each modality, and then concatenate all modal features into one. (3) **MultiNMF**: Multi-view NMF [Liu *et al.*, 2013] seeks a common latent subspace based on joint NMF, which can be approximately regarded as the complete-view case of PVC. For the synthetic data, there are few data points containing negative values. In order to successfully run the code, we make the input data nonnegative as preprocessing. (4) **PVC**: Partial multi-view clustering [Li *et al.*, 2014] is one of the most recent works in dealing with incomplete multi-modal data. This work can be considered as our proposed model without the complete graph Laplacian. One important parameter on regularizer λ is chosen from the parameter grid of $\{1e\text{-}4, 1e\text{-}3, 1e\text{-}2, 1e\text{-}1, 1e0, 1e1, 1e2\}$, including the default $1e$-2 used in the original paper.

[1] http://archive.ics.uci.edu/ml/datasets.html

Table 1: NMI/Precision results on synthetic data under different PER settings.

Method \ PER	0.1	0.3	0.5	0.7	0.9
BSV	0.4219 / 0.5233	0.3600 / 0.5439	0.1767 / 0.5147	0.1646 / 0.5118	0.0820 / 0.5109
Concat	0.4644 / 0.6922	0.4019 / 0.6436	0.3762 / 0.6159	0.3000 / 0.5711	0.2278 / 0.5965
MultiNMF	0.5767 / 0.8103	0.5699 / 0.8325	0.4430 / 0.7694	0.4298 / 0.7325	0.3677 / 0.6985
PVC	0.6194 / 0.8064	0.5820 / 0.8309	0.5512 / 0.8187	0.5142 / 0.7985	0.4185 / 0.6833
Ours	**0.8781 / 0.9585**	**0.8362 / 0.9303**	**0.7433 / 0.8816**	**0.7959 / 0.9176**	**0.4580 / 0.6947**

Table 2: NMI/Precision results on MSR Action Pairs dataset under different PER settings.

Method \ PER	0.1	0.3	0.5	0.7	0.9
BSV	0.4807 / 0.2687	0.4807 / 0.2687	0.3691 / 0.1660	0.2874 / 0.1190	0.2779 / 0.1085
Concat	0.6270 / 0.3538	0.5803 / 0.3306	0.5512 / 0.3030	0.5123 / 0.2750	0.4685 / 0.2268
MultiNMF	0.6033 / 0.4038	0.5149 / 0.2984	0.5008 / 0.2828	0.4816 / 0.2539	0.4463 / 0.2267
PVC	**0.6917** / 0.4490	0.6501 / 0.3998	0.6356 / 0.3734	0.6012 / 0.3662	0.5882 / 0.3629
Ours	0.6859 / **0.4504**	**0.6763 / 0.4431**	**0.6504 / 0.3836**	**0.6468 / 0.3774**	**0.6396 / 0.3734**

Table 3: NMI/Precision results on MSR Daily Activity dataset under different PER settings.

Method \ PER	0.1	0.3	0.5	0.7	0.9
BSV	0.2012 / 0.0826	0.1851 / 0.0765	0.1683 / 0.0680	0.1487 / 0.0641	0.1328 / 0.0626
Concat	0.2499 / 0.1137	0.2354 / 0.0997	0.2261 / 0.0843	0.2031 / 0.0755	0.1878 / 0.0758
MultiNMF	0.2077 / 0.0841	0.2057 / 0.0911	0.1924 / 0.0806	0.1823 / 0.0713	0.1655 / 0.0674
PVC	0.2605 / 0.1385	0.2487 / **0.1275**	0.2236 / 0.1086	0.2175 / 0.1049	0.2062 / 0.0902
Ours	**0.2807 / 0.1489**	**0.2554** / 0.1263	**0.2512 / 0.1241**	**0.2421 / 0.1108**	**0.2201 / 0.0907**

For the evaluation metric, we follow [Li *et al.*, 2014], using Normalized Mutual Information (NMI). Besides, precision of clustering result is also reported to give a comprehensive view. Same as [Li *et al.*, 2014], we test all the methods under different partial/incomplete example ratio (PER) varying from 0.1 to 0.9 with an interval of 0.2.

3.1 Experimental Result

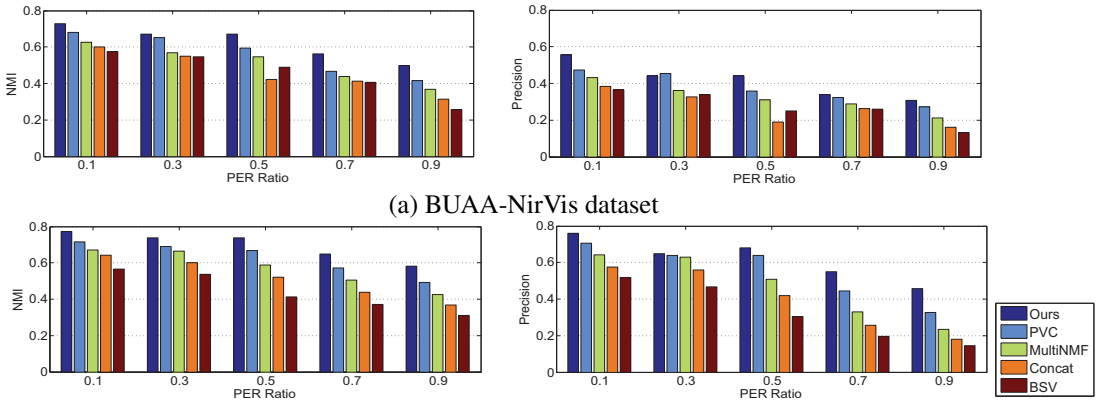
For each dataset, we randomly select samples from each modality as the missing ones. Note that our method not only learns a better low-dimensional representation but also learns a similarity matrix among samples iteratively. This naturally gives us two opportunities to do clustering, i.e., k-means clustering on the latent representation P, and spectral clustering on the learned affinity graph A. To make the fair comparison, k-means results on P are reported. For each experiment, we repeats 10 times to get the average performance, the standard deviations are omitted here because it is observed that these values are usually small.

Table 1,2,3 and Figure 2 report the NMI values and precision on synthetic, video and image datasets with different PER ratio settings. From these tables and bar graphs, the following observations and discussions are made.

- The proposed method performs superiorly to the other baselines in almost all the settings; Especially for the challenging synthetic data, we raise the performance bar by around 31.83% in NMI.
- With more missing modal samples (PER ratio increases), the performance of all the methods drops.
- With more missing modal samples, our method improves more compared with the state-of-the-art baseline. Specifically, our NMI improvement reaches 10.34% (PER=0.9) from 4.58% (PER=0.1) for five real-world video/image datasets.
- For the real-world data, as PER ratio grows, the extent that performance drops is less than that of synthetic data.

Discussion: The first observation experimentally demonstrates that the proposed complete graph Laplacian term works in both synthetic and video/image multi-modal data, especially when some modal data are missing to a large extent. Note that for the matrix factorization part, we use the simplest way with only Frobenius norm regularization on basis matrix. However, we still outperform the competitors with the help of complete graph Laplacian term. With better matrix factorization techniques, e.g. NMF in [Li *et al.*, 2014] or weighted NMF in [Shao *et al.*, 2015], we believe that a better performance will be achieved.

With no doubt, the problem becomes more challenging when the number of shared samples is fewer. However, one may be curious *why our method performs much better than others in synthetic data*. The possible reason is that compared with modal-1 data, modal-2 data are difficult to separate. When points in modal-1 are missing, the existing methods cannot do a good job with only modal-2 data even in latent space. However, thanks to the affinity graph built on all data points, the data points from different clusters are iteratively pulled towards their corresponding cluster centers by the influence of global constraint. This enlightens us that in real-world multi-modal visual data, if one modal data perform poorly (e.g. people blend in the background visually) than the others, our proposed complete graph Laplacian term is capable to make it up from other discriminative modal data (e.g. discriminative depth information between people and background). One may be also interested in



(a) BUAA-NirVis dataset

(b) UCI handwritten digits dataset

Figure 2: NMI/Precision results on (a) BUAA-NirVis dataset and (b) UCI handwritten digits dataset.

the reason *why our method has a considerable improvement when data suffer from a large incompleteness*. We believe, as PER ratio increases, the state-of-the-art method PVC degenerates dramatically because the common projection P_c becomes harder to be accurately estimated simply from the less shared multi-modal data. Nevertheless, our proposed complete graph Laplacian remedies the deviation by considering the global structure of incomplete multi-modal data in the latent space, which further leads to a robust grouping structure.

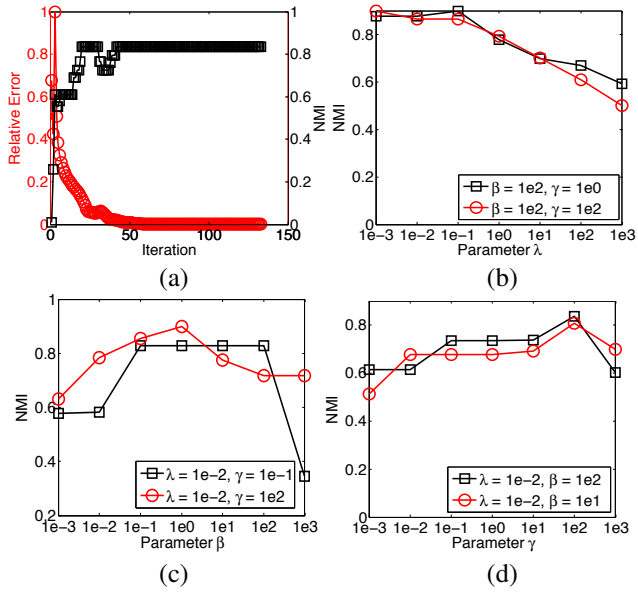
3.2 Convergence Study

To show the convergence property, we conduct an experiment on synthetic data with PER ratio set as 0.3 and parameters $\{\lambda,\ \beta,\ \gamma\}$ set as $\{1e\text{-}2,\ 1e2,\ 1e2\}$. The relative error of stop criterion $\|P_\tau - Q_\tau\|_\infty$ is computed in each iteration. The red curve in Figure 3(a) plots the convergence curve of our model. It is observed that after the first several iterations' bump, the relative error drops steadily, and then meets the convergence at around #40 iteration. The NMI value during each iteration is drawn in black. It can be seen that there are three stages before converging: the first stage (from #1 to #4), the NMI value grows dramatically; the second stage (from #5 to #40), the NMI bumps in a certain range but grows; the final stage (from #41 to the end), the NMI achieves the best at the convergence point.

3.3 Parameter Study

There are three major parameters in our approach, i.e., λ, β and γ. Same as convergence study, we conduct the parameter analytical experiments on synthetic data with PER ratio set as 0.3. Figure 3(b) shows the experiment of NMI result w.r.t. the parameter λ under two settings $\{[\beta{=}1e2,\ \gamma{=}1e0];\ [\beta{=}1e2,\ \gamma{=}1e2]\}$. We select the parameter α in the grid of $\{1e\text{-}3,\ 1e\text{-}2,\ 1e\text{-}1,\ 1e0,\ 1e1,\ 1e2,\ 1e3\}$. It is observed that our method has a relatively good performance when λ is in the range of $[1e\text{-}3,\ 1e\text{-}1]$, and drops when λ becomes larger.

The experiments shown in Figure 3(c,d) are designed to test the robustness of our model w.r.t. the trade-off parameters β and γ on the proposed graph Laplacian term. As we observe, the NMIs under different settings reach a relatively good performance when $\beta = [1e\text{-}1,\ 1e0,\ 1e1]$ and $\gamma = 1e2$.



(a) (b) (c) (d)

Figure 3: Convergence and parameter studies on synthetic data. (a) shows the relative error and NMI result w.r.t. iteration times. (b-d) plot the NMI results in terms of parameters λ, β and γ respectively. For each parameter analysis, we run two different settings shown in red circle and black cross.

4 Conclusion

In this paper, we proposed a method dealing with incomplete multi-modal visual data grouping problem with the consideration of the compact global structure via a novel graph Laplacian term. This practice bridged the connection of missing samples data from different modalities. Superior experimental results on synthetic data and four real-world multi-modal visual datasets compared with several baselines validated the effectiveness of our method.

Acknowledgments

This research is supported in part by the NSF CNS award 1314484, ONR award N00014-12-1-1028, ONR Young Investigator Award N00014-14-1-0484, and U.S. Army Research Office Young Investigator Award W911NF-14-1-0218.

References

[Bickel and Scheffer, 2004] S. Bickel and T. Scheffer. Multi-view clustering. In *IEEE International Conference on Data Mining (ICDM)*, 2004.

[Blaschko and Lampert, 2008] M.B. Blaschko and C.H. Lampert. Correlational spectral clustering. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008.

[Cai *et al.*, 2013] X. Cai, F. Nie, and H. Huang. Multi-view k-means clustering on big data. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 2013.

[Cao *et al.*, 2015] Xiaochun Cao, Changqing Zhang, Huazhu Fu, Si Liu, and Hua Zhang. Diversity-induced multi-view subspace clustering. In *IEEE Conference on Computer Vision and Pattern Recognition, (CVPR)*, pages 586–594, 2015.

[Chaudhuri *et al.*, 2009] Kamalika Chaudhuri, Sham M. Kakade, Karen Livescu, and Karthik Sridharan. Multi-view clustering via canonical correlation analysis. In *International Conference on Machine Learning (ICML)*, pages 129–136, 2009.

[Ding and Fu, 2014] Zhengming Ding and Yun Fu. Low-rank common subspace for multi-view learning. In *2014 IEEE International Conference on Data Mining (ICDM)*, pages 110–119. IEEE, 2014.

[Fred and Jain, 2005] A.L. Fred and A.K. Jain. Combining multiple clusterings using evidence accumulation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 27(6):835–850, 2005.

[Guo, 2015] Xiaojie Guo. Robust subspace segmentation by simultaneously learning data representations and their affinity matrix. In *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, (IJCAI)*, pages 3547–3553, 2015.

[Huang *et al.*, 2012] Di Huang, Jia Sun, and Yunhong Wang. The buaa-visnir face database instructions. *IRIP-TR-12-FR-001*, 2012.

[Li *et al.*, 2014] Shao-Yuan Li, Yuan Jiang, and Zhi-Hua Zhou. Partial multi-view clustering. In *AAAI Conference on Artificial Intelligence (AAAI)*, pages 1968–1974, 2014.

[Lin *et al.*, 2011] Zhouchen Lin, Risheng Liu, and Zhixun Su. Linearized alternating direction method with adaptive penalty for low-rank representation. In *Neural Information Processing Systems (NIPS)*, pages 612–620, 2011.

[Liu *et al.*, 2013] J. Liu, C. Wang, J. Gao, and J. Han. Multi-view clustering via joint nonnegative matrix factorization. In *SIAM International Conference on Data Mining (SDM)*, pages 252–260, 2013.

[Liu *et al.*, 2016] Tongliang Liu, Dacheng Tao, Mingli Song, and Stephen J. Maybank. Algorithm-dependent generalization bounds for multi-task learning. *DOI 10.1109/TPAMI.2016.2544314*, 2016.

[Lu *et al.*, 2012] Can-Yi Lu, Hai Min, Zhong-Qiu Zhao, Lin Zhu, De-Shuang Huang, and Shuicheng Yan. Robust and efficient subspace segmentation via least squares regression. In *European Conference on Computer Vision (ECCV)*, pages 347–360, 2012.

[Oreifej and Liu, 2013] Omar Oreifej and Zicheng Liu. HON4D: histogram of oriented 4d normals for activity recognition from depth sequences. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 716–723, 2013.

[Shao *et al.*, 2015] Weixiang Shao, Lifang He, and Philip S. Yu. Multiple incomplete views clustering via weighted nonnegative matrix factorization with $l_{2,1}$ regularization. In *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD*, pages 318–334, 2015.

[Singh and Gordon, 2008] A.P. Singh and G.J. Gordon. Relational learning via collective matrix factorization. In *ACM SIGKDD international conference on knowledge discovery and data mining (KDD)*, pages 650–658, 2008.

[Wang *et al.*, 2012] Jiang Wang, Zicheng Liu, Ying Wu, and Junsong Yuan. Mining actionlet ensemble for action recognition with depth cameras. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1290–1297, 2012.

[Zhang *et al.*, 2015] Changqing Zhang, Huazhu Fu, Si Liu, Guangcan Liu, and Xiaochun Cao. Low-rank tensor constrained multiview subspace clustering. In *2015 IEEE International Conference on Computer Vision, (ICCV)*, pages 1582–1590, 2015.

[Zhao and Fu, 2015] Handong Zhao and Yun Fu. Dual-regularized multi-view outlier detection. In *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, (IJCAI)*, pages 4077–4083, 2015.