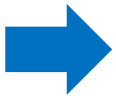
CCF-CV: University of Macau

Multi-View Fusion and Representation for Image Analysis

Zhe Xue（薛 哲）

Assistant Professor
School of Computer Science
Beijing University of Posts and Telecommunications

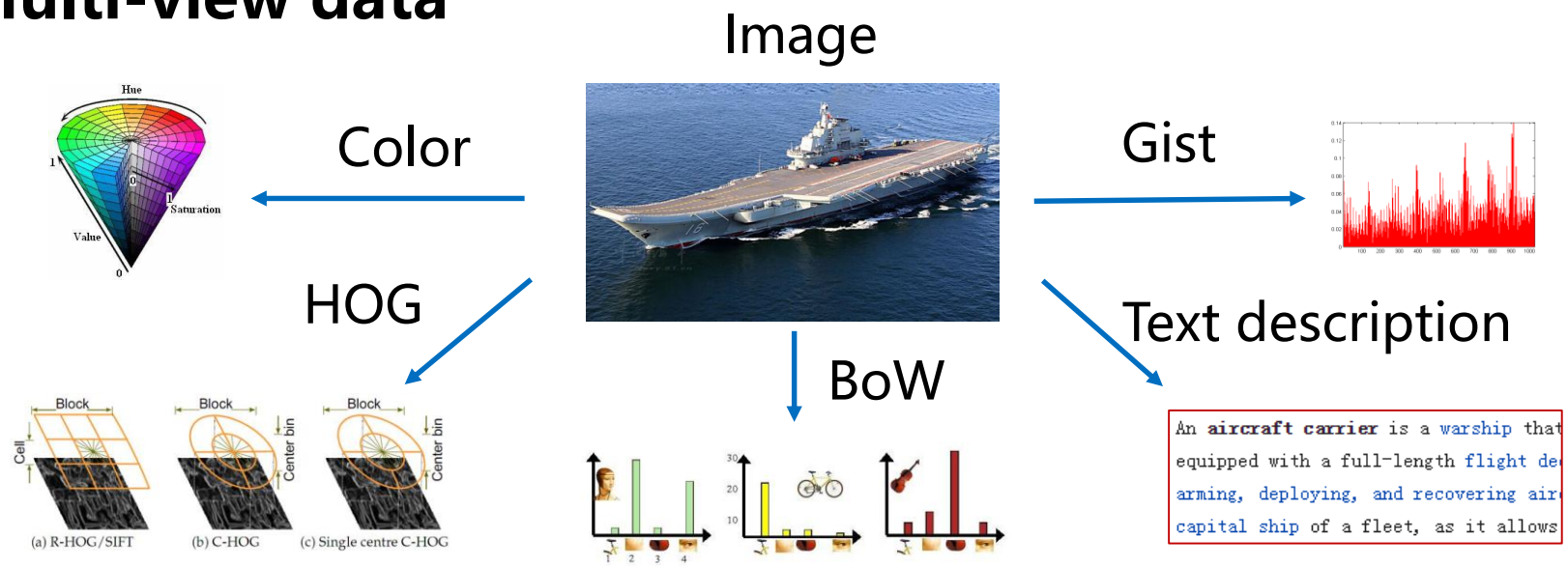
OUTLINE



1. Background
2. Related work
3. Research works
4. Summary

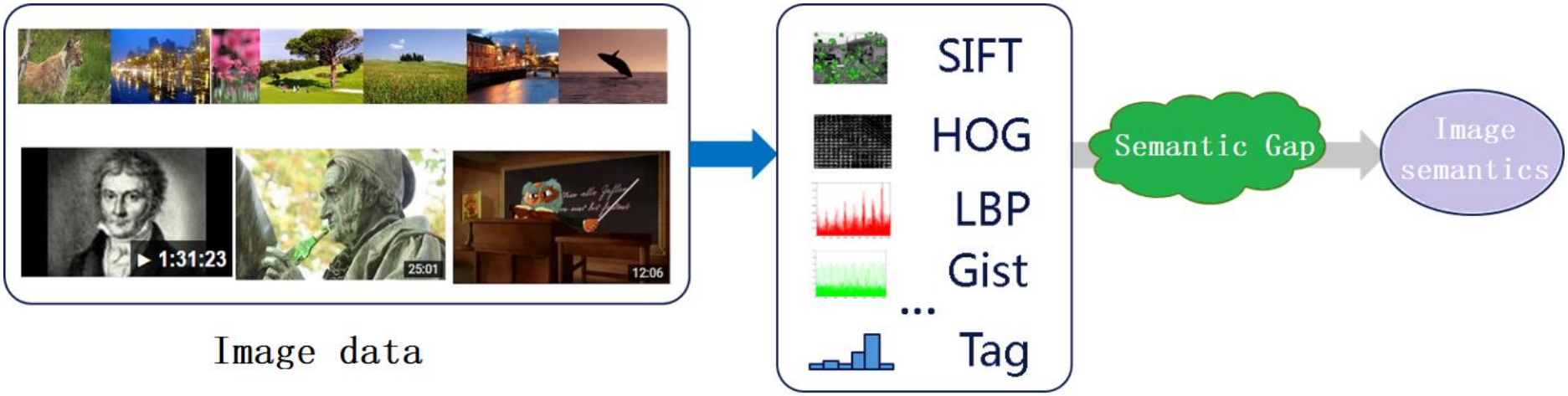
Background

Multi-view data



- In real world, data is often presented from various perspectives. For example, an image can be described by a variety of features
- Using a single view may miss some information, so that the semantic relationship between images cannot be described accurately. Multi-view features can complement each other, which should be used
- Multi-view learning is to effectively fuse information of different views and represent different views in a unified way, so that different views can complement each other and achieve better performance than using a single view

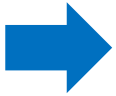
Challenges



Challenges in multi-view learning:

- There is a semantic gap between low-level multi-view features and high-level semantics
- Different views have different physical meanings, and they are difficult to fuse effectively
- The original multi-view high-dimensional features are heterogeneous and they lack a consistent representation

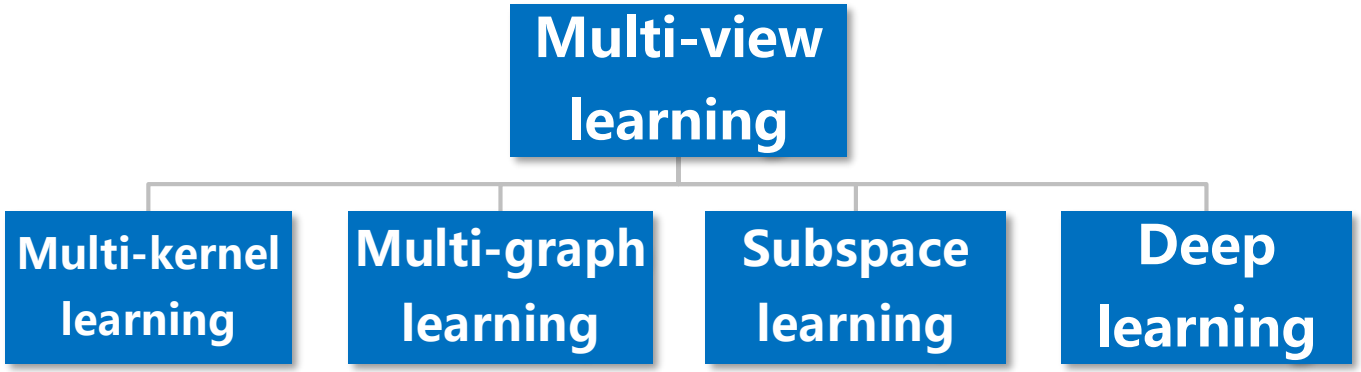
OUTLINE



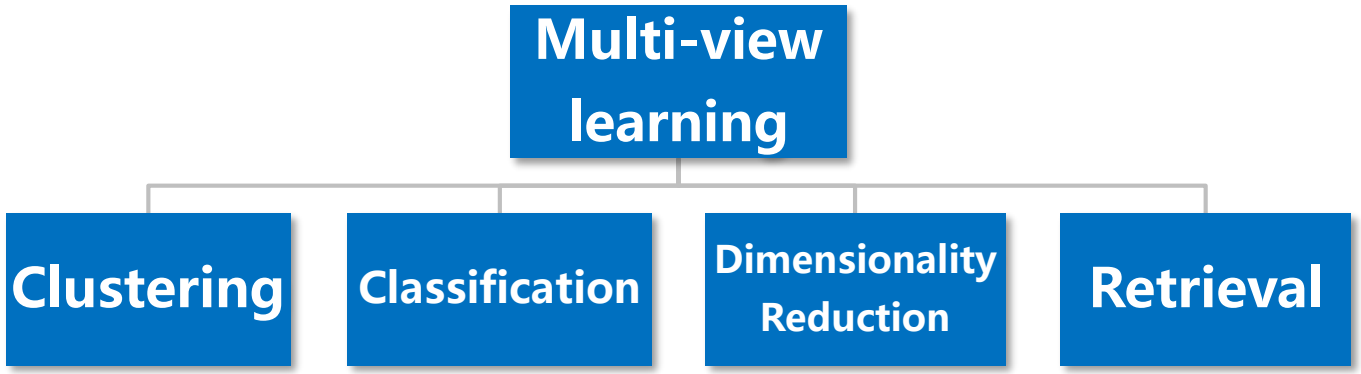
1. Background
2. Related work
3. Research works
4. Summary

Related work

Classify by methods

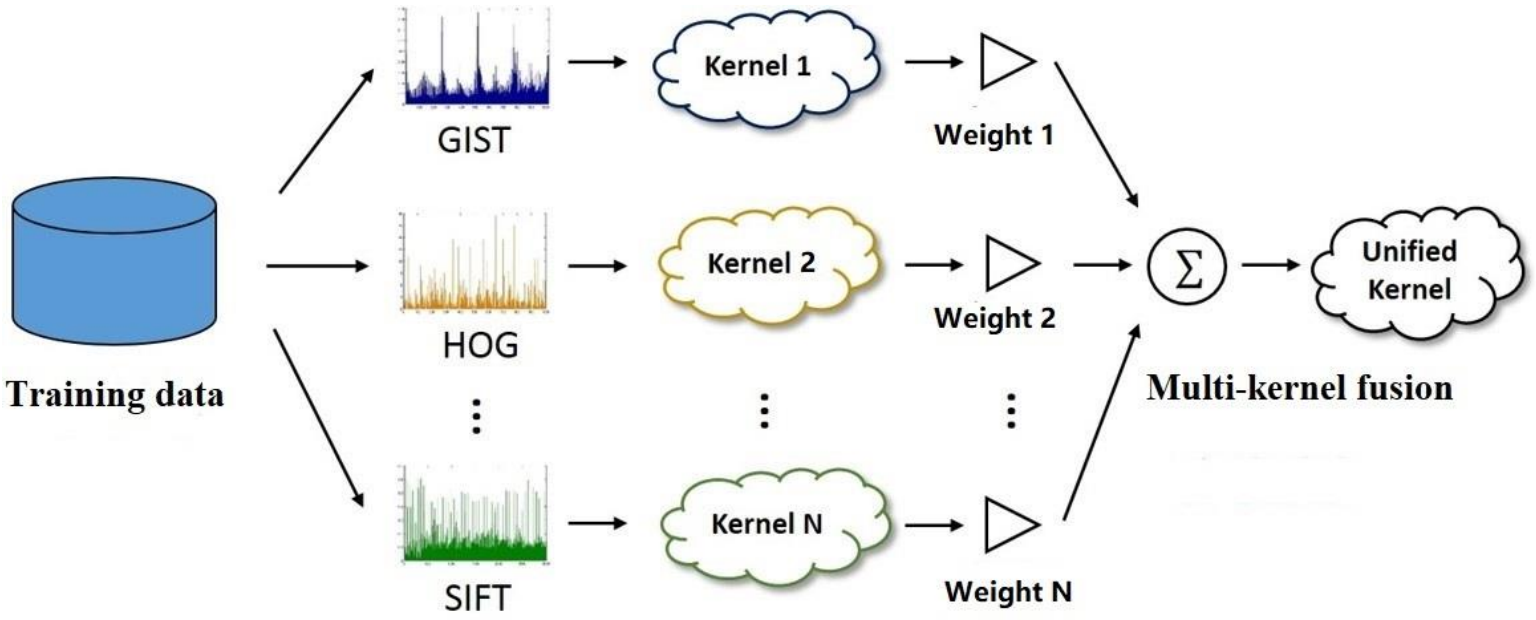


Classify by tasks



Related work

Multi-kernel learning



- Supervised multi-kernel learning:

[Joachims et al. 2001, Lanckriet et al. 2002, Lin et al. 2007, Gönen et al. 2008, Yang et al. 2012]

- Unsupervised multi-kernel learning:

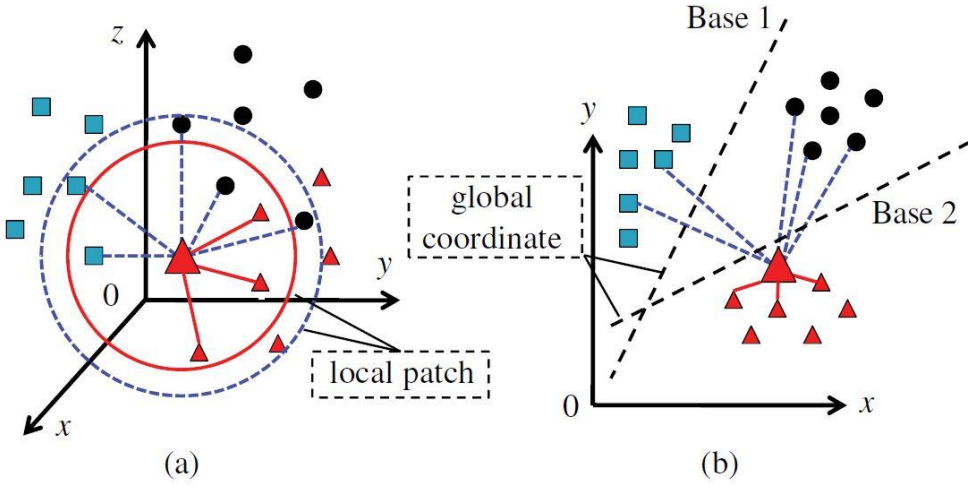
[Zhao et al. 2009, Lin et al. 2011, Tzortzis et al. 2012, Huang et al. 2012, Lu et al. 2014, Gönen et al. 2014]

Related work

Multi-graph learning：Image clustering

Use graphs to represent different views of data, and use graph learning method to fuse different views

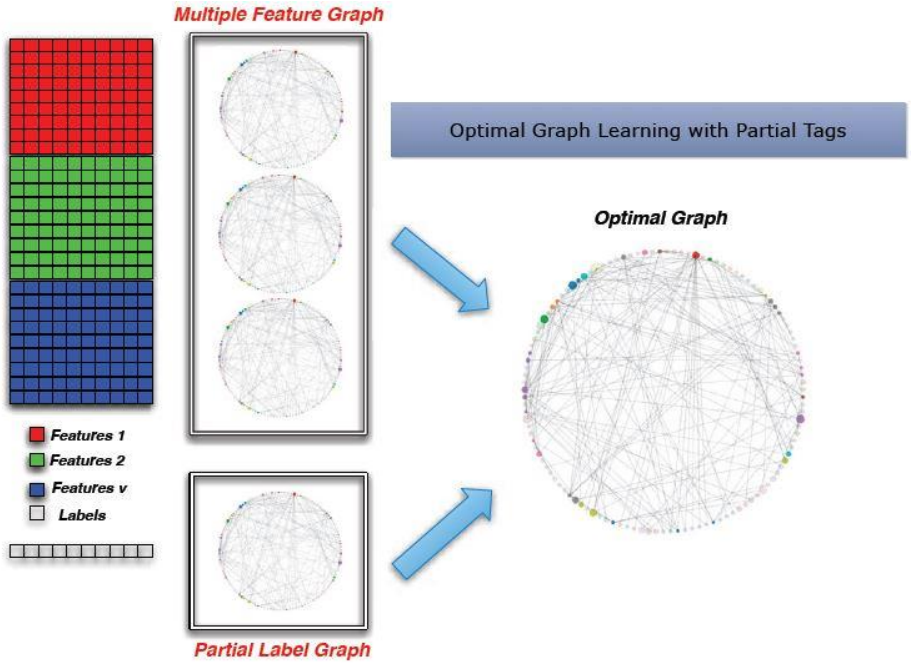
- Based on co-training：
  [Kumar et al. 2011(a), Kumar et al. 2011(b)]
- Based on weight learning to fuse graphs:
  [Huang et al. 2012, Tzortzis et al. 2012, Wang et al. 2014, Gui et al. 2014]



Related work

Multi-graph learning：Image annotation

- Using graphs to represent different views, and then the graphs are merged into a unified graph which preserves the relationship between data. Image annotation is conducted based on the graph

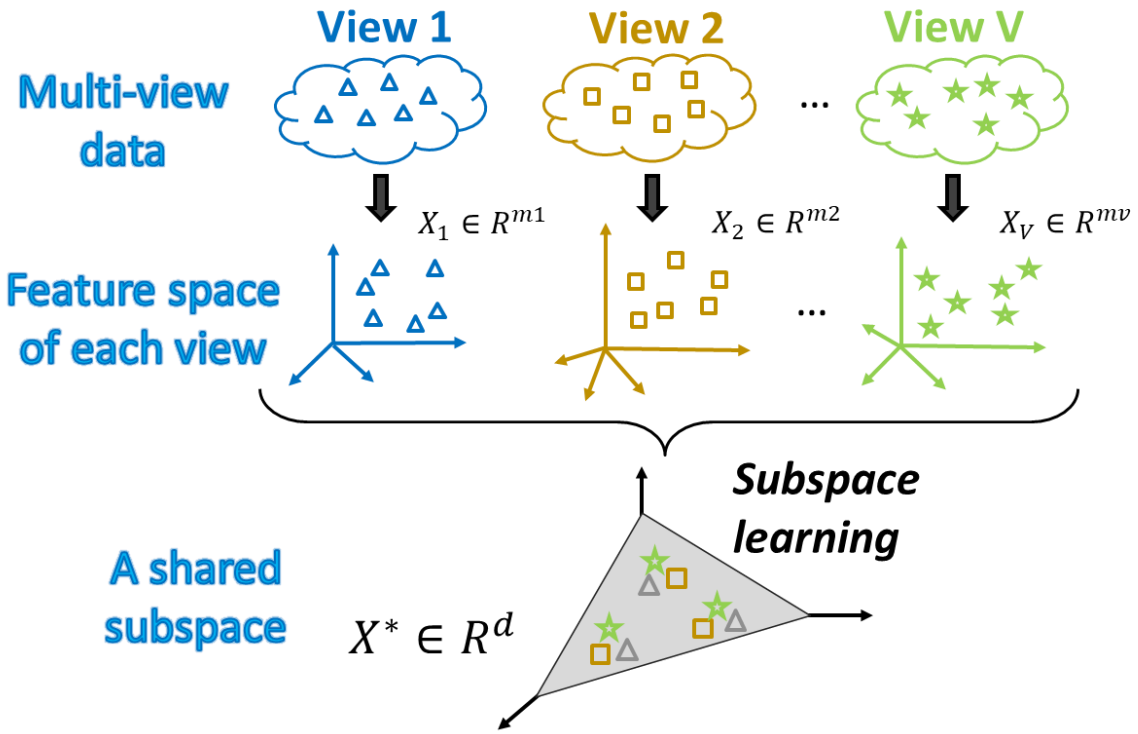


Gao L, Song J, Nie F, et al. “Optimal graph learning with partial tags and multiple features for image and video annotation”, *CVPR* 2015.

Related work

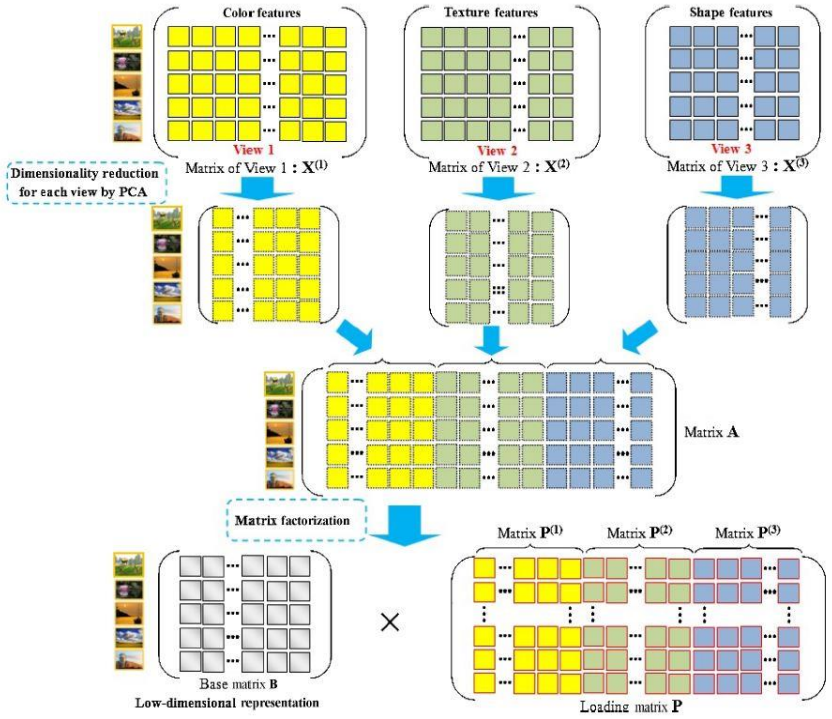
Subspace learning

- Canonical Correlation Analysis (CCA) [Hotelling et al. 1936]
- Kernelized Canonical Correlation Analysis (KCCA) [Akaho et al. 2006]



Related work

Subspace Learning: Dimension Reduction for Multi-view data

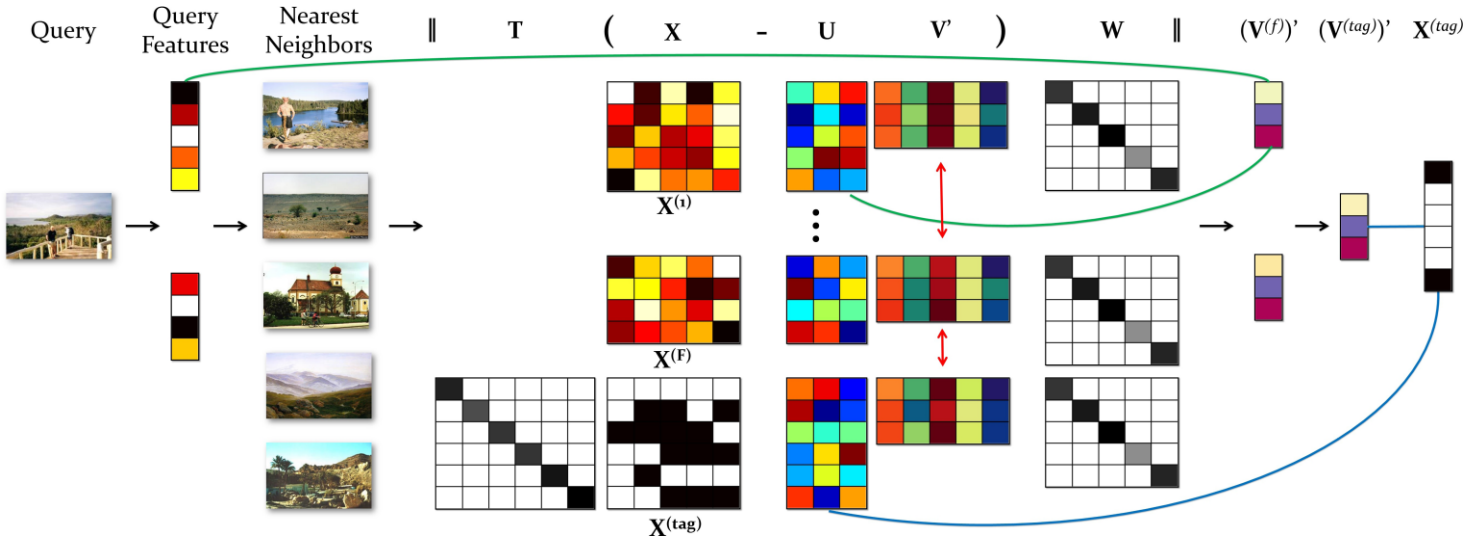


- First, PCA is performed for each view, and a unified representation is formed by concatenating them together
- Then, subspace learning is performed to obtain the final low-dimensional representation
- Sparse PCA learning is used to learn the subspace

Yahong Han, Fei Wu, Dacheng Tao, Jian Shao, Yueting Zhuang, Jianmin Jiang, "Sparse unsupervised dimensionality reduction for multiple view data." *TCSVT* 2012

Related work

Subspace learning：Image annotation



- First, a multi-view matrix is constructed by searching K-nearest neighbor images in the image database
- Then, image annotation for unlabeled data is achieved by conducting joint matrix factorization on multiple matrices

Kalayeh M M, Idrees H, Shah M. Nmf-knn: Image annotation using weighted multi-view non-negative matrix factorization, *CVPR 2014*

Related work

Deep learning

- Multi-view representation learning based on neural network

  [Andrew et al. 2013, Wang et al. 2015, Kan et al. 2016]

- Multi-view clustering based on deep matrix factorization

  [Zhao, et al. 2017]

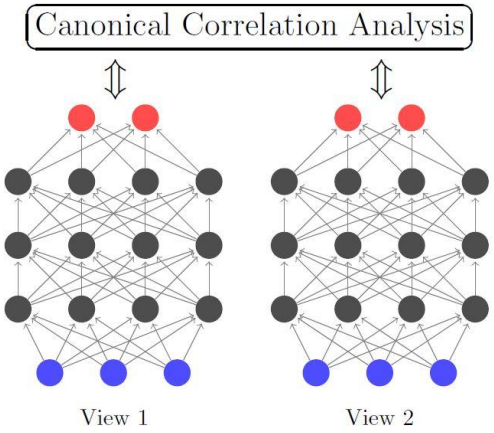
- Multi-view feature fusion based on convolution neural network CNN

  [Su et al. 2015, Wang et al. 2015, Zhu et al. 2016]

Related work

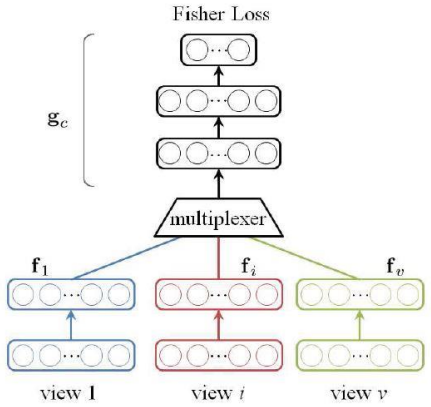
Deep learning：Multi-view representation learning

- Deep Canonical Correlation Analysis (DCCA) [Andrew et al. 2013]



- Extend CCA to multiple layers and maximize the data correlations at the top level
- More significant data correlations are obtained than CCA model

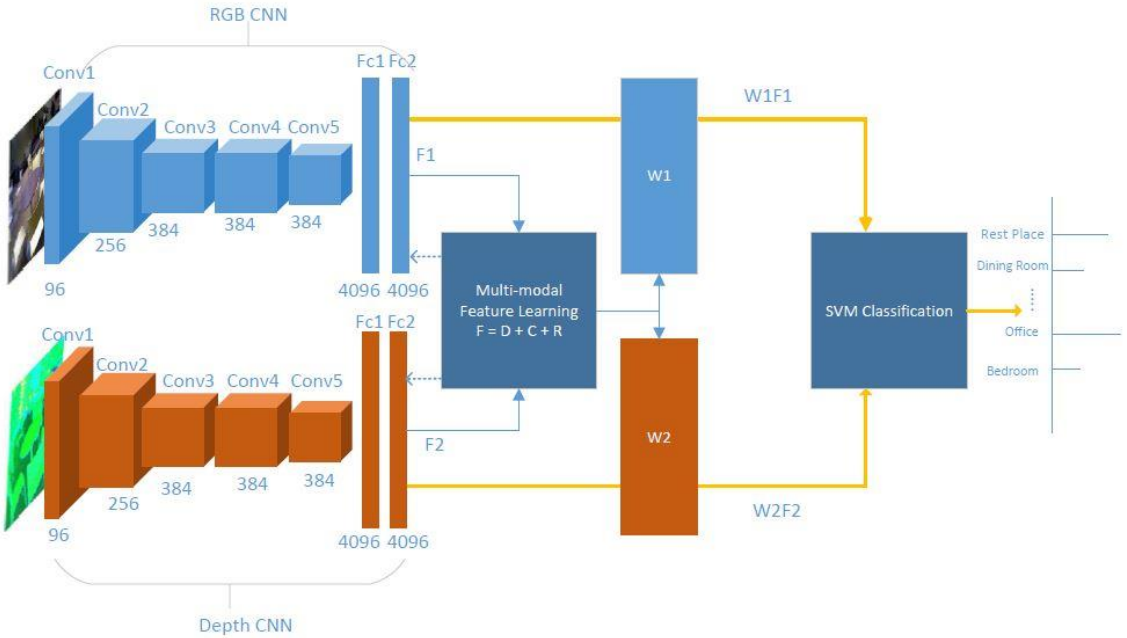
- Multi-view classification model based on deep neural network(MvDN) [Kan et al. 2016]



- Adding Fisher loss function at the top level to utilize discriminant information
- Multi-view data classification can be achieved by this network

Related work

Deep learning: RGBD scene recognition



- The color and depth features of images are used as input of the CNN network
- The outputs of two CNN channels are mapped to a new space, where class consistency and correlations of different views can be preserved

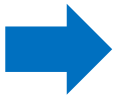
Hongyuan Zhu, Jean-Baptiste Weibel, Shijian Lu, Discriminative Multi-modal Feature Fusion for RGBD Indoor Scene Recognition, *CVPR 2016*

Related work

Summary

- Existing multi-view fusion methods usually adopt uniform fusion weights for all data. However, different data have different visual characteristics. Unified fusion weights cannot learn appropriate fusion weights
- Because different views have different physical meanings, they are not comparable. This problem is not considered in the multi-view learning methods
- Multi-view representation learning and the follow-up tasks are closely related. The existing methods conduct the two tasks independently, and their correlations are ignored

OUTLINE



1. Background
2. Related work
3. Research works
4. Summary

Research works

3.1 Multi-view clustering

✓ A group-aware multi-view fusion approach for image clustering

3.2 Multi-view dimensionality reduction

✓ Bi-level multi-view latent space learning

3.3 Multi-view classification for complete data

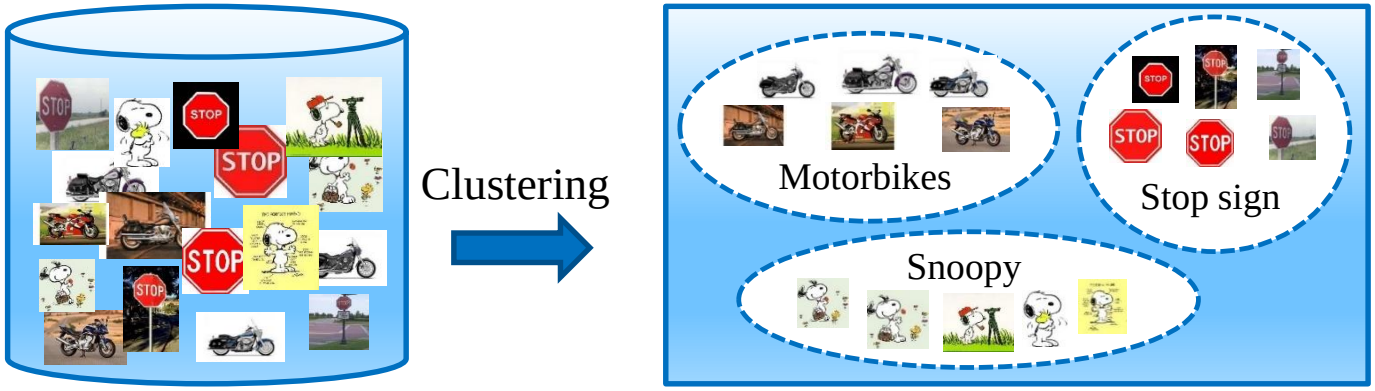
✓ Joint multi-view representation learning and image tagging

3.4 Multi-view semi-supervised classification for Incomplete data

✓ Deep Correlated Predictive Subspace Learning for Incomplete Multi-View Semi-Supervised Classification

3.1 A group-aware multi-view fusion approach for image clustering

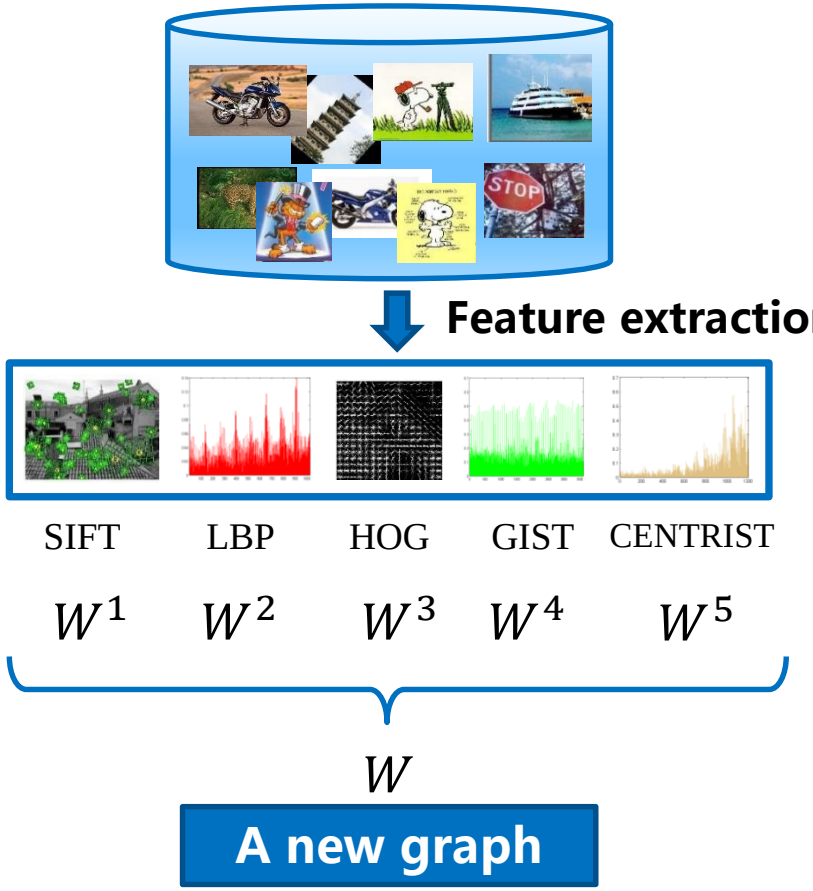
Background



- Image clustering is to divide the unlabeled image data into several image clusters. We expect the clustering results keep Intra-class consistency and inter-class diversity
- Image clustering can help people better organize and manage image data. It can automatically construct image categories without labeled samples

3.1 A group-aware multi-view fusion approach for image clustering

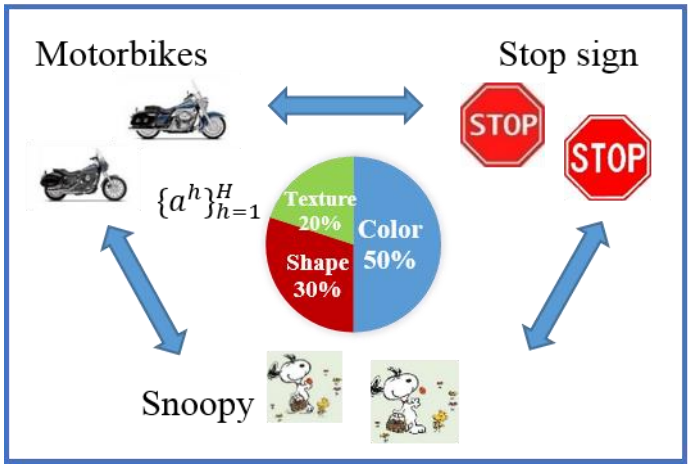
The procedure of traditional multi-view clustering methods

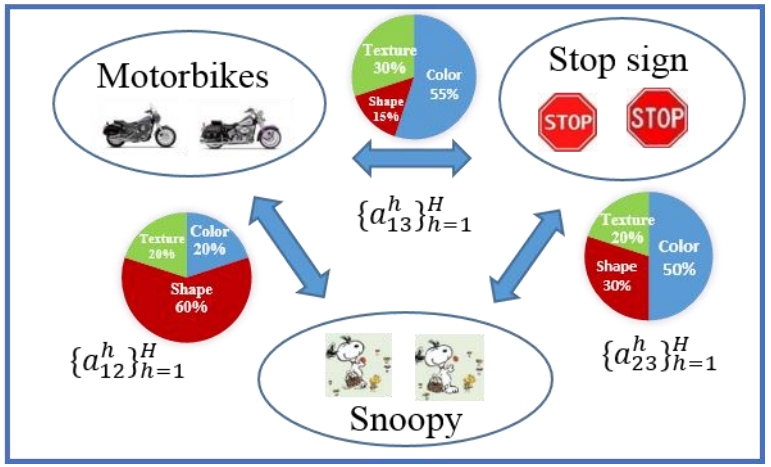


- Extract various features from images and obtain multiple views for describing images
- Build a similarity graph for each view
- Fuse multiple graphs into a new graph and conduct image clustering on this graph

3.1 A group-aware multi-view fusion approach for image clustering

Motivation





Unified fusion weights cannot accurately capture the similarity between images

If the fusion weight can be adjusted according to the characteristics of the image, then more accurate fusion results can be obtained

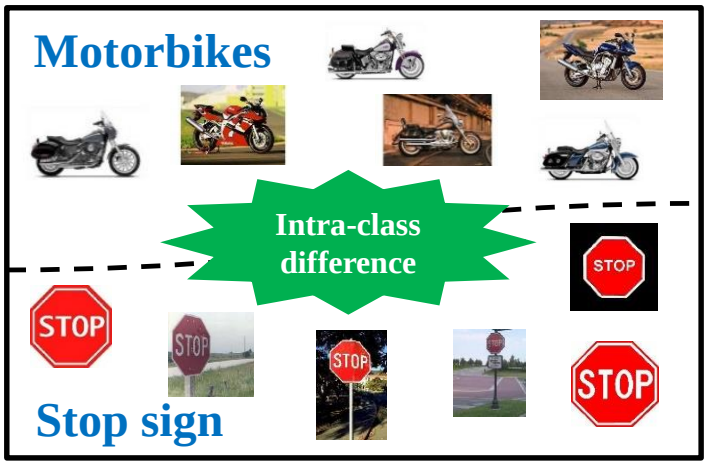
$$W = \sum_{h=1}^{H} a^h W^h$$

$$W_{pq} = \sum_{h=1}^{H} a_{ij}^h W_{pq}^h$$

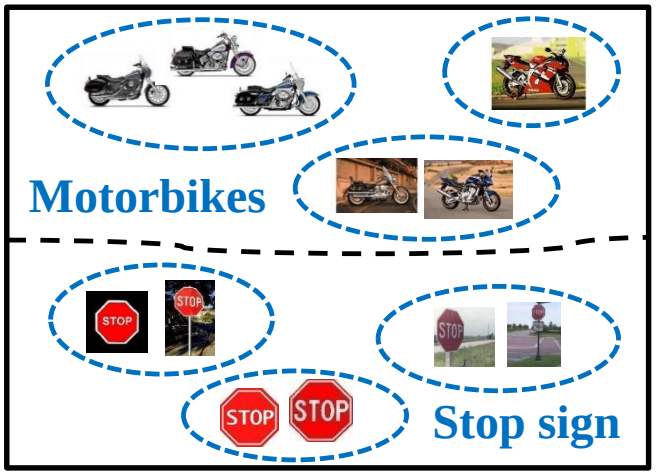
- An image clustering method based on group-aware multi-view fusion (**GOMES**) is proposed

3.1 A group-aware multi-view fusion approach for image clustering

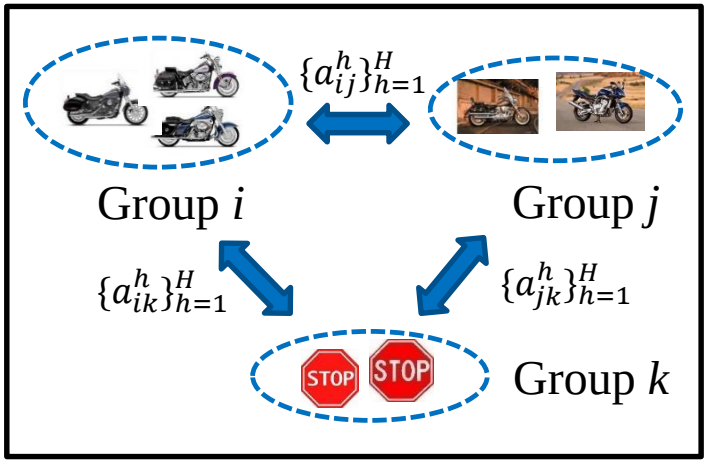
Key steps of GOMES



grouping



fusion



- The same group of images share similar visual properties
- The fusion weight a_{ij}^h reflects the importance of view h when describing the similarity between image group i and group j
- The key issue is to solve the fusion weight a_{ij}^h and obtain image groups

Weight learning

- More accurate and reliable views should be assigned with higher fusion weights
- For any two groups i and j, two criterions are proposed to learn the fusion weights $\{a_{ij}^h\}_{h=1}^H$

The consensus criterion:

$$Con(h,i,j) = \frac{||F_i^t F_j^{t^T} - F_i^h F_j^{h^T}||^2}{\sum_{h=1}^H ||F_i^t F_j^{t^T} - F_i^h F_j^{h^T}||^2}$$

- F_i^t and F_i^h are cluster indicator matrices of group i which are obtained by conducting NNSC on the fused graph and the h-th view graph respectively
- The view h with smaller value of Con is more important

The discrimination criterion:

- Similarity of two groups i and j is defined as:

$$Sim(h,i,j) = \sum_{p\in i}\sum_{q\in j} W_{pq}^h$$

- The most discriminative view d is found by:

$$d = \begin{cases} arg\max_h Sim(h,i,j) & if\ lb(i) = lb(j) \\ arg\min_h Sim(h,i,j) & otherwise \end{cases}$$

- The cost function is defined as:

$$Dis(h,i,j) = \begin{cases} 0 & if\ h = d \\ 1 & otherwise \end{cases}$$

Weight learning

The two criterions are integrated together and the final optimization problem is formulized as:

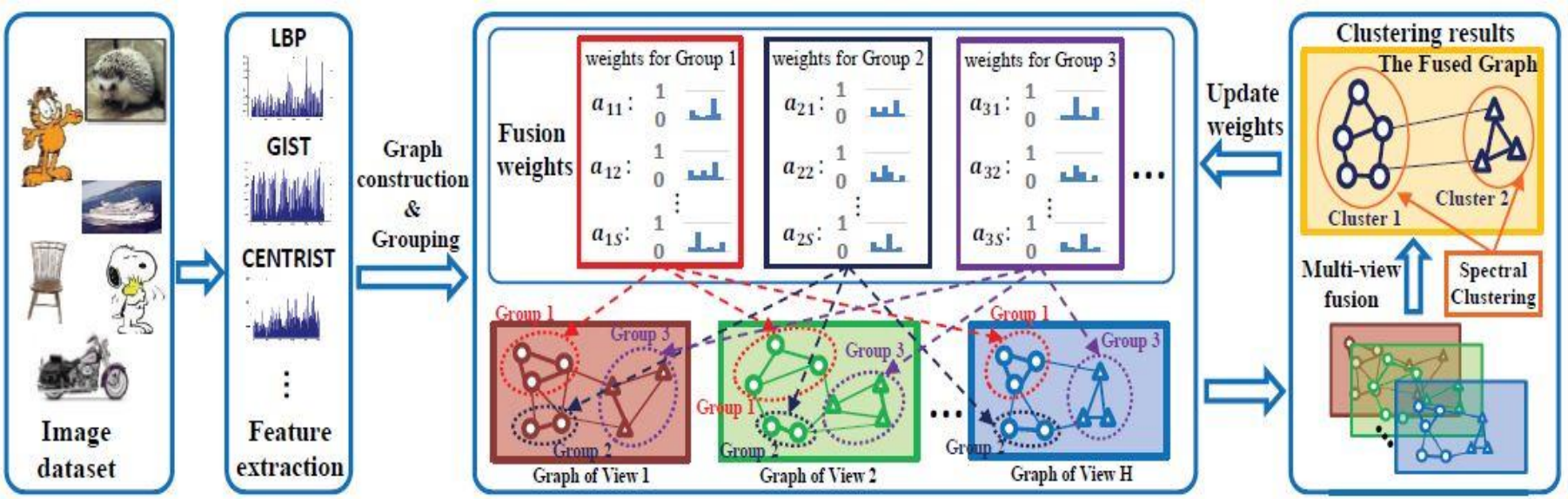
$$\min \sum_{i=1}^{S}\sum_{j=i}^{S}\sum_{h=1}^{H}(a_{ij}^{h})^{r}\left[\beta Con(h,i,j)+(1-\beta)Dis(h,i,j)\right]$$

$$s.t. \quad \sum_{h=1}^{H} a_{ij}^{h}=1,\ 1\le i\le j\le S$$

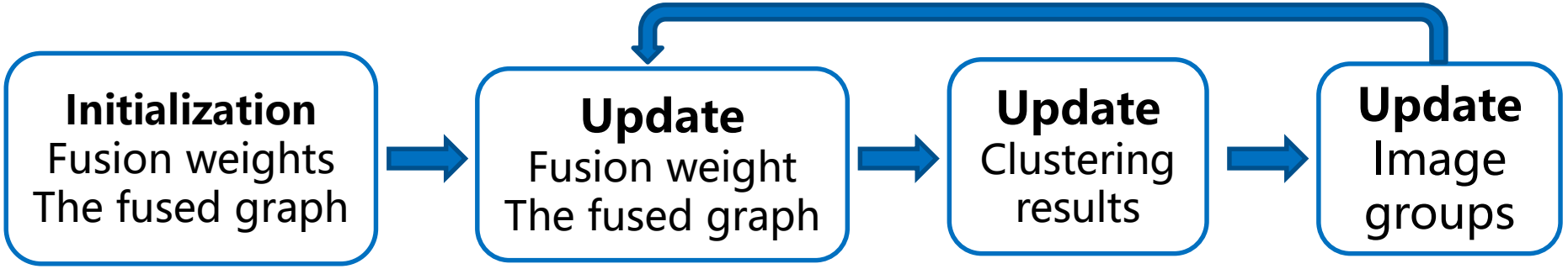
$$a_{ij}^{h}\ge 0,\ 1\le i\le j\le S,\ h=1,...,H$$

- The parameter $\beta \in [0,1]$ provides a tradeoff between the two criterions. $r \in [1,\infty]$ is the parameter to control the sparseness of the solution
- The above problem could be effectively solved by Lagrangian multiplier method

3.1 A group-aware multi-view fusion approach for image clustering



The framework of the proposed method



The flowchart of the algorithm

北京郵電大學
Beijing University of Posts and Telecommunications

Experiments

We compare GOMES with the following methods:

- Single view spectral clustering (**SC**(#)): original spectral clustering method using single view graph
- Equally combining affinity matrices spectral clustering (**EASC**)
- Multi-modal spectral clustering (**MMSC**)
  - Xiao Cai, Feiping Nie, Heng Huang, and Farhad Kamangar, "Heterogeneous image feature integration via multi-modal spectral clustering," in CVPR, 2011.
- Affinity aggregation spectral clustering (**AASC**)
  - Hsin-Chien Huang, Yung-Yu Chuang, and Chu-Song Chen, "Affinity aggregation for spectral clustering," in CVPR, 2012.
- Multi-feature spectral clustering with minimax optimization (**MSCMO**):
  - Hongxing Wang, Chaoqun Weng, and Junsong Yuan, "Multifeature spectral clustering with minimax optimization," in CVPR, 2014.

3.1 A group-aware multi-view fusion approach for image clustering

To be fair, we select four datasets that are adopted in **MMSC**, **AASC** and **MSCMO** as our datasets: **Caltech-101** (**7 and 20 classes**), Microsoft Research Cambridge Volume 1 (**MSRC**) and **Oxford Flowers**.

Dataset	Sample Number	Class Number	Feature
Caltech-7	441	7	LBP/GIST/SIFT/ HOG/CENTRIST
Caltech-20	1230	20	LBP/GIST/SIFT/ HOG/CENTRIST
MSRC	210	7	LBP/GIST/SIFT/ HOG/CENTRIST
Oxford Flowers	1360	17	Color/shape/ texture

Statistics of each dataset

Caltech dataset

MSRC dataset

Example images from two datasets

3.1 A group-aware multi-view fusion approach for image clustering

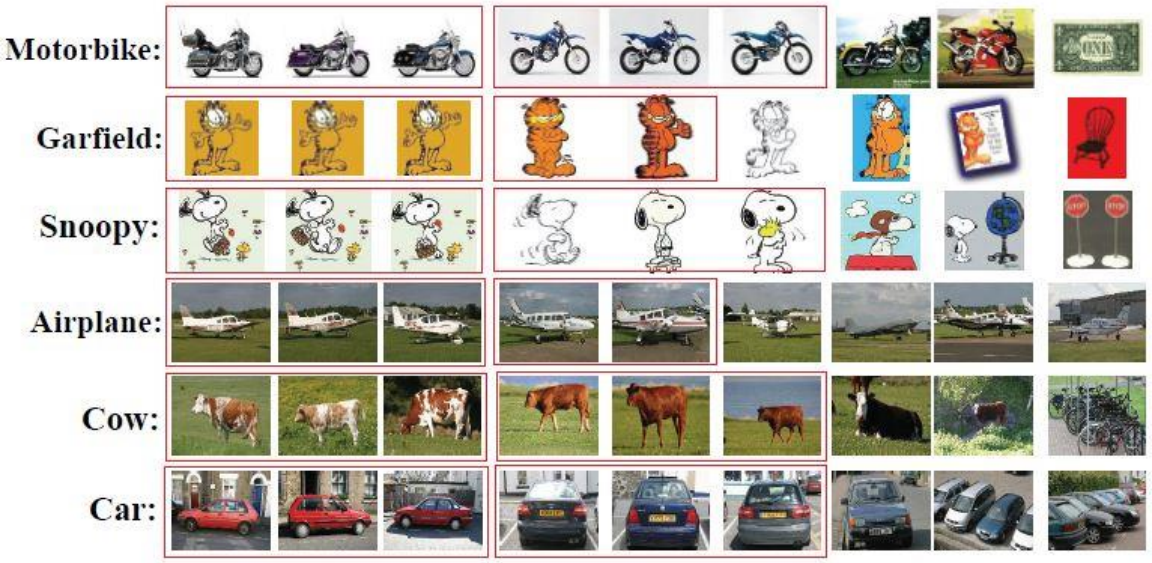
Clustering results on different datasets

Method	Caltech-101 (7 classes)			Caltech-101 (20 classes)			MSRC			Oxford Flowers		
	AMI	NMI	ARI	AMI	NMI	ARI	AMI	NMI	ARI	AMI	NMI	ARI
SC(1)	0.4583	0.4781	0.4020	0.4241	0.4743	0.2861	0.4466	0.4976	0.3546	0.3403	0.3658	0.1753
SC(2)	0.5601	0.5813	0.4448	0.5335	0.5651	0.3644	0.4888	0.5673	0.3478	0.3782	0.4121	0.1976
SC(3)	0.5112	0.5296	0.4416	0.5105	0.5427	0.3276	0.4909	0.5847	0.3078	0.1438	0.2049	0.0538
SC(4)	0.5397	0.5693	0.4372	0.4655	0.5048	0.2881	0.4554	0.5008	0.3529	–	–	–
SC(5)	0.4629	0.4869	0.3540	0.5101	0.5529	0.3529	0.5033	0.5404	0.4040	–	–	–
EASC	0.6355	0.6544	0.5551	0.5880	0.6220	0.4421	0.7332	0.7540	0.6585	0.3896	0.4145	0.2170
MMSC	N/A	0.6792	N/A	N/A	0.6329	N/A	N/A	0.7745	N/A	N/A	0.4270	N/A
AASC	0.6747	0.6853	0.6692	0.6202	0.6458	0.5110	0.7588	0.7806	0.7244	0.4031	0.4291	0.2363
MSCMO	0.6825	0.6922	0.6428	0.5965	0.6331	0.4164	0.6890	0.7166	0.6116	N/A	0.4840	N/A
GOMES	**0.7365**	**0.7456**	**0.6896**	**0.6852**	**0.7044**	**0.5713**	**0.8694**	**0.8770**	**0.8578**	**0.4870**	**0.5069**	**0.3351**

- All the multi-view clustering methods achieve better clustering performance than single view methods
- GOMES accurately learns the fusion weights and achieves the best clustering performance on each dataset compared with all the baseline methods

3.1 A group-aware multi-view fusion approach for image clustering

Clustering results



Some clustering results from Caltech-7 and MSRC. The images belonging to the same group are put in the red box.

- The images belong to the same group share similar visual properties such as background (Garfield and Snoopy) and viewpoints (Airplane and Car)
- GOMES can capture the intra-class variance and generate more accurate description

This work has been published in *ICME 2015* and *Information Sciences 2019*

Background & Motivation

Objective: Learning low-dimensional and discriminative data representation for high-dimensional multi-view data

Problem definition:

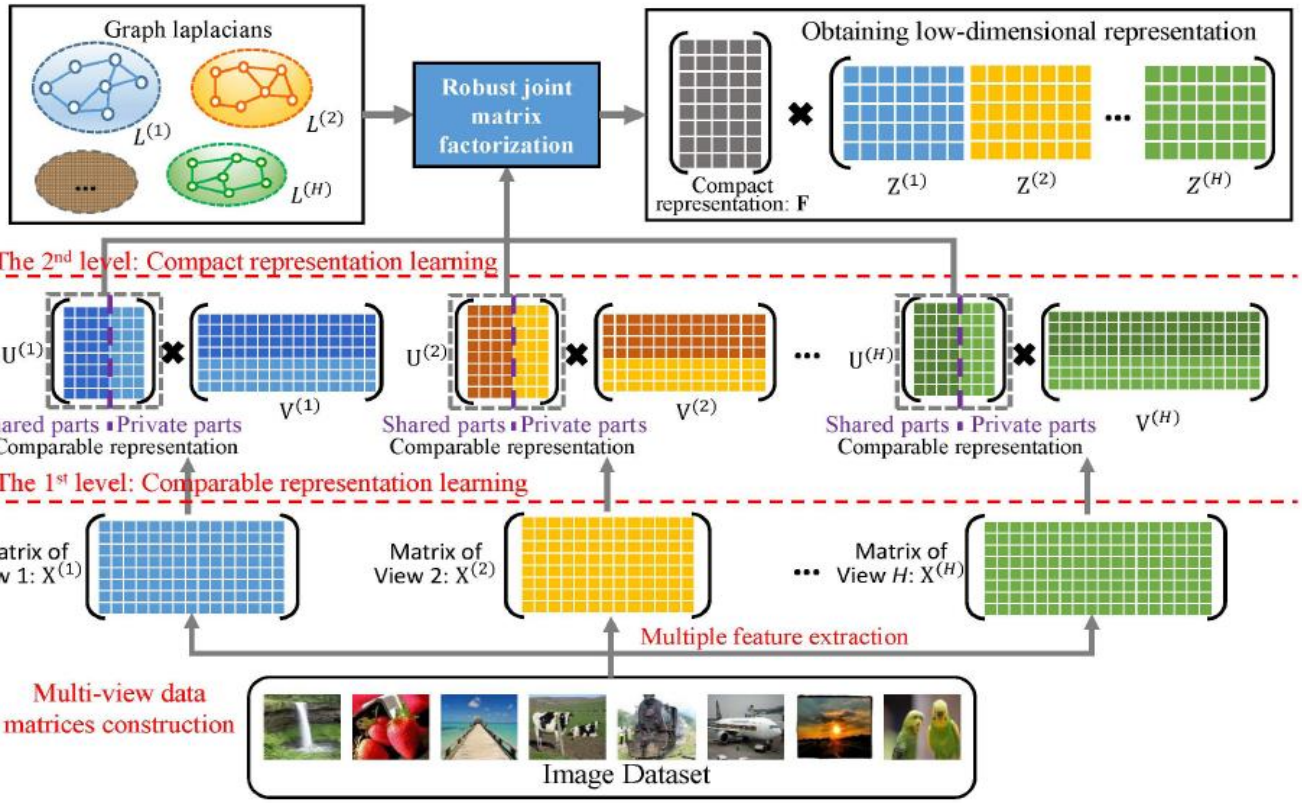
- Given N images with H views: $\{X^{(i)} \in R^{N \times M_i}\}_{i=1}^{H}$, we aim to learn the low-dimensional representation for images: $\mathrm{F} \in R^{N \times R}$, where $R < M_i$

Motivation:

- Different views have different physical meanings, and they cannot be compared directly. Therefore, it is necessary to learn comparable representation
- The information of each view has share part and independent part. Only considering the two parts of information, can we accurately capture the information of each view
- The low-dimensional representation should preserve the information of each view and robust to noise

3.2 Bi-level multi-view latent space learning

The proposed framework



Second level: Low-dimensional representation learning

First level: Comparable representation learning

Input: Multi-view data

Merits: Based on the first level, the heterogeneous multi-view features are made comparable with each other. In the second level, our method can overcome the influence of noise in multi-view data, making the learned low-dimensional representation more effective

3.2 Bi-level multi-view latent space learning

The first level：Comparable representation learning

- Learning comparable representation $U^{(i)} \in R^{N\times K}$ for each view, both shared and private part of each view are considered

$$\mathcal{O}_{cp}(\mathbf{U}^{(i)}, \mathbf{V}^{(i)}) = \min \sum_{i=1}^{H} \left(\|\mathbf{X}^{(i)} - (\mathbf{U}_S^{(i)}, \mathbf{U}_P^{(i)})(\mathbf{V}_S^{(i)}, \mathbf{V}_P^{(i)})^T\|_F^2 + \eta \|\mathbf{V}_P^{(i)}\|_{2,1} \right) + \lambda \sum_{i=1}^{H-1} \sum_{j=i+1}^{H} \|\mathbf{U}_S^{(i)} - \mathbf{U}_S^{(j)}\|_F^2$$

$$s.t.\ \mathbf{U}^{(i)} \geq 0,\ \mathbf{V}^{(i)} \geq 0,\ \forall i = 1, 2, ..., H,$$

The shared parts of basis and coefficient matrix

The private parts of basis and coefficient matrix

The constraint imposed on the private part

The consensus constraint imposed on shared part

3.2 Bi-level multi-view latent space learning

The second level: low-dimensional representation learning

- Based on the learned comparable representation, the low-dimensional representation $F \in R^{N \times R}$ is learned. Considering the importance of different views and noise in data, we propose the following objective function:

$$\mathcal{O}_{pt}(\mathbf{U}^{(i)}, \mathbf{F}, \mathbf{Z}^{(i)}, \gamma_i) = \min \sum_{i=1}^{H} \gamma_i^P \left[\left\| \mathbf{U}^{(i)} - \mathbf{F}\mathbf{Z}^{(i)} \right\|_{2,1} + \beta Tr(\mathbf{F}^T \mathbf{L}^{(i)} \mathbf{F}) \right]$$

$$s.t.\ \mathbf{F} \geq 0, \mathbf{Z}^{(i)} \geq 0, \gamma_i \geq 0, \forall i = 1, ..., H, \sum_{i=1}^{H} \gamma_i = 1,$$

The ultimate objective function:

$$\mathcal{O}_U(\mathbf{U}^{(i)}, \mathbf{V}^{(i)}, \mathbf{F}, \mathbf{Z}^{(i)}, \gamma_i) = \mathcal{O}_{cp}(\mathbf{U}^{(i)}, \mathbf{V}^{(i)}) + \mathcal{O}_{pt}(\mathbf{U}^{(i)}, \mathbf{F}, \mathbf{Z}^{(i)}, \gamma_i).$$

3.2 Bi-level multi-view latent space learning

Experiments

- **Dataset**

✓ NUS dataset and PASCAL VOC'07 dataset

- **Experimental settings**

✓ 5 visual features are extracted: Color moments, SIFT, HOG, GIST, LBP

✓ Compared methods： MSCMO (CVPR'14), GSMVPA (TIP'14), SSMVD (TCSVT'12), CONMF (WWW'14), EMLSRA (PR'15)

✓ Based on the learned low-dimensional representation, we conduct image classification to verify the effectiveness of each method

NUS dataset

VOC dataset

3.2 Bi-level multi-view latent space learning

The classification results with 1000 training samples and dimensionality is 100:

(a) Performance comparison on NUS dataset.

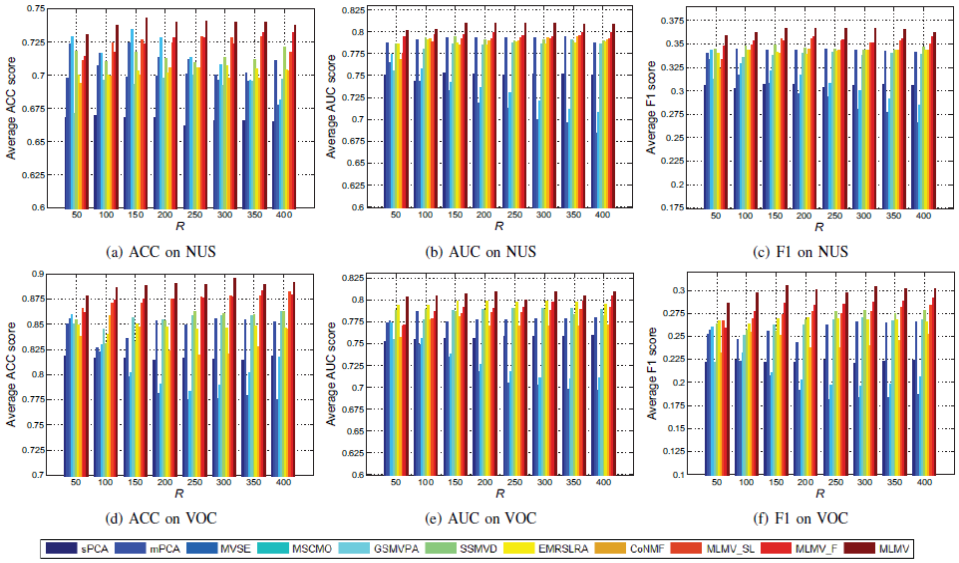
Method	ACC Score	AUC Score	F1 Score
sPCA	0.669 ± 0.004	0.744 ± 0.005	0.303 ± 0.004
mPCA	0.707 ± 0.007	0.791 ± 0.003	0.345 ± 0.003
MVSE	0.701 ± 0.012	0.749 ± 0.003	0.317 ± 0.006
MSCMO	0.716 ± 0.010	0.758 ± 0.004	0.330 ± 0.003
GSMVPA	0.696 ± 0.006	0.780 ± 0.003	0.336 ± 0.004
SSMVD	0.710 ± 0.005	0.793 ± 0.003	0.348 ± 0.004
EMRSLRA	0.700 ± 0.006	0.792 ± 0.002	0.343 ± 0.004
CoNMF	0.699 ± 0.014	0.793 ± 0.008	0.343 ± 0.005
BLMV_SL	0.724 ± 0.013	0.788 ± 0.008	0.349 ± 0.007
BLMV_F	0.717 ± 0.010	0.797 ± 0.006	0.354 ± 0.005
BLMV	$\mathbf{0.737 \pm 0.007}$	$\mathbf{0.803 \pm 0.006}$	$\mathbf{0.363 \pm 0.006}$

(b) Performance comparison on VOC dataset.

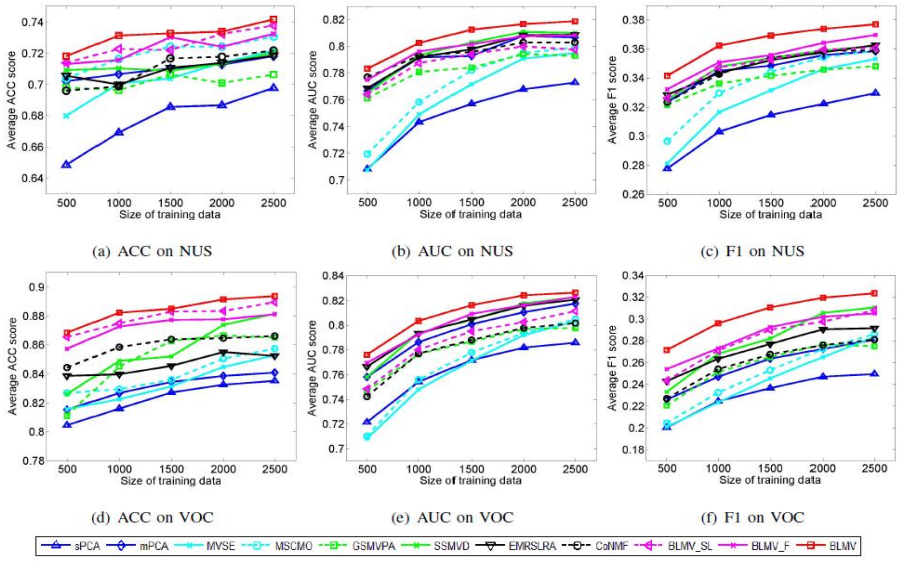
Method	ACC Score	AUC Score	F1 Score
sPCA	0.816 ± 0.011	0.754 ± 0.007	0.224 ± 0.005
mPCA	0.827 ± 0.006	0.786 ± 0.003	0.247 ± 0.004
MVSE	0.823 ± 0.005	0.748 ± 0.004	0.223 ± 0.004
MSCMO	0.830 ± 0.009	0.756 ± 0.006	0.232 ± 0.004
GSMVPA	0.845 ± 0.010	0.777 ± 0.007	0.251 ± 0.003
SSMVD	0.849 ± 0.006	0.794 ± 0.003	0.268 ± 0.002
EMRSLRA	0.840 ± 0.005	0.793 ± 0.004	0.263 ± 0.005
CoNMF	0.859 ± 0.007	0.777 ± 0.003	0.254 ± 0.003
BLMV_SL	0.871 ± 0.008	0.778 ± 0.005	0.269 ± 0.005
BLMV_F	0.873 ± 0.006	0.786 ± 0.003	0.277 ± 0.006
BLMV	$\mathbf{0.886 \pm 0.005}$	$\mathbf{0.804 \pm 0.004}$	$\mathbf{0.297 \pm 0.006}$

- Compared with other methods, the performance of our method BLMV is better than the other methods
- BLMV_SL is removing the first level of BLMV. This model cannot accurately capture the shared and private components of each view, so it cannot accurately encode the information of each view
- BLMV_F uses F-norm instead of $L_{2,1}$ norm in BLMV. This model cannot handle the noise in data, so the performance is degraded

3.2 Bi-level multi-view latent space learning



Classification results for different dimensions R



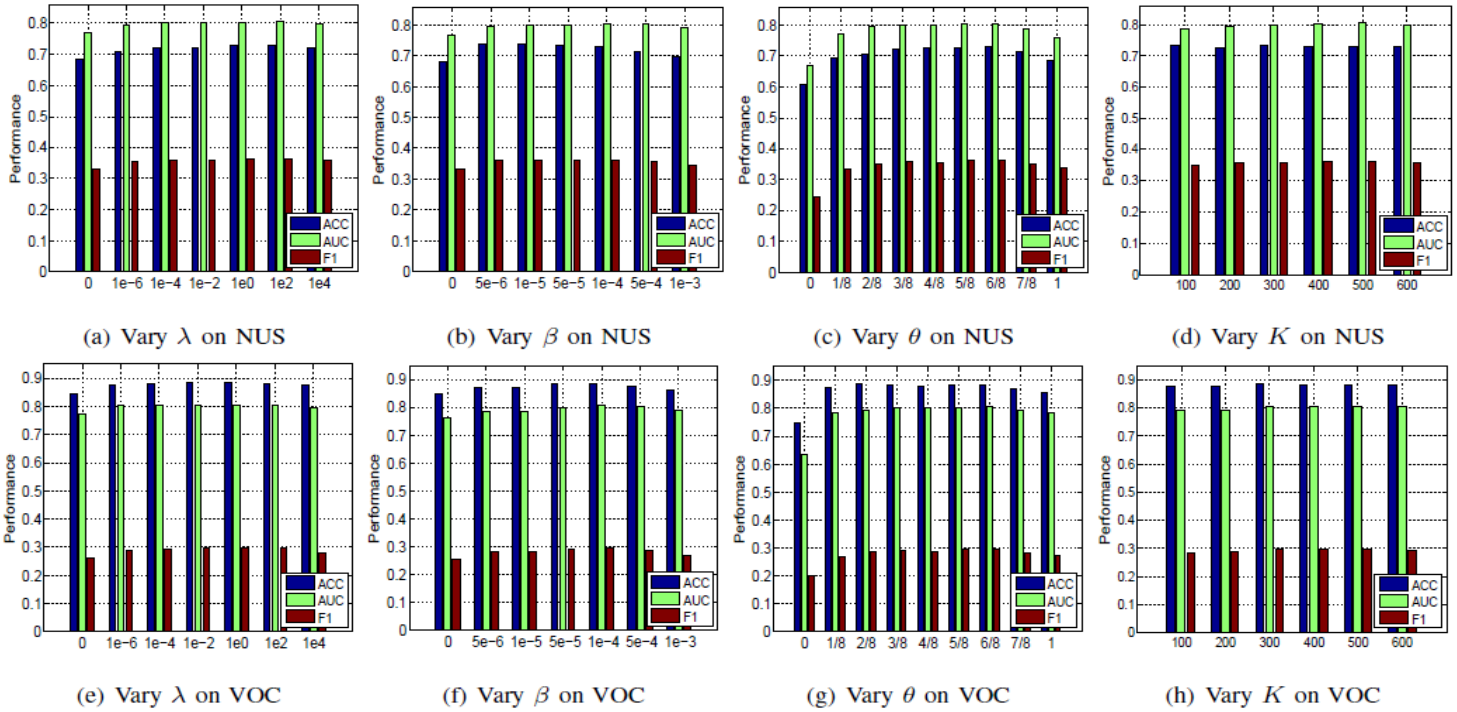
Classification results for different number of training samples

- Our method achieves better classification performance than other methods for different dimensions, and our method is insensitive to the dimensions
- For different training samples, our method also achieves better results. As the number of training samples increases, the classification performance improves steadily

北京郵電大學
Beijing University of Posts and Telecommunications

3.2 Bi-level multi-view latent space learning

Parameter sensitivity



(a) Vary λ on NUS (b) Vary β on NUS (c) Vary θ on NUS (d) Vary K on NUS

(e) Vary λ on VOC (f) Vary β on VOC (g) Vary θ on VOC (h) Vary K on VOC

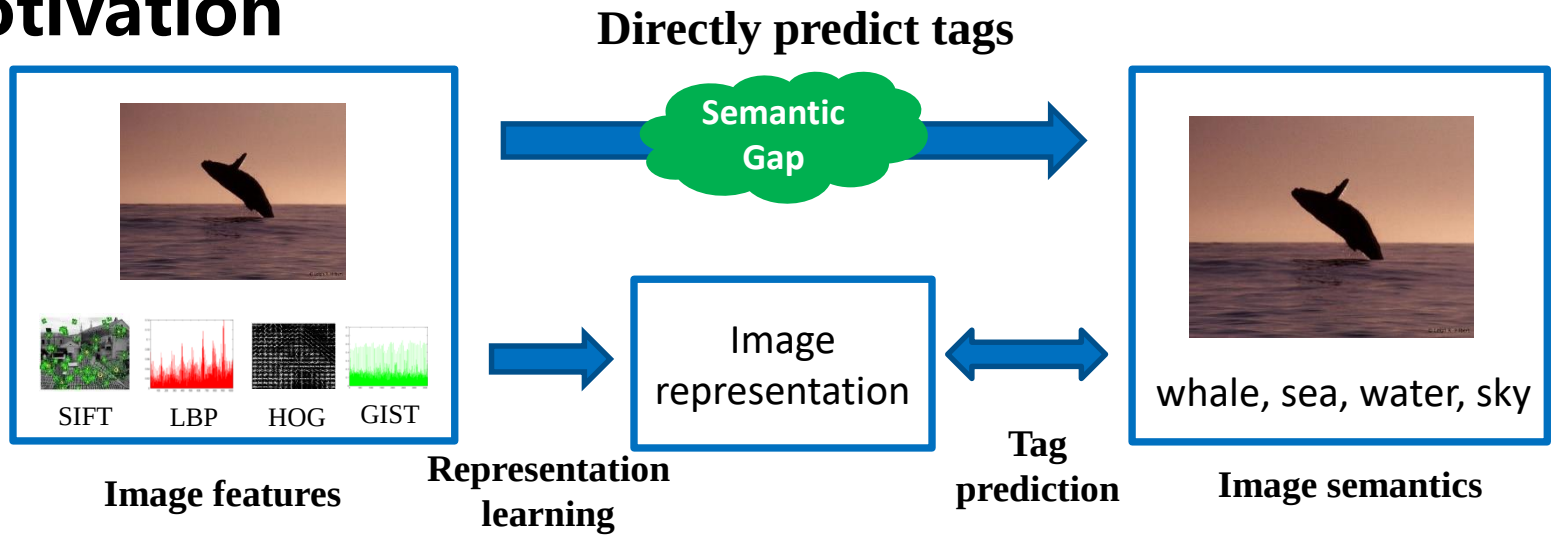
- In general, our method is insensitive to parameters and it can achieve good learning performance in a wide range

This work has been published in *IEEE Transactions on Circuits and Systems for Video Technology 2018*

北京邮电大学 Beijing University of Posts and Telecommunications

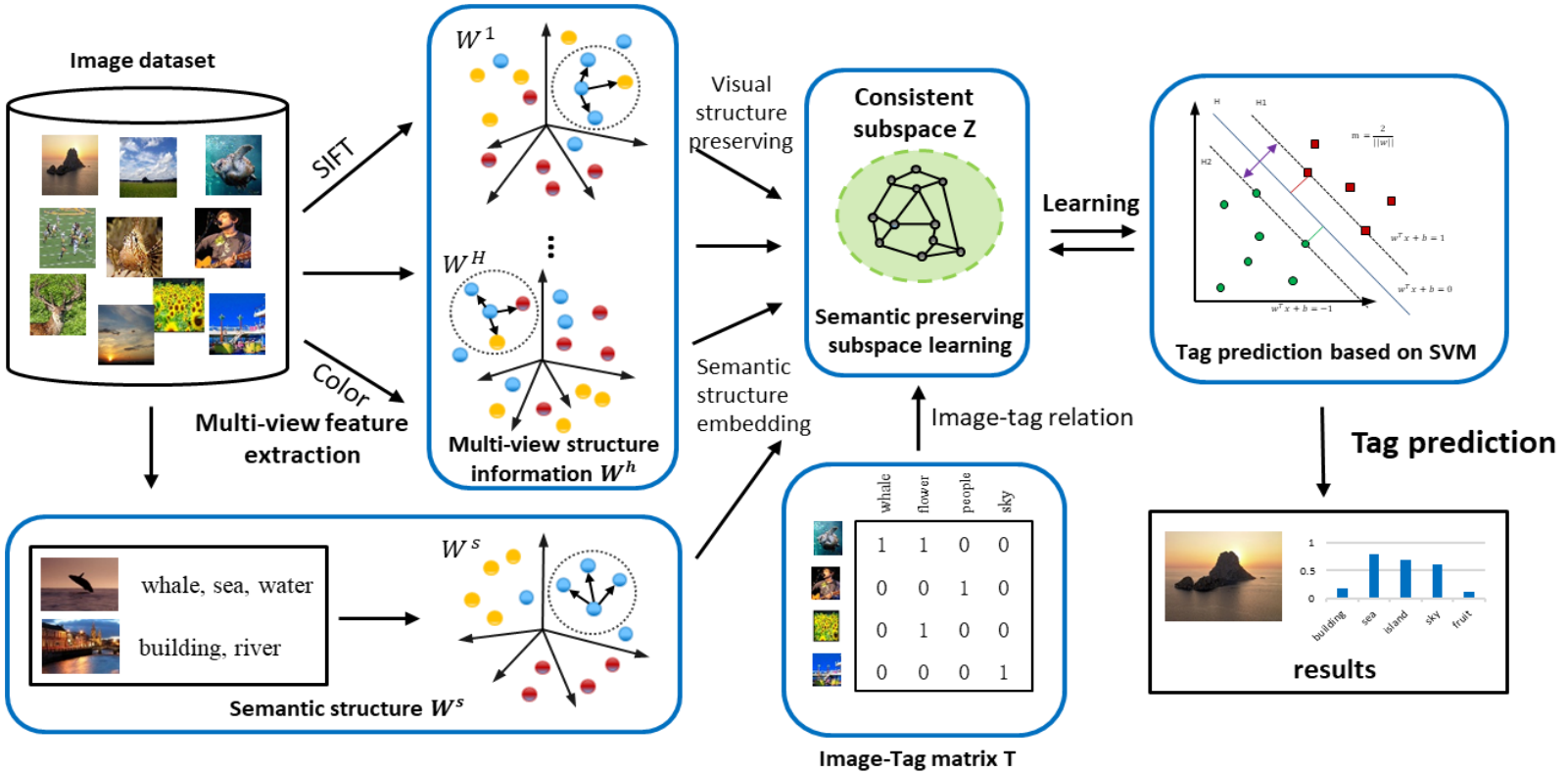
3.3 Joint multi-view representation learning and image tagging

Motivation



- It is difficult to accurately predict the image tags based on the original image features. We hope to learn a more suitable image representation for image annotation tasks
- Image representation learning and image annotation are two closely related tasks: proper image representation can predict labels more accurately, and label prediction can guide the learning of representation
- If we integrate image representation learning and label prediction, and make two tasks promote each other, then the performance of label prediction can be improved

3.3 Joint multi-view representation learning and image tagging



The proposed framework

- Merits: By integrating multi-view representation learning and label prediction into a unified framework and utilizing the correlation between the two tasks, the discriminating power of classifiers and the accuracy of image representation can be improved

3.3 Joint multi-view representation learning and image tagging

The proposed method

- **Subproblem 1：Semantic preserved multi-view subspace learning**

$$\min\ f(Z)=\frac{1}{\gamma}\log\left\{\sum_{h=1}^{H}\exp\left[\gamma\,\|\,Z-(W^h\odot W^s)Z\,\|_F^2\right]\right\}+\eta\,\|\,T-ZZ^{\mathrm{T}}T\,\|_F^2$$

$$s.t.\ Z^TZ=I$$

Softmax function is used to find the views with higher embedding losses, so that multi-view information can be fully preserved

Semantic information guided visual information preserving

Semantic information preserving

- **Subproblem 2: Image tag prediction**

$$\min_{Z}\max_{\alpha_t} g(Z,\alpha_t)=\sum_{t=1}^{m}\left[\alpha_t^{\mathrm{T}}1-0.5Tr\left(ZZ^{\mathrm{T}}Y_t\alpha_t(Y_t\alpha_t)^{\mathrm{T}}\right)\right]$$

$$s.t.\quad \alpha_t^{\mathrm{T}}y_t=0, 0\le\alpha_t\le C, t=1,\ldots,m$$

Joint learning SVM and image representation Z

- **Subproblem 3：Projection function learning**

$$\min h(Z,P)=\|\,XP-Z\,\|_F^2+\beta\,\|\,P\,\|_{2,1}$$

The group sparsity constraint is imposed on P to select discriminative features

- **The objective function** $O(Z,P,\alpha_t)=\min_{Z,P}\max_{\alpha_t}\left(f(Z)+\mu_1 g(Z,\alpha_t)+\mu_2 h(Z,P)\right)$

3.3 Joint multi-view representation learning and image tagging

Experiments

Dataset	View #	Training #	Testing #
Corel5k	7	4500	500
ESP Game	7	18000	2000
NUS	6	10000	3000

- Image annotation datasets Corel5k, ESP Game and NUS-WIDE are used for image annotation experiments
- Evaluation metrics: Precision (P), Recall (R), F-measure (F1), MAP

Id	Compared methods
1	Fast Image Tagging (FastTag) [ICML'13]
2	NMF-KNN: Image Annotation using Weighted Multi-view Non-negative Matrix Factorization (NMF-KNN) [CVPR'14]
3	Image Tag Completion via Image-Specific and Tag-Specific Linear Sparse Reconstructions (LSR) [CVPR'13]
4	Tag Completion for Image Retrieval (TMC) [TPAMI'13]
5	Optimal Graph Learning with Partial Tags and Multiple Features for Image and Video Annotation (OGL) [CVPR'15]
6	Low-Rank Multi-View Learning in Matrix Completion for Multi-Label Image Classification (lrMVL) [AAAI'15]
7	A Closed Form Solution to Multi-View Low-Rank Regression (MVLR) [AAAI'15]
8	The proposed method with equal weight of each view (OPSL-V)

Experimental results

Method	Corel5k				ESP Game				NUS			
	P	R	F1	MAP	P	R	F1	MAP	P	R	F1	MAP
FastTag	32.2	45.7	37.8	25.3	29.0	32.1	30.5	12.2	58.0	26.6	36.5	11.2
NMF-KNN	35.0	49.6	41.0	26.2	28.4	31.6	29.4	13.7	51.6	23.8	32.5	10.5
LSR	33.1	46.8	38.8	24.8	28.5	32.4	30.3	14.9	52.8	24.2	33.2	13.6
TMC	31.7	37.1	33.9	17.3	21.1	23.2	22.1	9.8	39.2	17.9	24.6	9.4
OGL	34.7	49.0	40.7	27.5	31.0	34.1	32.5	**17.0**	57.2	26.2	35.9	13.3
lrMVL	29.9	42.0	34.9	20.4	25.9	28.5	27.1	10.3	48.6	22.3	30.6	9.4
MVLR	25.9	37.2	30.5	16.9	24.5	27.2	25.8	9.5	37.7	17.3	23.7	7.9
OPSL-V	36.0	50.7	42.1	28.8	31.3	34.5	32.8	15.3	59.1	27.1	37.2	13.9
OPSL	**37.0**	**52.1**	**43.3**	**29.6**	**32.3**	**35.6**	**33.9**	16.1	**60.9**	**28.0**	**38.4**	**14.5**

- Compared with other methods, our method OPSL achieves better image annotation performance on all the datasets
- OPSL-V is a comparison method without using SoftMax activation function, and its performance is decreased
- The proposed method OPSL can learn suitable image representation for image annotation task, therefore, better image annotation performance are obtained

Some image annotation results

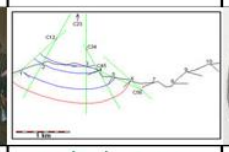
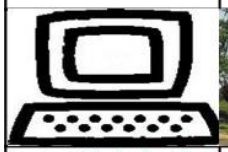
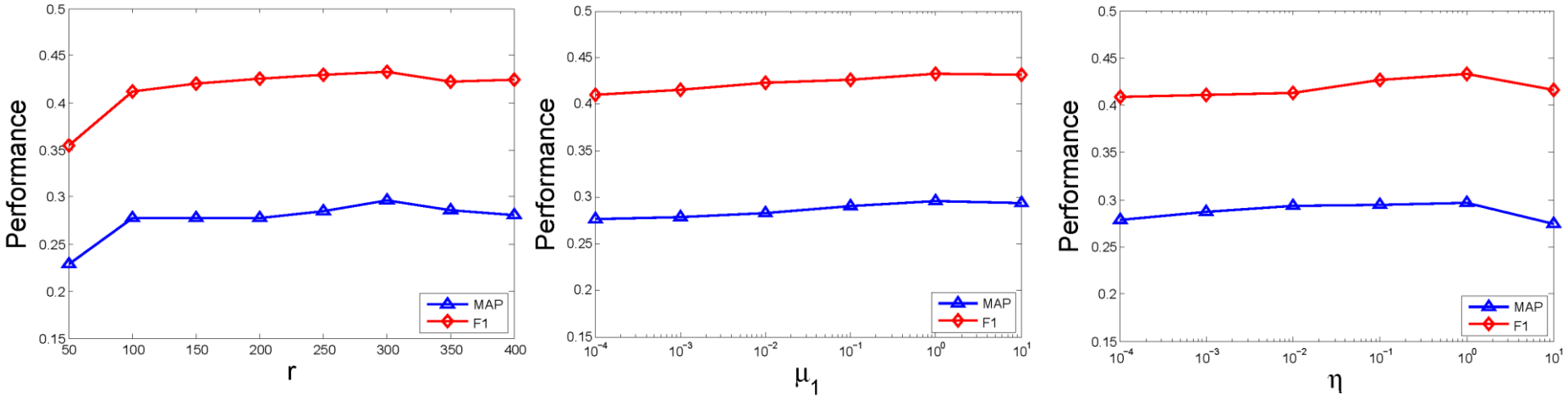
					
red, *logo*, *white*, letter, *black*	people, *group*, man, crowd, *woman*	sky, plane, airplane, red, fly	*circle*, *logo*, *blue*, *round*, *black*	sky, *sand*, desert, *man*, grass	green, tree, leaf, *plant*, sky
letter, red	crowd, man, people	airplane, fly, plane, red, sky, wheel	music, sign, square	brown, desert, grass, sand, sky	green, leaf, sky, tree
					
man, people, *black*, *party*, white,	graph, chart, map, line, green	coin, money, old, silver, round	*red*, band, *man*, light, sing	*red*, *people*, *black*, *cartoon*, *drawing*	black, space, star, white, *tree*
man, people, white	black, chart, graph, green, line, map, red	circle, coin, head, man, money, old, round, silver	band, light, music, sing	hat, man, street	black, planet, space, star, white
					
screen,black,drawing, computer,*logo*	tree, house, building, sky, home	green, leaf, *tree*, plant, pink	circle, *yellow*, green, logo, round	people, man, woman, *hat*, group	sky, *mountain*, tree, green, cloud
black, computer, drawing, screen	building, grass, home, house, light, sky, tree, window	flower, grass, green, leaf, pink, plant	circle, green, logo, round, sign	asian, blue, chinese, group, man, people, photo, woman	city, cloud, green, sky, tree

Image annotation results on ESP GAME dataset. Green tags are the labels that are accurately predicted, red tags are the labels that are incorrectly predicted, and the ground truth tags are in black

3.3 Joint multi-view representation learning and image tagging

Parameter sensitivity



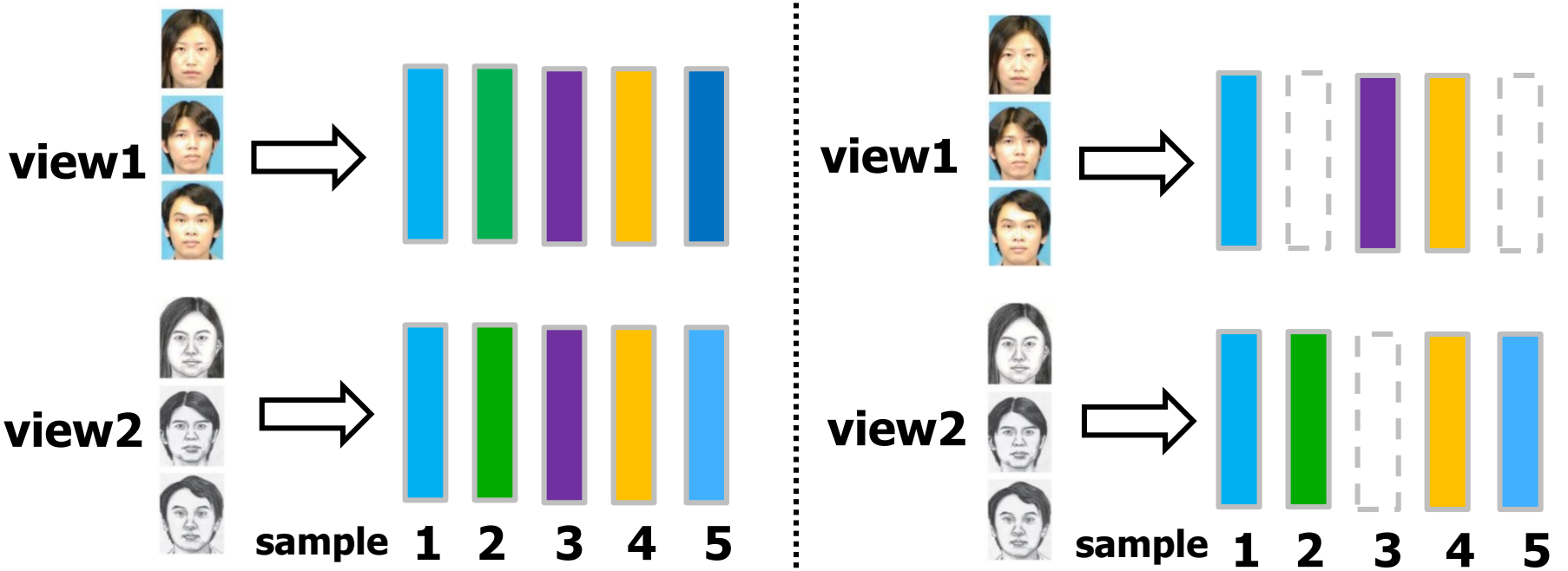
- When the dimension r is greater than 100, the performance do not change significantly
- When increase the value of μ_1 and η, the discriminating power of the learned representation improves
- In general, the proposed method is insensitive to different parameter settings

This work has been published in *AAAI 2016*

北京郵電大學
Beijing University of Posts and Telecommunications

3.4 Deep Correlated Predictive Subspace Learning for Incomplete Multi-View Semi-Supervised Classification

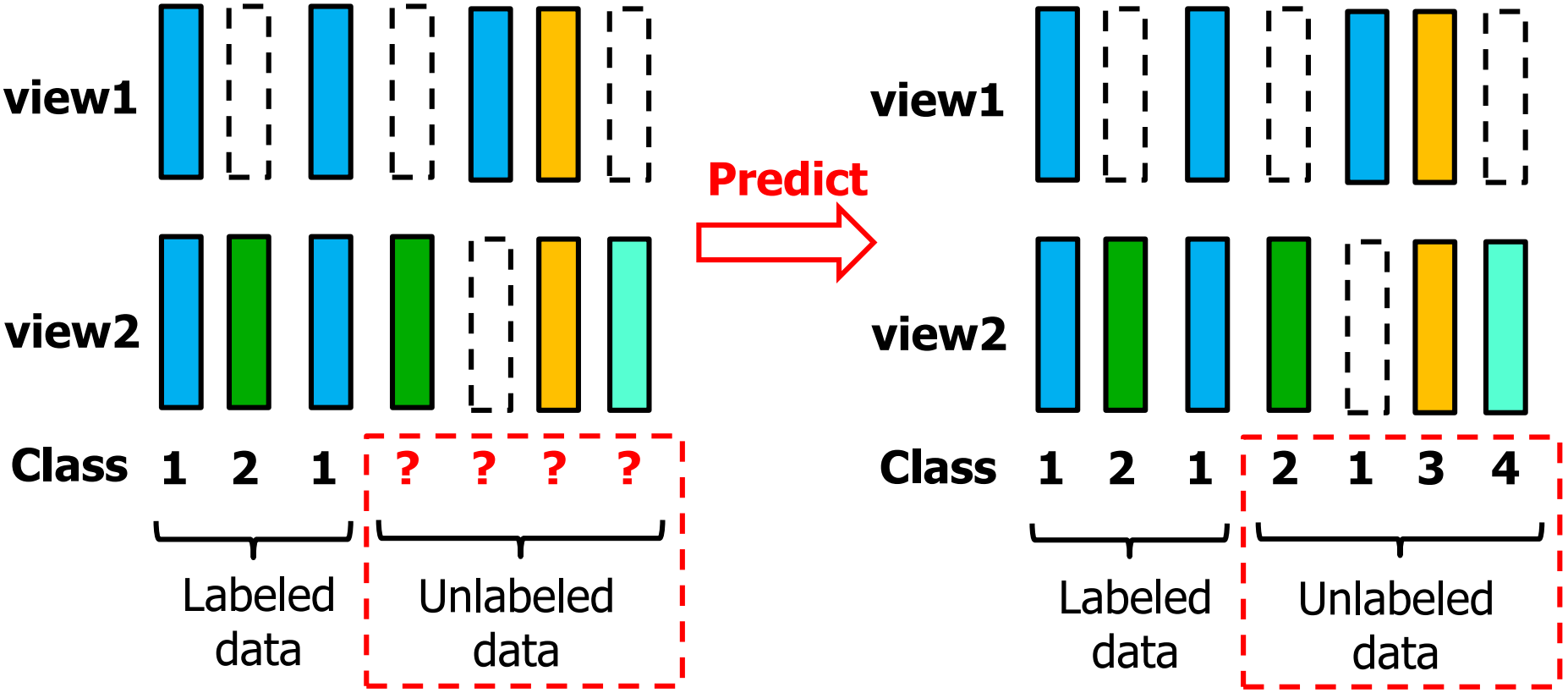
Incomplete multi-view data



- In real-world applications, some views may suffer from several missing samples, resulting in partial multi-view data

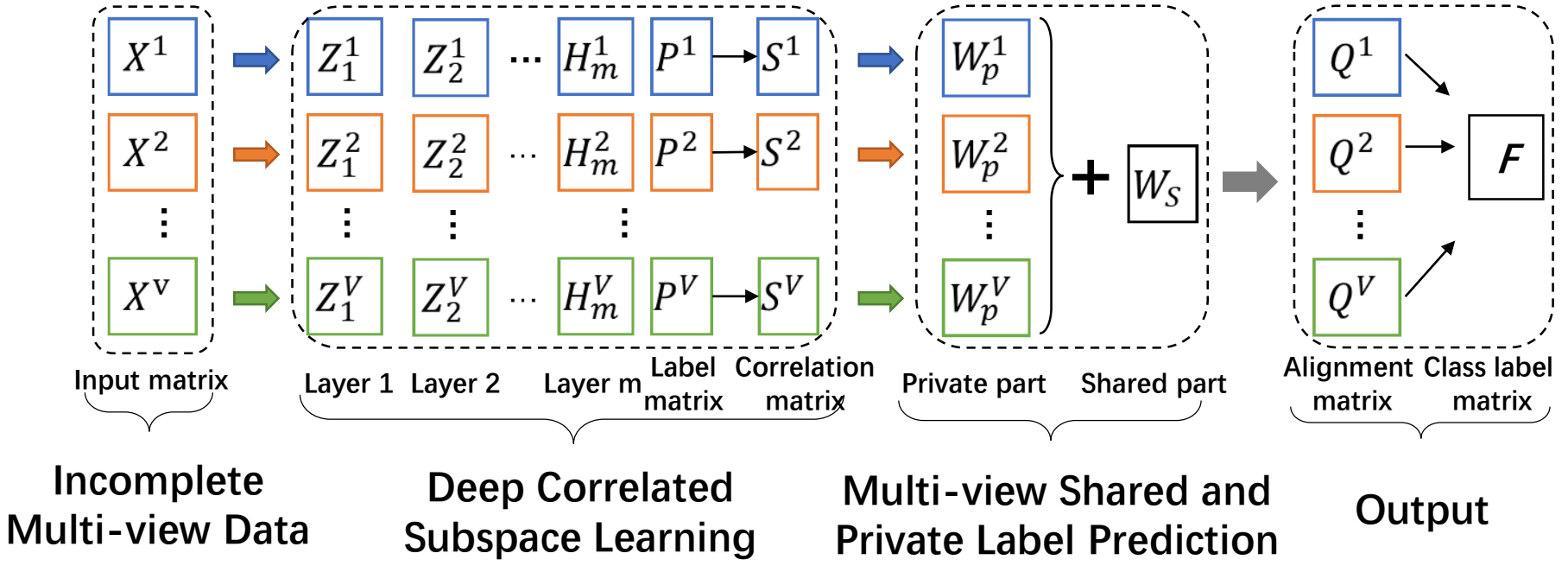
3.4 Deep Correlated Predictive Subspace Learning for Incomplete Multi-View Semi-Supervised Classification

- Objective
  - Given incomplete and partially labeled multi-view data, we expect to predict the labels for unlabeled data



3.4 Deep Correlated Predictive Subspace Learning for Incomplete Multi-View Semi-Supervised Classification

The framework



- The proposed method: Deep Correlated Predictive Subspace Learning (DCPSL)
  - Part 1: Deep Correlated Subspace Learning
  - Part 2: Multi-view Shared and Private Label Prediction
  - Integrate Part 1 and Part 2 into a unified framework

- Data representation
  - We have totally n samples with V views
  - Incomplete multi-view data is represented by $\{X^{(v)} \in R^{d_v \times n_v}\}_{v=1}^{V}$, n_v is the number of samples in the v-th view, d_v is the dimension of the v-th view
  - Due to the incomplete problem, $n_v < n$
  - Given l labeled samples, $Y \in R^{c \times l}$ is the label matrix and c is the number of classes
  - If x_j belongs to class i, then $Y_{ij}=0$, otherwise, $Y_{ij}=1$

3.4 Deep Correlated Predictive Subspace Learning for Incomplete Multi-View Semi-Supervised Classification

(1) Deep Correlated Subspace Learning: J_1

$$\min J_1(Z_i^{(v)}, H_m^{(v)}, S^{(v)})$$

$$= \sum_{v=1}^{V} \{ \underbrace{\| X^{(v)} - Z_1^{(v)} Z_2^{(v)} \ldots Z_m^{(v)} H_m^{(v)} P^{(v)} \|_F^2}_{\text{Label constrained deep matrix factorization}} + \underbrace{\| H_m^{(v)} P^{(v)} - H_m^{(v)} P^{(v)} S^{(v)} \|_F^2 + \alpha \| S^{(v)} \|_*}_{\text{Low-rank subspace learning}}$$

$$s.t.\ H_m^{(v)} \geq 0,\ S^{(v)} \geq 0$$

- In deep matrix factorization, $Z_k^{(v)}$ is the basis matrix of the k-th layer, $H_m^{(v)}$ is the coefficient matrix
- $P^{(v)}$ is the label constraint matrix of the v-th view. If $x_1^{(v)}$ and $x_2^{(v)}$ belong to class 1, $x_3^{(v)}$ and $x_4^{(v)}$ belong to class 2, then $P^{(v)}$ is defined as:

$$P^{(v)} = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & I_{n_v-4} \end{pmatrix}$$

x_1^v x_2^v x_3^v x_4^v Unlabeled samples

(2) Multi-view Shared and Private Label Prediction: J_2

$$\min J_2(W_s, W_p^{(v)}, F)$$

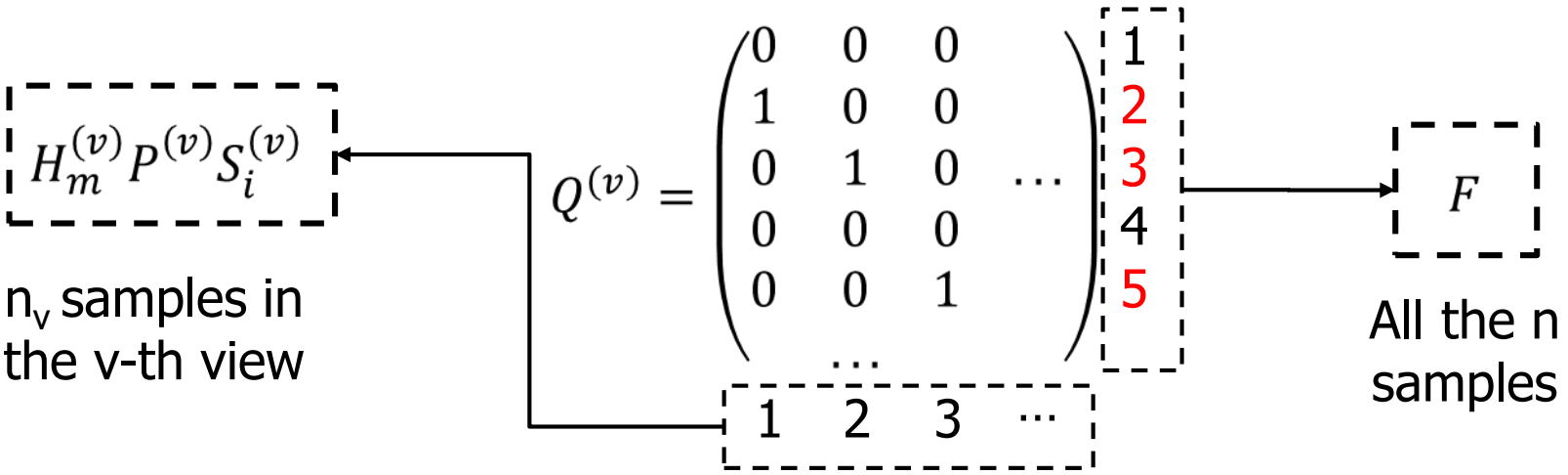
$$= \sum_{v=1}^{V} \{ \underbrace{\| (W_s + W_p^{(v)}) H_m^{(v)} P^{(v)} S^{(v)} - F Q^{(v)} \|_F^2}_{\text{Label prediction}} + \underbrace{\beta_1 \| W_p^{(v)} \|_1 + \beta_2 \| W_s \|_F^2}_{\text{Parameter regularization}} \}$$

$$s.t.\ F_{\pi_l} = Y$$

- The classifier for the v-th view is $W^{(v)} = W_s + W_p^{(v)}$. W_s is the shared part of classifier, $W_p^{(v)}$ is the private part for view v
- $F \in R^{c \times n}$ is the label matrix of all samples to be learned. The labeled samples F are made to be consistent with ground truth Y
- How to establish the relationship between $H^{(v)}P^{(v)}S^{(v)}$ and F?

(2) Multi-view Shared and Private Label Prediction: J_2

- $Q^{(v)} \in R^{n \times n_v}$ is the alignment matrix of the v-th view.
- If the first three sample in $H_m^{(v)} P^{(v)} S^{(v)}$ correspond to the 2nd, 3rd, 5th sample in F, then $Q^{(v)}$ is defined as:



$Q^{(v)}$ is the alignment matrix between $H_m^{(v)} P^{(v)} S_i^{(v)}$ and F

(3) Ultimate Objective Function

- We propose the objective function to jointly conduct deep correlated subspace learning and multi-view shared and private label prediction:

$$\min J(Z_i^{(v)}, H_m^{(v)}, S^{(v)}, W_s, W_p^{(v)}, F)$$

$$= \sum_{v=1}^{n_v} \{ \| X^{(v)} - Z_1^{(v)} Z_2^{(v)} \ldots Z_m^{(v)} H_m^{(v)} P^{(v)} \|_F^2 + \| H_m^{(v)} P^{(v)} - H_m^{(v)} P^{(v)} S^{(v)} \|_F^2$$

$$+ \lambda \| (W_s + W_p^{(v)}) H_m^{(v)} P^{(v)} S^{(v)} - F Q^{(v)} \|_F^2 + \Phi(S^{(v)}, W_p^{(v)}, W_s) \}$$

$$s.t.\ H_m^{(v)} \geq 0,\ S^{(v)} \geq 0, F_{\pi l} = Y,$$

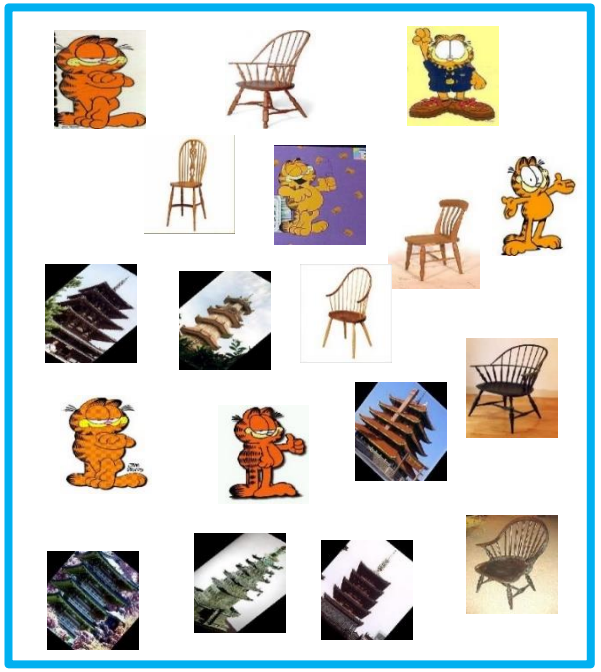
$$where\ \ \Phi(S^{(v)}, W_p^{(v)}, W_s) = \alpha \| S^{(v)} \|_* + \beta_1 \| W_p^{(v)} \|_1 + \beta_2 \| W_s \|_F^2$$

Datasets

- We select four famous image datasets as our datasets: NUS, SUN, Caltech, and Flowers
- Different visual features are extracted to represent image

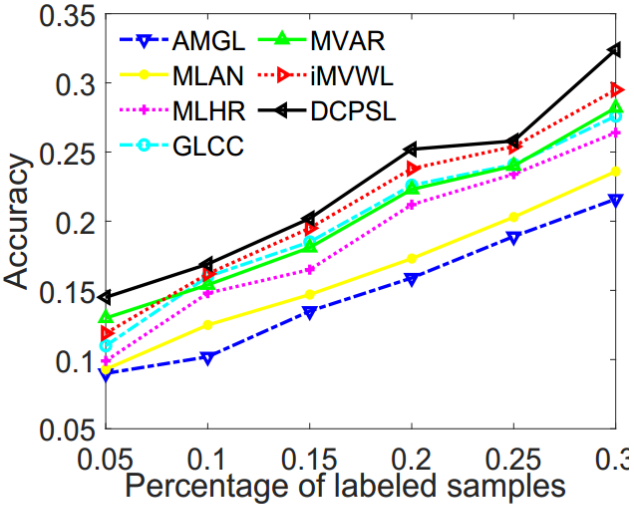
Dataset	Sample Number	Class Number	NO. of Features
NUS	3100	31	5
SUN	3000	30	5
Caltech	1230	20	5
Flowers	1360	17	3

Statistics of each dataset

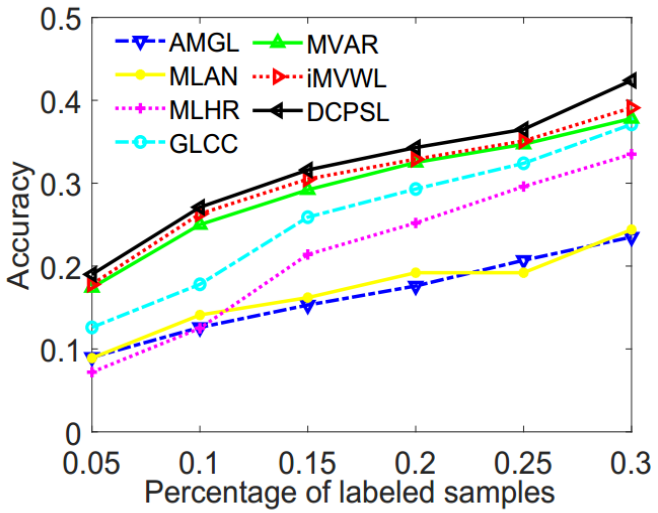


Example images from Caltech dataset

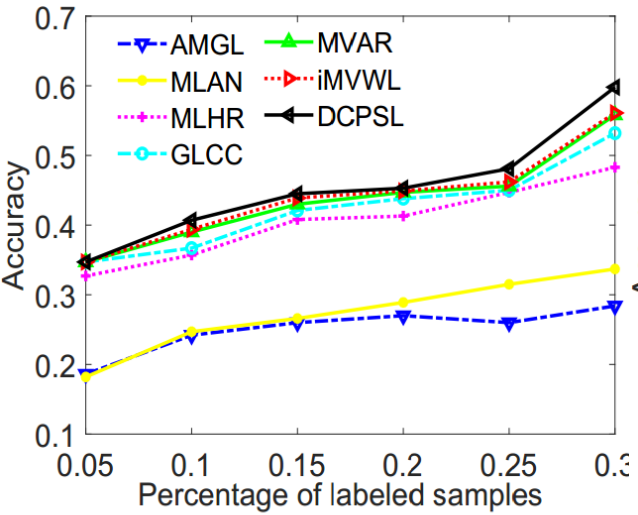
3.4 Deep Correlated Predictive Subspace Learning for Incomplete Multi-View Semi-Supervised Classification



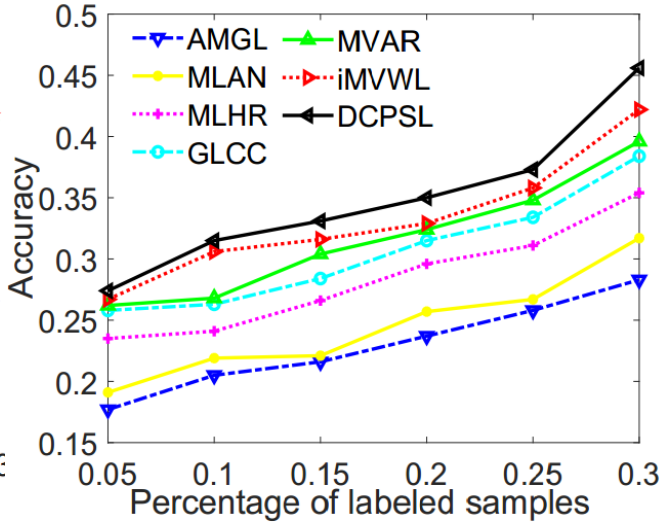
(a) NUS



(b) SUN



(c) Cal20

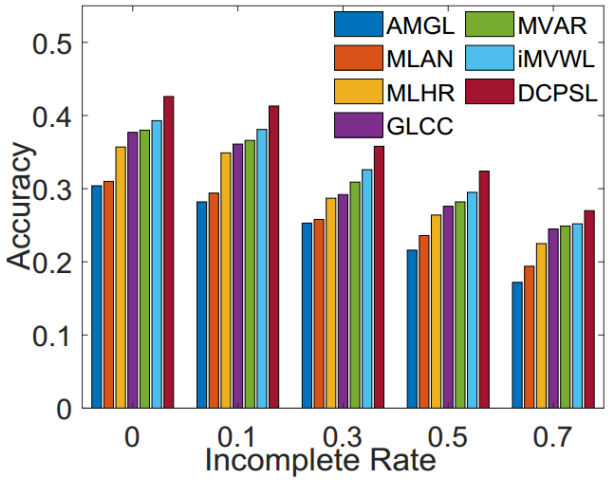


(d) Flowers

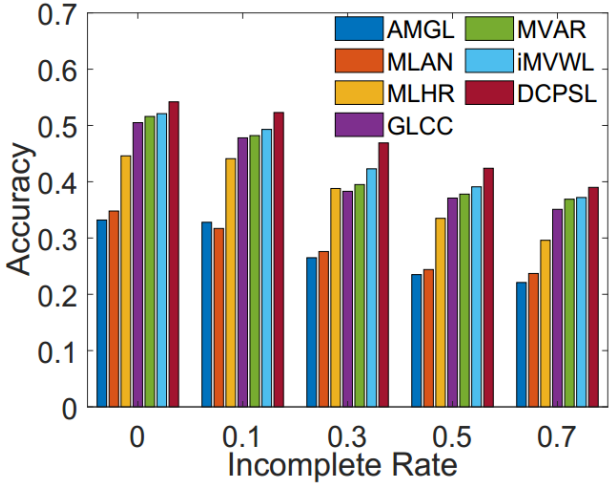
- Fix the incomplete rate to 0.5, and test the classification accuracy with different number of labeled samples
- We can see that DCPSL achieves better classification accuracy compared to all the other methods on each dataset
- The largest performance improvements on the four datasets are: 2:9%, 3:3%, 3:7% and 3:4%, respectively

北京郵電大學
Beijing University of Posts and Telecommunications

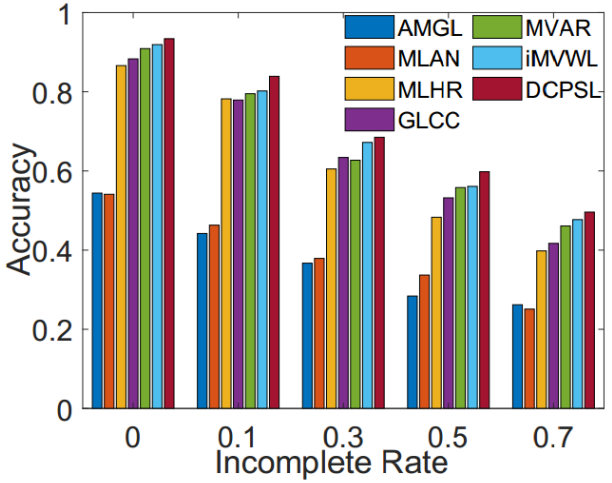
3.4 Deep Correlated Predictive Subspace Learning for Incomplete Multi-View Semi-Supervised Classification



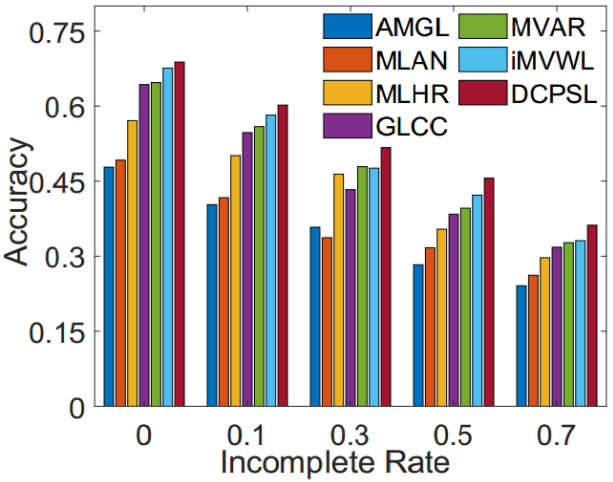
(a) NUS



(b) SUN



(c) Cal20



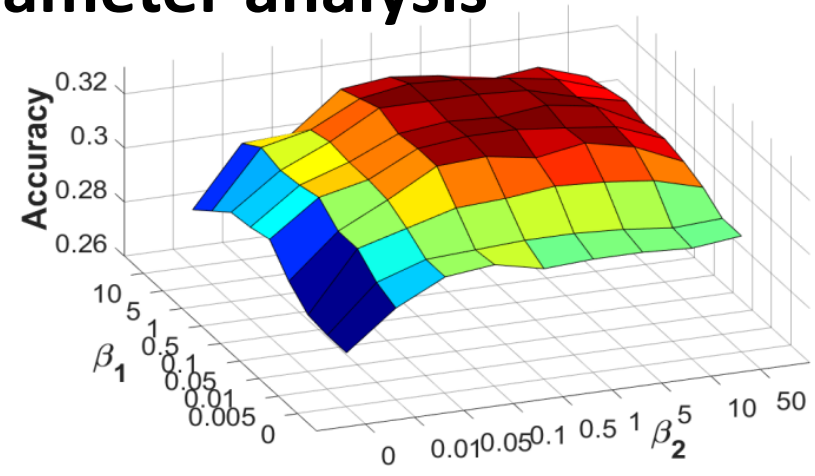
(d) Flowers

- Fix the number of labeled samples to 0.3, and test the classification accuracy with different incomplete rates
- Due to the influence of incomplete views, the performance of all the methods are declined with the increase of incomplete rate
- DCPSL achieves better performance than the others by leveraging both data correlation and multi-view complementary information

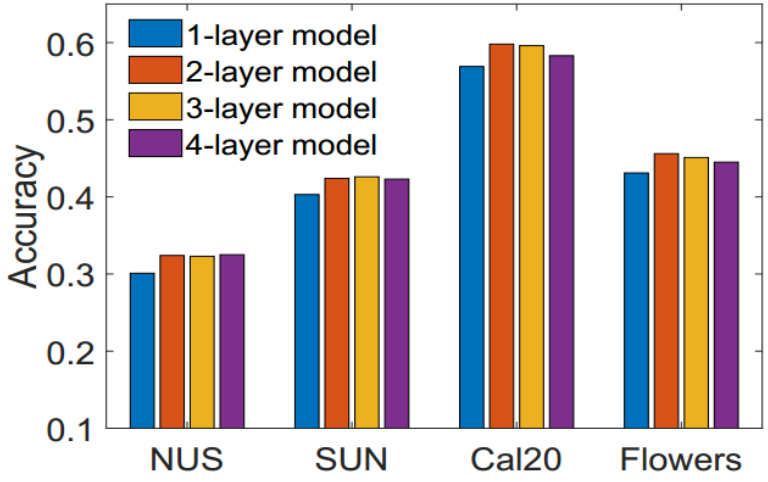
北京郵電大學
Beijing University of Posts and Telecommunications

3.4 Deep Correlated Predictive Subspace Learning for Incomplete Multi-View Semi-Supervised Classification

Parameter analysis



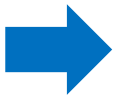
(a) Sensitivity analysis of β_1, β_2 on NUS.



(b) Sensitivity analysis of m on each dataset

- DCPSL is not sensitive w.r.t different parameter settings. It obtains competitive performance when β_1 = {0.05,…,5}, and β_2 = {0.1,…,10}
- Better classification results can be obtained when m = {2, 3} for most of the datasets. The shallow model (m = 1) fails to learn the discriminative subspace representation, so its classification performance is limited

OUTLINE



1. Background
2. Related work
3. Research works
4. Summary

Summary

Four multi-view fusion and representation methods for image analysis are introduced:

- A group-aware multi-view fusion approach for image clustering is proposed
  - ✓ The problem of inaccurate multi-view fusion caused by global fusion method is solved
- A bi-level multi-view latent space learning method is proposed
  - ✓ It overcomes the incomparability of multi-view data and the influence of noise. More compact and discriminative multi-view representation can be obtained
- Two classification methods for complete multi-view data and incomplete multi-view data are proposed
  - ✓ By jointly learning multi-view representation and image classification, the accuracy of multi-view representation and the discriminating power of classifiers are improved

Thank you!
Q & A

References

- Zhiwu Lu and Yuxin Peng, Unified Constraint Propagation on Multi-View Data, *AAAI* 2013
- Yang Wang, Xuemin Lin, Lin Wu, Wenjie Zhang, Qing Zhang, and Xiaodi Huang, Robust Subspace Clustering for Multi-View Data by Exploiting Correlation Consensus, *TIP* 2015
- Abhishek Kumar, Hal Daume III, A Co-training Approach for Multi-view Spectral Clustering, *ICML* 2011
- Abhishek Kumar, Piyush Rai, Hal Daume III, Co-regularized Multi-view Spectral Clustering, *NIPS* 2011
- Grigorios Tzortzis and Aristidis Likas, Kernel-based Weighted Multi-view Clustering, *ICDM* 2012
- Mehmet Gonen, Adam A. Margolin, Localized Data Fusion for Kernel k-Means Clustering with Application to Cancer Biology, *NIPS* 2014
- Mehmet Gonen, Ethem Alpaydn, Localized Multiple Kernel Learning, *ICML* 2008
- Hsin-Chien Huang, Yung-Yu Chuang, Chu-Song Chen, Affinity Aggregation for Spectral Clustering, *CVPR* 2015
- Anran Wang, Jianfei Cai, Jiwen Lu, and Tat-Jen Cham, MMSS: Multi-modal Sharable and Specific Feature Learning for RGB-D Object Recognition, *ICCV* 2015
- Handong Zhao, Zhengming Ding, Yun Fu, Multi-View Clustering via Deep Matrix Factorization, *AAAI* 2017