Multi-modal Subspace Learning with Joint Graph Regularization for Cross-modal Retrieval

Kaiye Wang, Wei Wang, Ran He, Liang Wang, Tieniu Tan

{kaiye.wang, wangwei, rhe, wangliang, tnt}@nlpr.ia.ac.cn

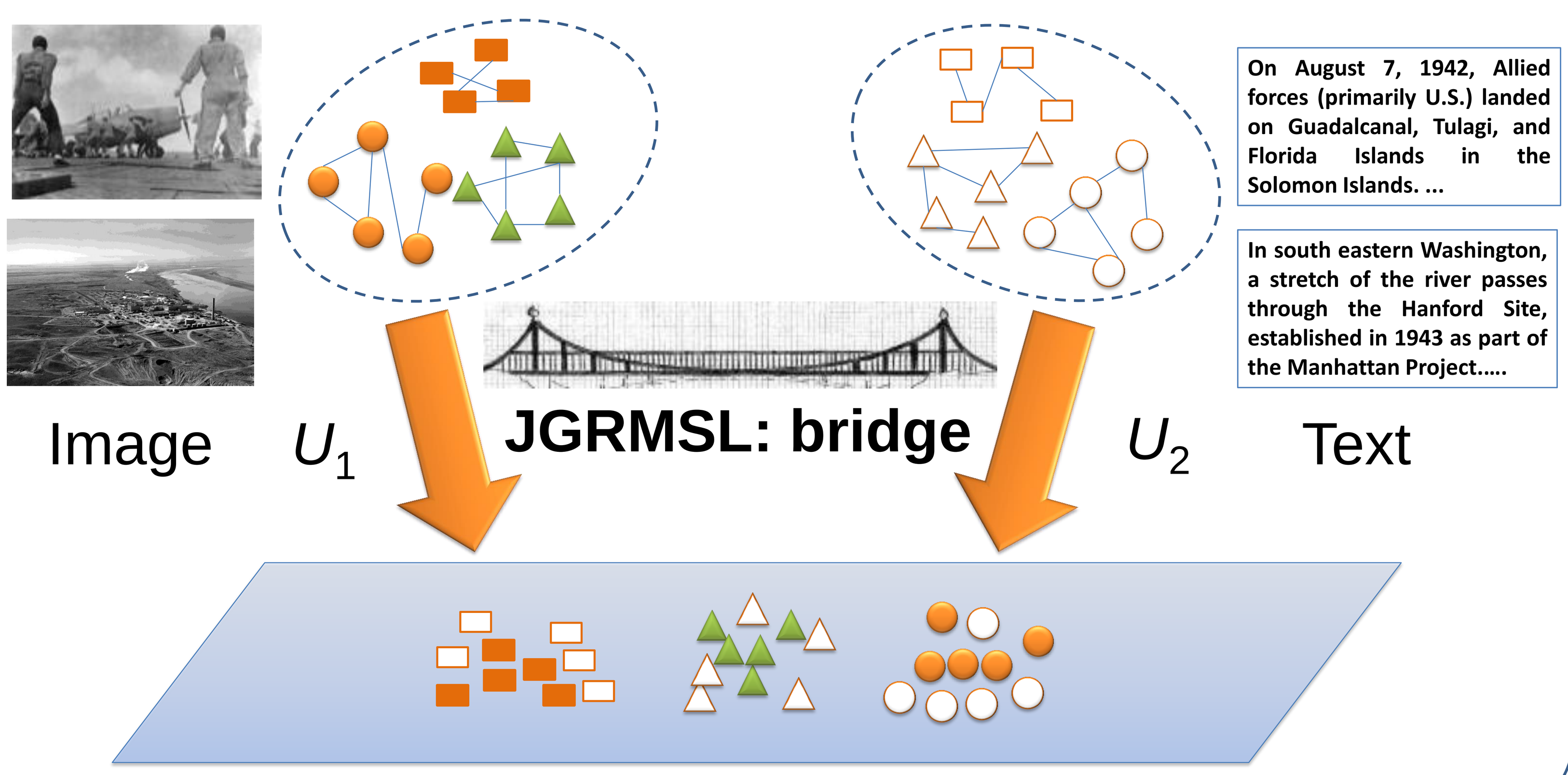
Abstract

Goal: search results across various modalities of data.

Challenge: bridge the heterogeneity gap.

Contribution: we propose a joint graph regularization multi-modal subspace learning(**JGRMSL**) method, which well explores the inter-modality similarity and intra-modality similarity. It also has good discriminability.

Overview



Our method

JGRMSL = inter-modality similarity (similar pairs)
+ intra-modality similarity (neighborhood)
+ discrimination (class information)

Joint graph regularization term

$$J(\mathbf{u}_1,\mathbf{u}_2)=\underline{\sum_{i,j=1}^{n} z_{ij}(\mathbf{u}_1^T\mathbf{x}_i^{(1)}-\mathbf{u}_2^T\mathbf{x}_j^{(2)})^2}+\underline{\frac{\lambda_1}{2}\sum_{i,j=1}^{n} s_{ij}^{(1)}(\mathbf{u}_1^T\mathbf{x}_i^{(1)}-\mathbf{u}_1^T\mathbf{x}_j^{(1)})^2}$$

$$+\underline{\frac{\lambda_2}{2}\sum_{i,j=1}^{n} s_{ij}^{(2)}(\mathbf{u}_2^T\mathbf{x}_i^{(2)}-\mathbf{u}_2^T\mathbf{x}_j^{(2)})^2}$$

Inter-modality similarity

Intra-modality similarity

Reformulation: $X=\begin{bmatrix} X_1 & \mathbf{0} \\ \mathbf{0} & X_2 \end{bmatrix};\mathbf{u}=\begin{bmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \end{bmatrix};W=\begin{bmatrix} \lambda_1 S_1 & Z \\ Z^T & \lambda_2 S_2 \end{bmatrix}$

$$J(\mathbf{u})=\frac{1}{2}\sum_{i,j=1}^{2n} W_{ij}(\mathbf{u}^T X_{(i)}-\mathbf{u}^T X_{(j)})^2$$

$$=\frac{1}{2}\mathbf{u}^T X(D-W)X^T\mathbf{u}$$

$$=\frac{1}{2}\mathbf{u}^T XLX^T\mathbf{u}$$

Inter-modality similarity: project similar pairs as close as possible

Intra-modality similarity: preserve local manifold structure

Objective function - discrimination

$$\arg\max_{\mathbf{u}} \frac{S_B}{S_W+\alpha J(\mathbf{u})} \qquad S_W=\sum_{i=1}^{c}\sum_{v=1}^{2}\sum_{k=1}^{n_i^{(v)}}(\mathbf{y}_{ik}^{(v)}-\boldsymbol{\mu}_i)(\mathbf{y}_{ik}^{(v)}-\boldsymbol{\mu}_i)^T$$

$$\Rightarrow \arg\max_{\mathbf{u}} \frac{S_B}{S_W+\alpha\mathbf{u}^T XLX^T\mathbf{u}} \qquad S_B=\sum_{i=1}^{c} n_i(\boldsymbol{\mu}_i-\boldsymbol{\mu})(\boldsymbol{\mu}_i-\boldsymbol{\mu})^T$$

Discriminability: different-class samples should be mapped far apart while the same-class samples lie as close as possible.

Algorithmic view

Step1: input data from different modalities.

Step2: learn the projection matrices using JGRMSL .

Step3: map data into latent space using learnt projections.

Step4: conduct cross-modal ranking in the latent space.

Experimental results

Evaluation: MAP, PS curve

Compared Methods:

CCA, PLS, BLM (CVPR'11): similar pairs

GMLDA, GMMFA (CVPR'12): similar pairs + label

Results on Pascal image-tag data

20 classes, 2808 / 2841 training/testing samples

Image: 512-dim Gist, Text: 399-dim word frequency

Methods	Image query	Text query	Average
PLS	0.275	0.199	0.237
BLM	0.266	0.240	0.253
CCA	0.265	0.221	0.243
GMMFA	0.309	0.230	0.269
GMLDA	0.242	0.204	0.223
JGRMSL	**0.346**	**0.265**	**0.305**

Table 1. Comparison of MAP for different methods

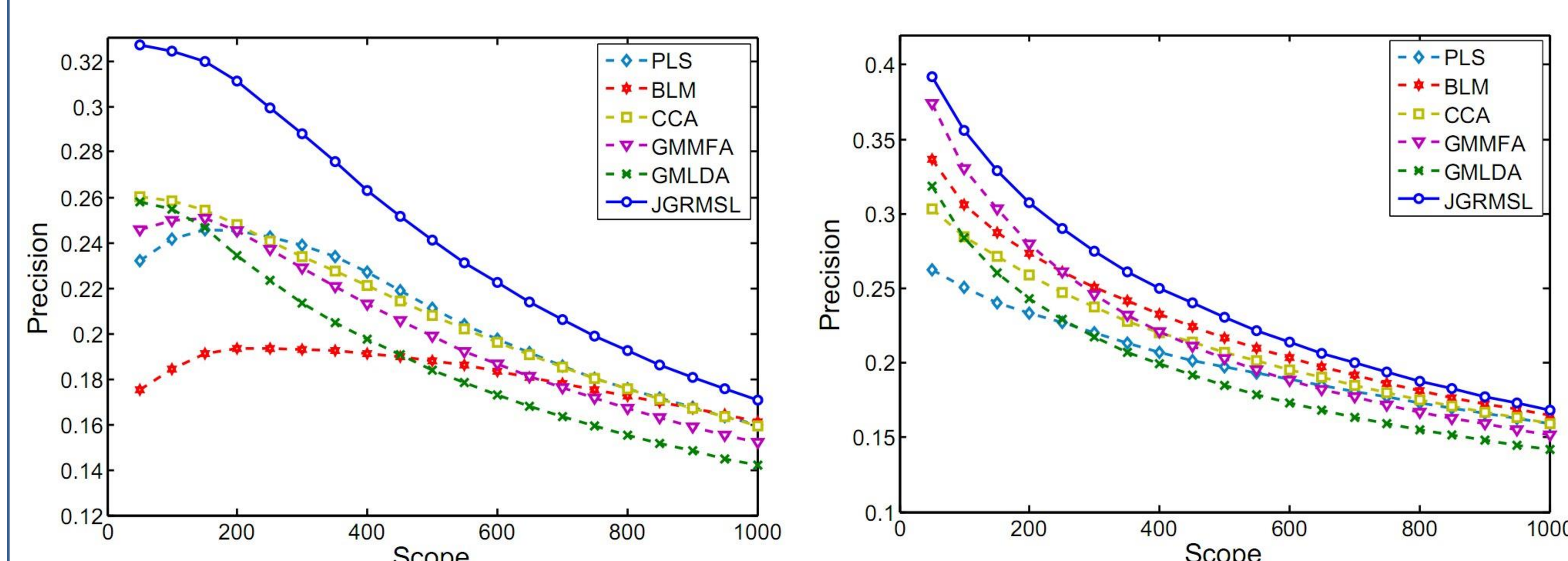


Figure 1. Precision-scope curves of different methods. **Left:** Image as query, **Right:** Text as query

Results on Wikipedia image-text data

10 classes, 1300 / 1566 training/testing samples

Image: 128-dim bags of SIFT, Text: 10-dim LDA

Methods	Image query	Text query	Average
PLS	0.240	0.163	0.202
BLM	0.256	0.202	0.229
CCA	0.254	0.184	0.219
GMMFA	0.276	**0.213**	0.245
GMLDA	0.275	0.210	0.243
JGRMSL	**0.304**	0.211	**0.258**

Table 2. Comparison of MAP for different methods

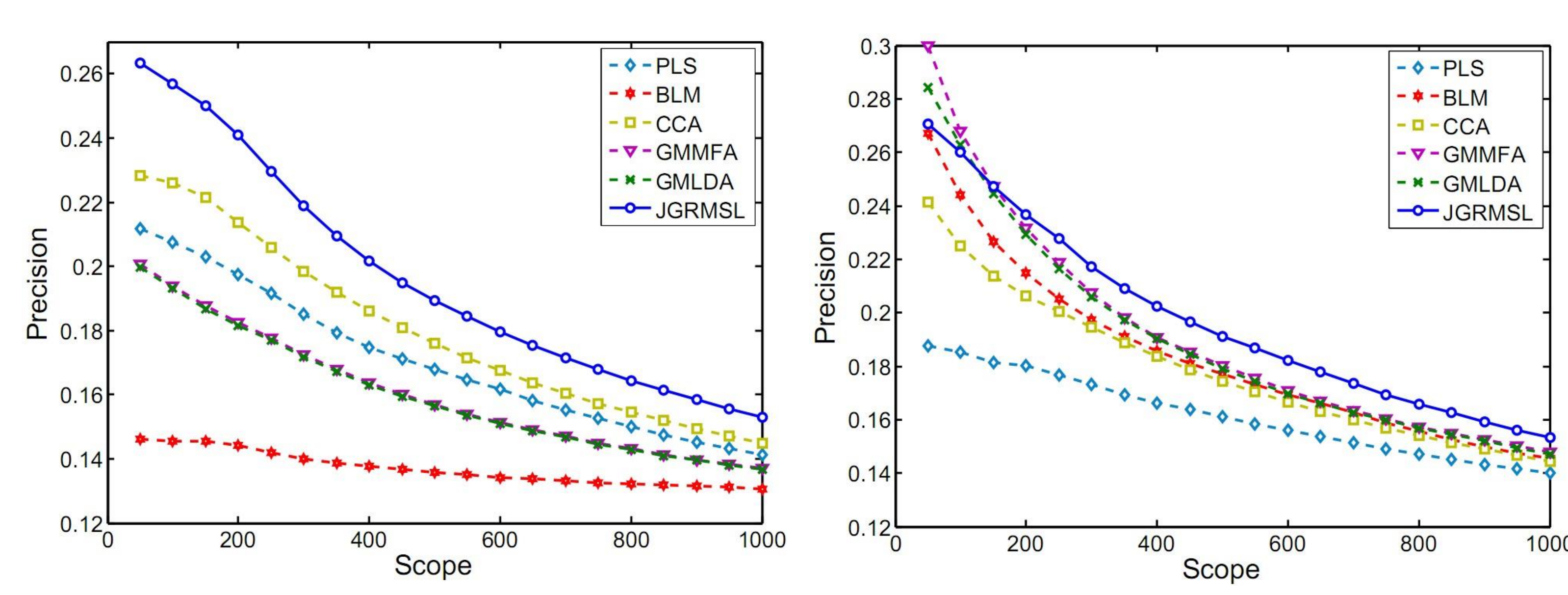


Figure 2. Precision-scope curves of different methods. **Left:** Image as query, **Right:** Text as query

Center for Research on Intelligent Perception and Computing (CRIPAC)
National Laboratory of Pattern Recognition (NLPR)
Institute of Automation, Chinese Academy of Sciences (CASIA)