Dimensionality Reduction for Data in Multiple Feature Representations

Yen-Yu Lin[1,2], Tyng-Luh Liu[1], and Chiou-Shann Fuh[2]

[1]Institute of Information Science, Academia Sinica, Taiwan; [2]Dept. of CSIE, National Taiwan University, Taiwan

1. Summary

- In solving complex visual learning tasks, we establish an approach in which multiple kernel learning (MKL) is incorporated into the training process of dimensionality reduction (DR) methods.
- Our approach
  - Based on MKL, data described by various descriptors are jointly considered, and projected into a unified space of low dimension.
  - Built on graph embedding, any DR methods explainable by graph embedding can be generalized by our approach.
  - Via integrating with different DR methods, MKL can address not only supervised learning problems but also semi-supervised and unsupervised ones.

2. Background

- Multiple kernel learning and graph embedding are two important components in the establishment of the approach.

2.1 Multiple Kernel Learning

- In complex vision applications, adopting multiple descriptors to characterize data is a feasible way for improving performances.
- Kernel matrix as a unified feature representation:
  - Represent the pairwise relationships among data under each descriptor by a kernel matrix.
- Multiple kernel learning: Learning an optimal ensemble kernel over a given convex set of base kernels:

$$K = \sum_{m=1}^{M} \beta_m K_m, \quad \beta_m \geq 0 \tag{1}$$

$$\text{or } k(\mathbf{x}_i, \mathbf{x}_j) = \sum_{m=1}^{M} \beta_m k_m(\mathbf{x}_i, \mathbf{x}_j), \quad \beta_m \geq 0 \tag{2}$$

- Multiple kernel learning for finding optimal coefficients $\{\beta_m\}_{m=1}^{M}$ can be interpreted as descriptor combination.

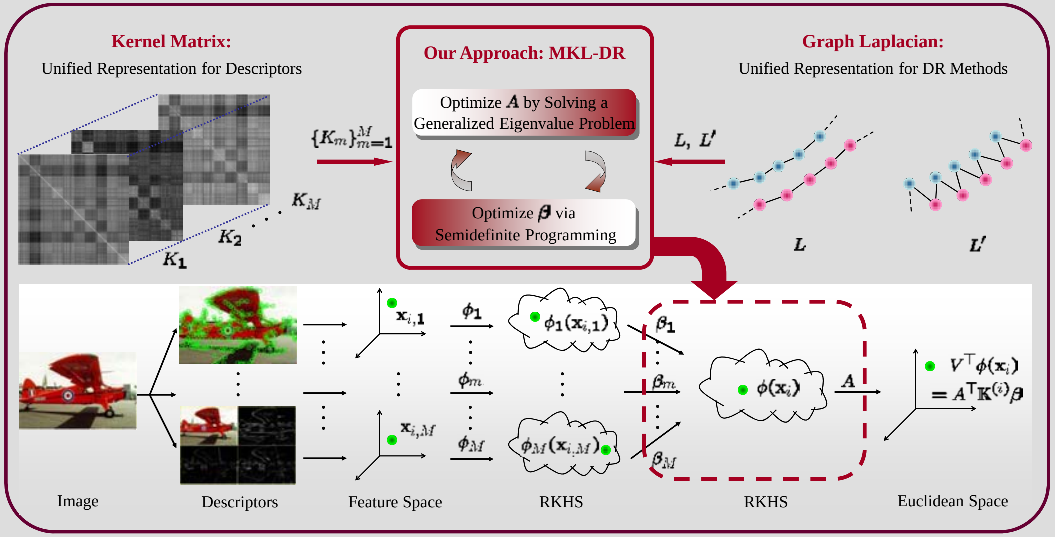
2.2 Graph Embedding

- Many DR methods focus on modeling pairwise relationships among data points via graph structures.
- The framework of graph embedding [Yan et al. PAMI07] provides a unified formulation for a set of graph-based DR methods:

$$\mathbf{v}^* = \underset{\substack{\mathbf{v}^\top XDX^\top \mathbf{v}=1, \text{ or} \\ \mathbf{v}^\top XL'X^\top \mathbf{v}=1}}{\arg\min} \mathbf{v}^\top XLX^\top \mathbf{v}, \tag{3}$$

where $X = [\mathbf{x}_1\ \mathbf{x}_2\ \cdots\ \mathbf{x}_N]$ is the data matrix,
L and L' are graph Laplacian,
D is a diagonal matrix,
$\mathbf{v}^*$ is the optimized linear embedding.

- By specifying particular L and L' (or L and D), a set of DR methods can by expressed by (3), such as PCA, LPP, LDA, LDE, or SDA.
- These DR methods include supervised, semi-supervised, and unsupervised ones.



3. Problem Definition

- Our goal is to make DR methods that can be expressed by (3) to consider multiple base kernels simultaneously, such that
  - Data characteristics captured in different descriptors can be jointly utilized to accomplish the objectives of these DR methods.
  - The diversity in the objectives of these DR methods enhances the applicability of MKL.
  - We could provide more prior knowledge to benefit the analysis of given data.
- With some derivation, sample i in the projected space can be expressed as

$$\mathbf{v}^\top \mathbf{x}_i = \boldsymbol{\alpha}^\top \mathbb{K}^{(i)} \boldsymbol{\beta} \in \mathbb{R},$$

where $\boldsymbol{\alpha} = [\alpha_1\ \alpha_2\ \cdots\ \alpha_N]^\top$ is the sample coefficient vector,
$\boldsymbol{\beta} = [\beta_1\ \beta_2\ \cdots\ \beta_M]^\top$ is the kernel weight vector,
$\mathbb{K}^{(i)} = [K_1^{(i)}\ K_2^{(i)}\ \cdots\ K_M^{(i)}] \in \mathbb{R}^{N \times M}$.

- The resulting constrained optimization problem is

$$\min_{A,\boldsymbol{\beta}} \sum_{i,j=1}^{N} \|A^\top \mathbb{K}^{(i)} \boldsymbol{\beta} - A^\top \mathbb{K}^{(j)} \boldsymbol{\beta}\|^2 w_{ij} \tag{4}$$

$$\text{subject to } \sum_{i,j=1}^{N} \|A^\top \mathbb{K}^{(i)} \boldsymbol{\beta} - A^\top \mathbb{K}^{(j)} \boldsymbol{\beta}\|^2 w'_{ij} = 1,$$

$$\beta_m \geq 0, \text{ for } m = 1, 2, \ldots, M,$$

where $W = [w_{ij}]$ is the affinity matrix of L in (3),
$W' = [w'_{ij}]$ is the affinity matrix of L' in (3),
$A = [\boldsymbol{\alpha}_1\ \boldsymbol{\alpha}_2\ \cdots\ \boldsymbol{\alpha}_P]$.

4. Optimization

- An alternative optimization procedure is used to solve (4).
  - Variable A is optimized by solving a generalized eigenvalue problem.
  - Variable $\boldsymbol{\beta}$ is optimized via semidefinite programming.

4.1 On Optimizing A

- By fixing $\boldsymbol{\beta}$, the optimization problem (4) can be expressed as

$$\min_A \operatorname{trace}(A^\top S_W^{\boldsymbol{\beta}} A)$$

$$\text{subject to } \operatorname{trace}(A^\top S_{W'}^{\boldsymbol{\beta}} A) = 1$$

where $S_W^{\boldsymbol{\beta}}$ and $S_{W'}^{\boldsymbol{\beta}}$ are two fixed square matrices.

- The columns of the optimized A can be obtained by solving a generalized eigenvalue problem.

4.2 On Optimizing $\boldsymbol{\beta}$

- By fixing A, (4) becomes a nonconvex QCQP with respect to $\boldsymbol{\beta}$.
- Thus, we instead consider its SDP relaxation:

$$\min_{\boldsymbol{\beta},B} \operatorname{trace}(S_W^A B)$$

$$\text{subject to } \operatorname{trace}(S_{W'}^A B) = 1,$$

$$\beta_m \geq 0, \text{ for } m = 1, 2, \ldots, M,$$

$$\begin{bmatrix} 1 & \boldsymbol{\beta}^T \\ \boldsymbol{\beta} & B \end{bmatrix} \succeq 0,$$

where S_W^A and $S_{W'}^A$ are two fixed square matrices.

- The optimal $\boldsymbol{\beta}$ can be obtained via semidefinite programming.

5. Experimental Results

- The Caltech 101 image dataset is used in the experiments.
  - It consists of 101 object categories and one additional class of background images.
  - All 102 classes are used in the application to object recognition
  - 30 classes are chosen in the application to image clustering

- We implement seven kinds of image descriptors that result in the following seven base kernel matrices:
  - GB-1 / GB-2: Based on geometric blur descriptor.
  - SIFT-Dist / SIFT- Grid: Based on the SIFT descriptor.
  - C2-SWP / C2-ML: Based on biologically inspired features.
  - PHOG: Based on PHOG descriptor.

5.1 Supervised Application to Recognition

- We adopt two supervised DR schemes for the applications:
  - Linear discriminant analysis (LDA): Assume data of a class can be modeled by a Gaussian.
  - Local discriminant embedding (LDE) [Chen et al. CVPR05]: Assume data of a class spread as a submanifold.
- To express LDA and LDE in the form of (3), we need specify their corresponding graph Laplacian matrices, i.e., L and L'.
- The recognition rates are listed as follows

Table 1: Recognition rates (mean ± std %) for Caltech-101 dataset

kernel(s)	method	number of classes		method	number of classes	
		102	101		102	101
GB-1	KFD	57.3 ± 2.5	57.7 ± 0.7	KLDE	57.1 ± 1.4	57.7 ± 0.8
GB-2		60.0 ± 1.5	60.6 ± 1.5		60.9 ± 1.4	61.3 ± 2.1
SIFT-Dist		53.0 ± 1.4	53.2 ± 0.8		54.2 ± 0.5	54.6 ± 1.5
SIFT-Grid		48.8 ± 1.9	49.6 ± 0.7		49.5 ± 1.3	50.1 ± 0.3
C2-SWP		30.3 ± 1.2	30.7 ± 1.5		31.1 ± 1.5	31.3 ± 0.7
C2-ML		46.0 ± 0.6	46.8 ± 0.9		45.8 ± 0.2	46.7 ± 1.5
PHOG		41.8 ± 0.6	42.1 ± 1.3		42.2 ± 0.6	42.6 ± 1.3
-	KFD-*Voting*	68.4 ± 1.5	68.9 ± 0.3	KLDE-*Voting*	68.4 ± 1.4	68.7 ± 0.8
-	KFD-*SAMME*	71.2 ± 1.4	72.1 ± 0.7	KLDE-*SAMME*	71.1 ± 1.9	71.3 ± 1.2
All	MKL-LDA	**74.6 ± 2.2**	**75.3 ± 1.7**	MKL-LDE	**75.3 ± 1.5**	**75.5 ± 1.7**

5.2 Unsupervised Application to Clustering

- We apply our approach to locality preserving projections (LPP) [He & Niyogi NIPS03] to serve as a data preprocessing tool.
- Data in various representations are projected into a unified space.
- Consider clustering methods affinity propagation and *k*-means.

Table 2: Clustering performance (NMI / ERR %) on the 20-class image dataset

kernel(s)	preprocessing method	affinity propagation		*k*-means clustering	
		without data preprocessing	with data preprocessing	without data preprocessing	with data preprocessing
GB-1	kernel LPP	0.553 / 50.8	0.609 / 38.3	–	0.611 / 44.7
GB-2		0.577 / 48.0	0.624 / 43.7	–	0.640 / 43.0
SIFT-Dist		0.627 / 43.7	0.651 / 31.0	–	0.671 / 41.7
SIFT-Grid		0.598 / 41.3	0.631 / 45.7	–	0.642 / 43.0
C2-SWP		0.383 / 70.3	0.379 / 60.5	0.383 / 68.5	0.379 / 66.7
C2-ML		0.499 / 54.5	0.488 / 56.0	0.525 / 53.0	0.507 / 55.8
PHOG		0.455 / 57.3	0.482 / 52.7	–	0.510 / 52.7
All	MKL-LPP	–	**0.714 / 25.0**	–	**0.758 / 25.7**