Learning and Mining with Visual Data on the Web

Dacheng Tao

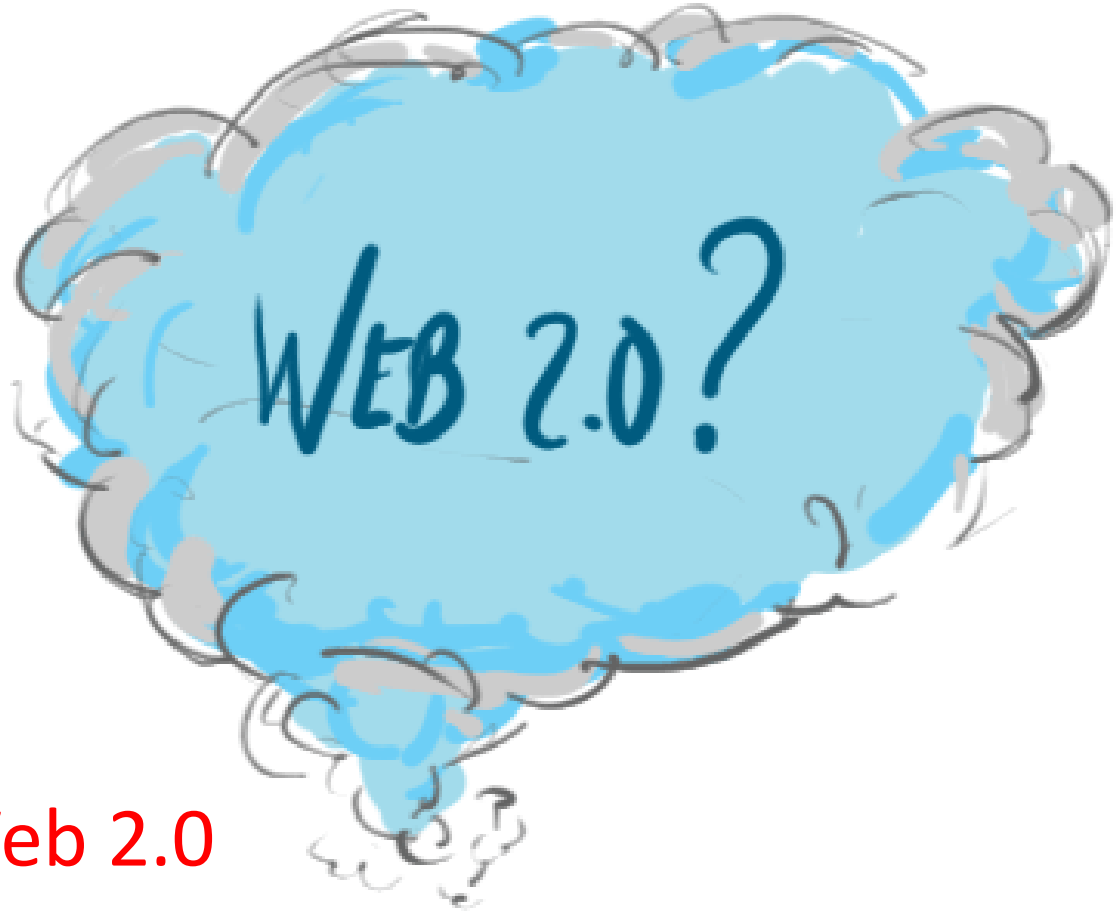
UTS:QCIS
QUANTUM COMPUTATION & INTELLIGENT SYSTEMS

Outline

- Part I: Introduction to Web 2.0
- Part II: Data Driven Approaches to Multimedia Content Understanding
- Part III: Patch Align Framework for Improving the Performance

Short break: 30 minutes

- Part IV: Extensions of Patch Alignment Framework
- Part V: Applications, Challenges and Future Directions



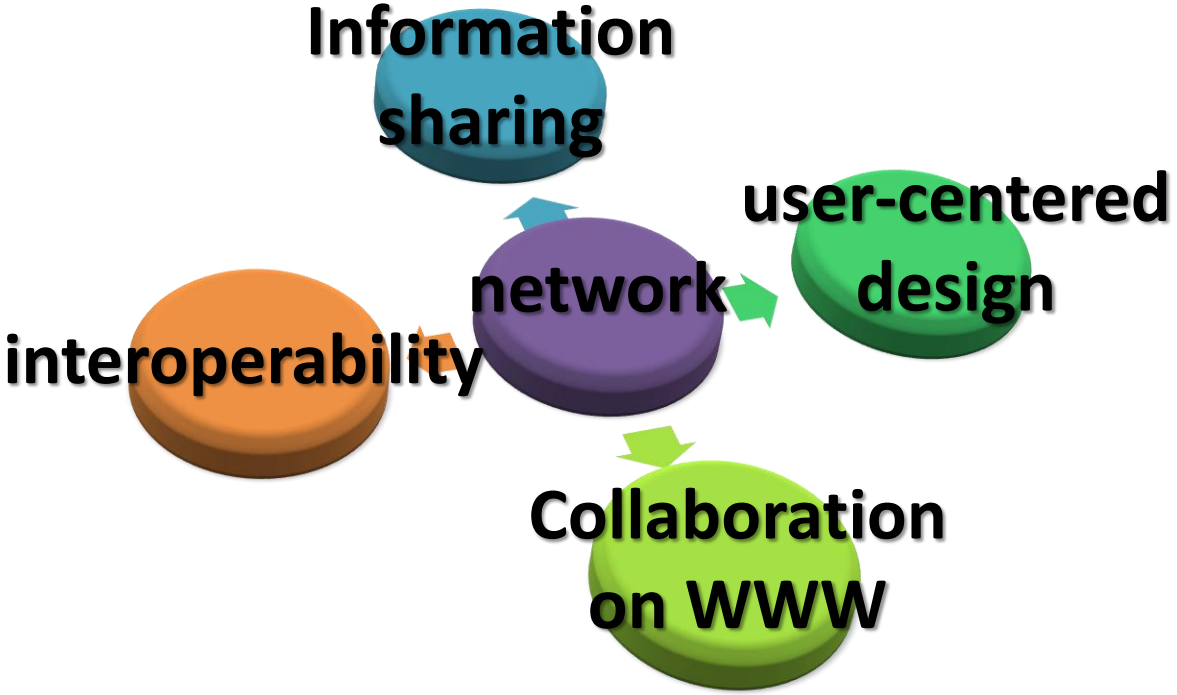
Introduction to Web 2.0

PART ONE

UTS:QCIS
QUANTUM COMPUTATION & INTELLIGENT SYSTEMS

What is Web 2.0?

Web 2.0 is a concept that takes the network as a platform for:



What is Web 2.0?

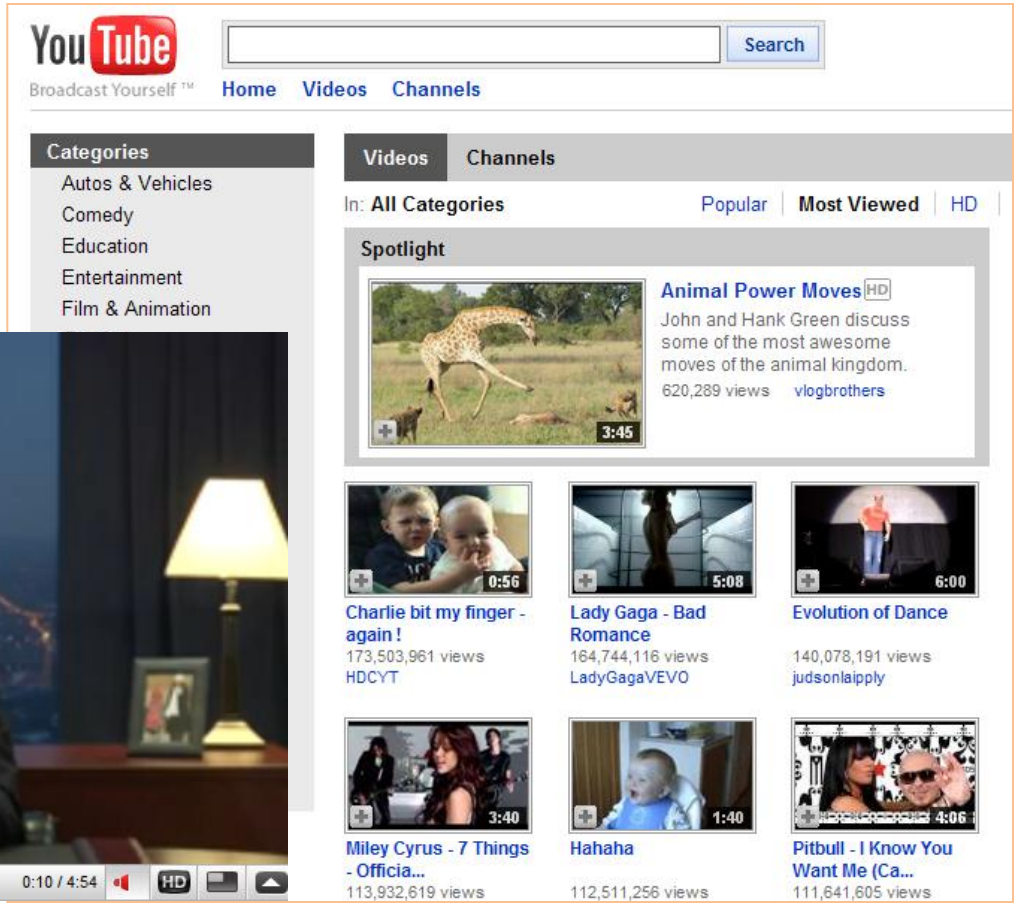
in a social media dialogue

Examples of Web 2.0

- social network
- blogs
- wikis
- media sharing
- hosted services
- web application
- …

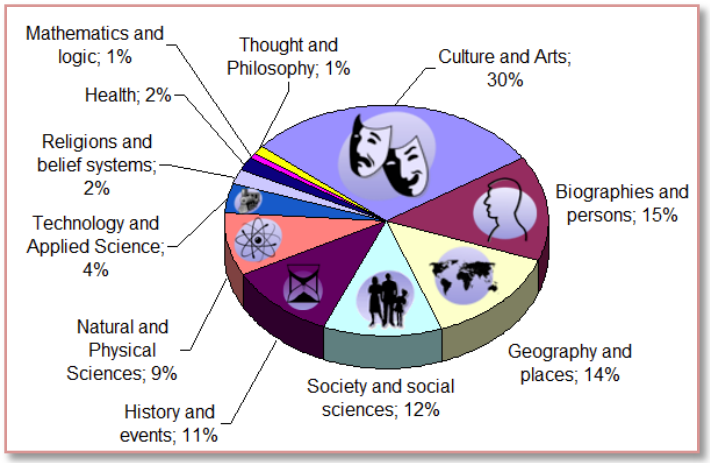
YouTube™

- Video sharing website which allows users to upload and share videos
- Launched in 2005
- Reached 1 billion views per day in 2009!



Wikipedia

- Collaborative dictionary being edited in real time by millions of users around the world.
- Wikipedia covers a wider variety of topics than any print resource in existence!

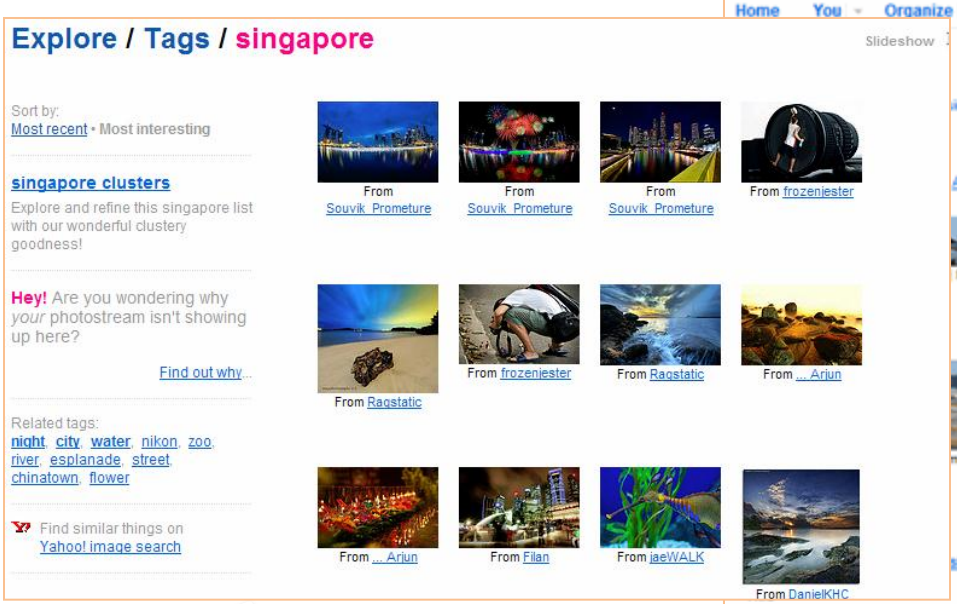


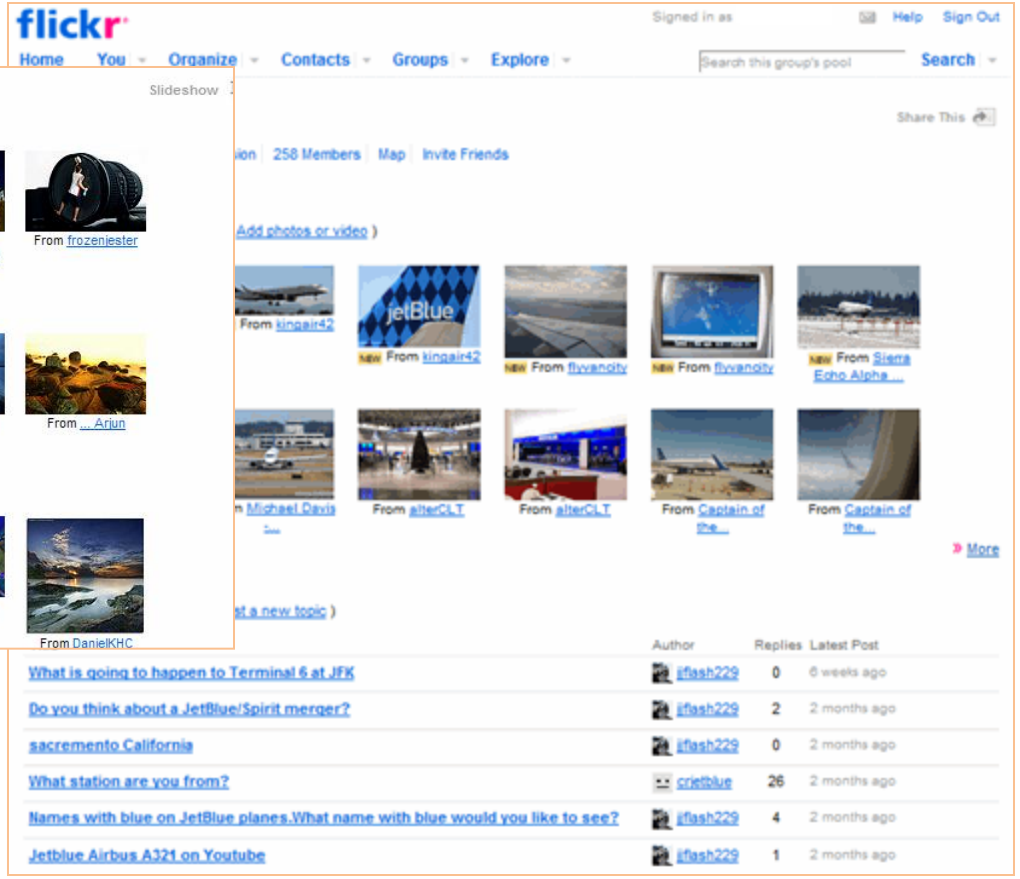
Wikipedia content by subject

Sample Wiki article

- Social networking website with over 350,000,000 users
- Basic features include networking with others and posting on a "wall" or "commenting" on pictures.

- Online photo management and sharing website
- 4+ billion users uploaded pictures, 2+ million uploads per day





Enable searching by Tags

Characteristics of Web 2.0

- Rich user experience
- User participation
- Dynamic content
- Scalability
- Openness
- Web standards

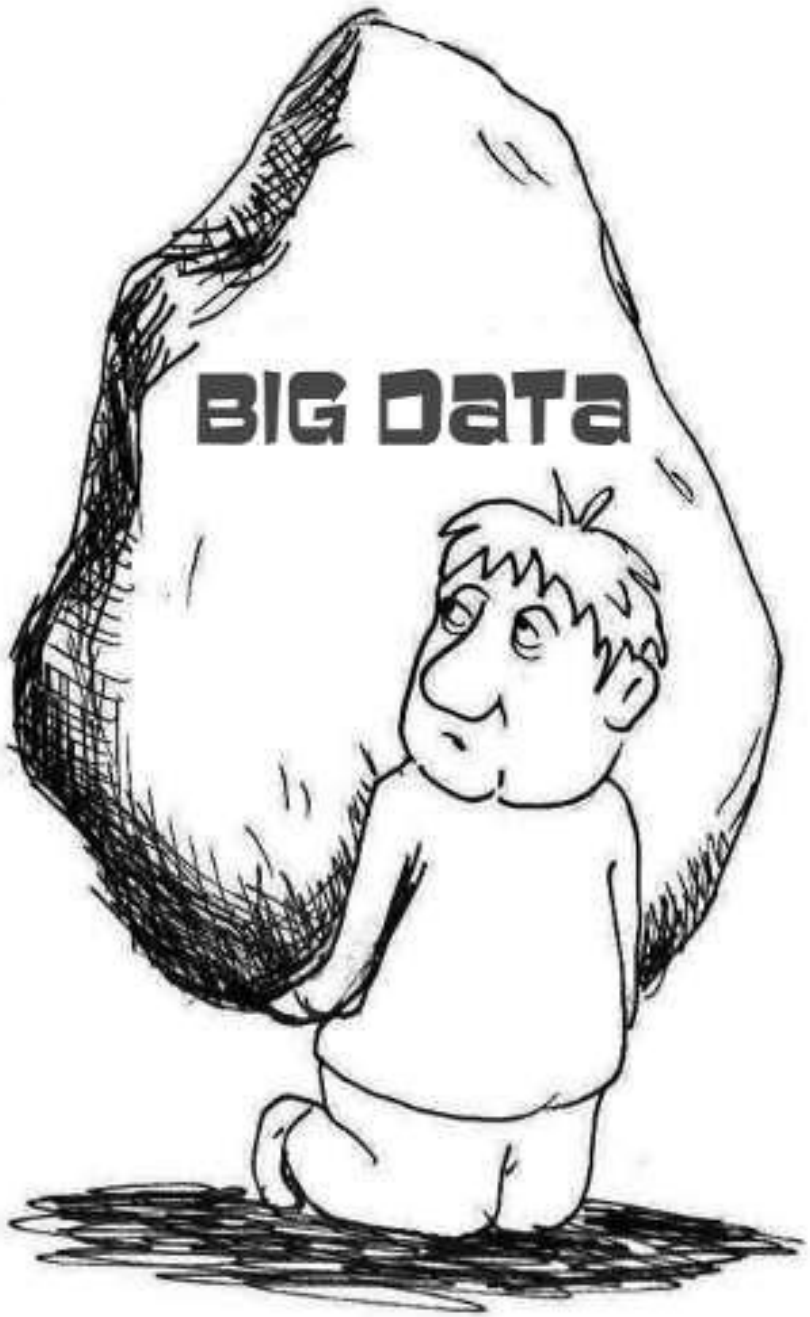
Future of Web 2.0: Web 3.0

- Definition varies greatly…
- **Intelligent Web**: natural language processing, machine-based learning and reasoning, intelligent applications
- **Semantic Web**: computers can understand meaning (semantics) of information and become capable of analysing all the data on the Web
- **Personalization** and **Behavioural advertising**

Data Driven Approaches to Multimedia Content Understanding

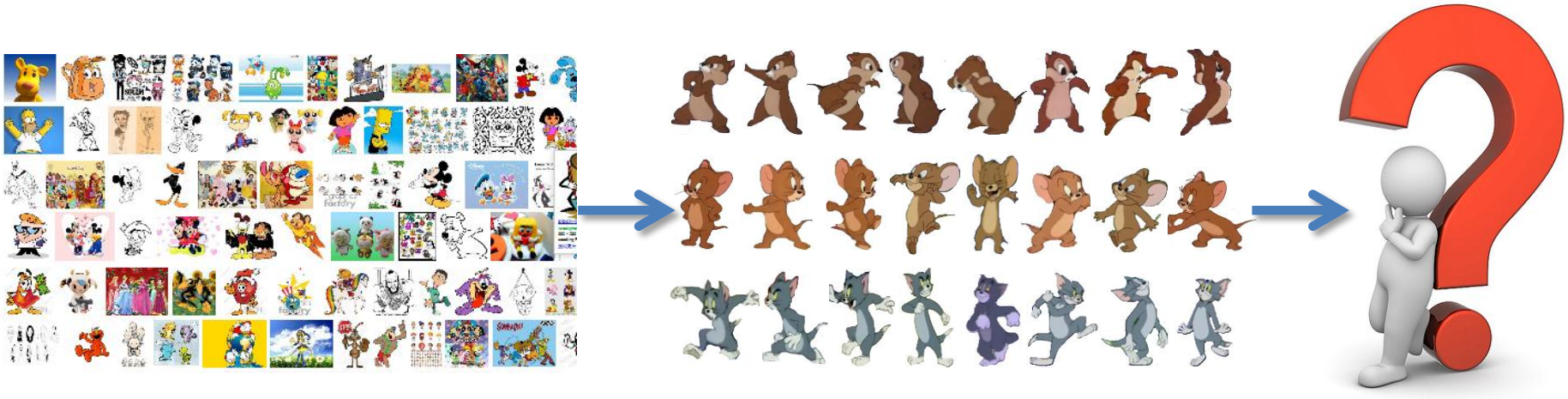
PART TWO

UTS:QCIS
QUANTUM COMPUTATION & INTELLIGENT SYSTEMS



- Previously: **small**
  - Image processing
  - Image understanding
  - Image database management
- Currently: **big**
  - Image processing
  - Image understanding
  - Image database management
- More problems to study
  - Serve people better**!**

Cartoon Inbetweening



Cartoon demo

Scene Completion

[Hays and Efros. Scene Completion Using Millions of Photographs. SIGGRAPH 2007]

Scene Completion

Diffusion Result

Efros and Leung result

Criminisi et al. result

Scene Completion

Image Superresolution

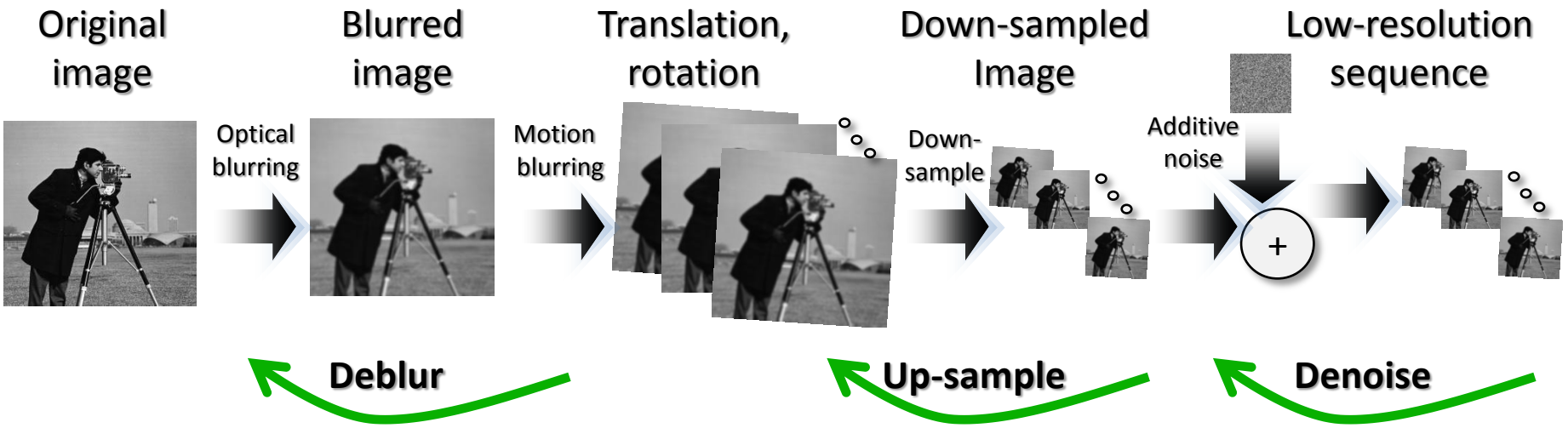
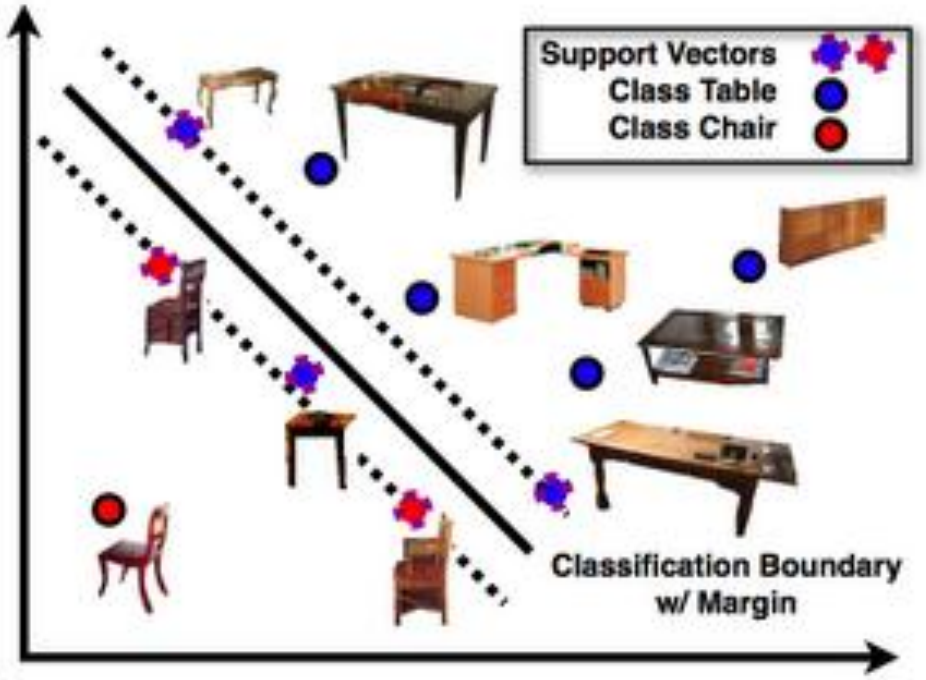


Image Super-Resolution

Deblur → **Up-sample** → **Denoise**

Image Classification



?

Motivations

- Semantic gap is still the key challenge

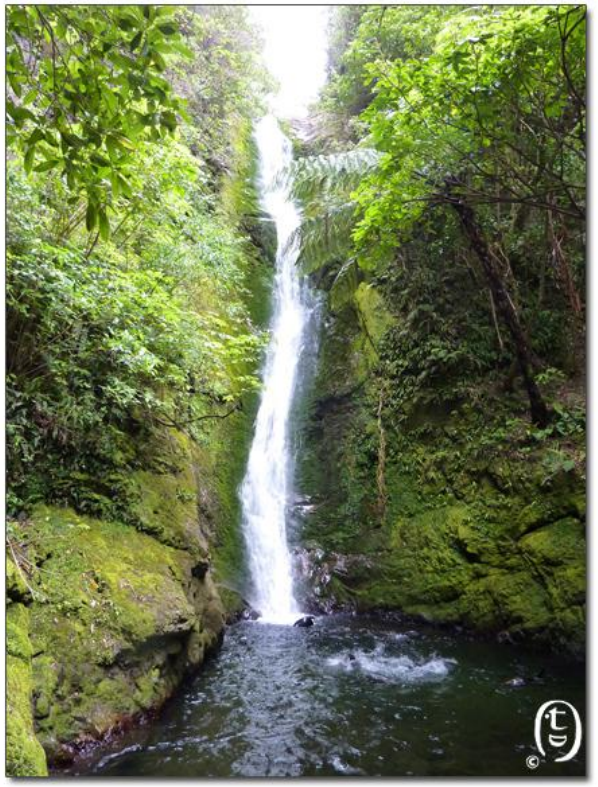
- Attempts have been tried
  - Better feature extraction algorithms
  - Better relevance feedback algorithms
  - Better image annotation (classification) results

- A data driven approach for image annotation

Image Representation

- Global features
  - Color histogram, color coherence vector …
- Local features
  - BoW: SIFT, SURF…
- Biologically inspired features
  - GIST, C1/C2/C4…

$$\left\{\begin{array}{c} x \in \mathbb{R}^n, X \in \mathbb{R}^{n_1 \times n_2}, \\ \mathcal{X} \in \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_k}, \\ or\ \{x^v | 1 \leq v \leq V\} \end{array}\right\}$$



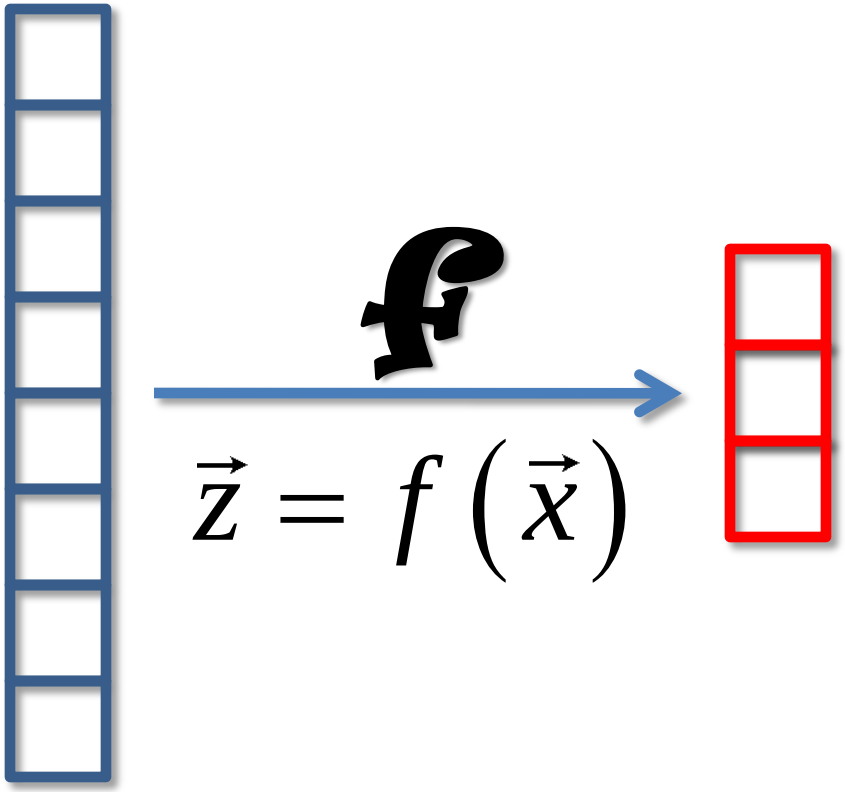
Learning Models

- Supervised learning
- Unsupervised learning
- Semi-supervised learning

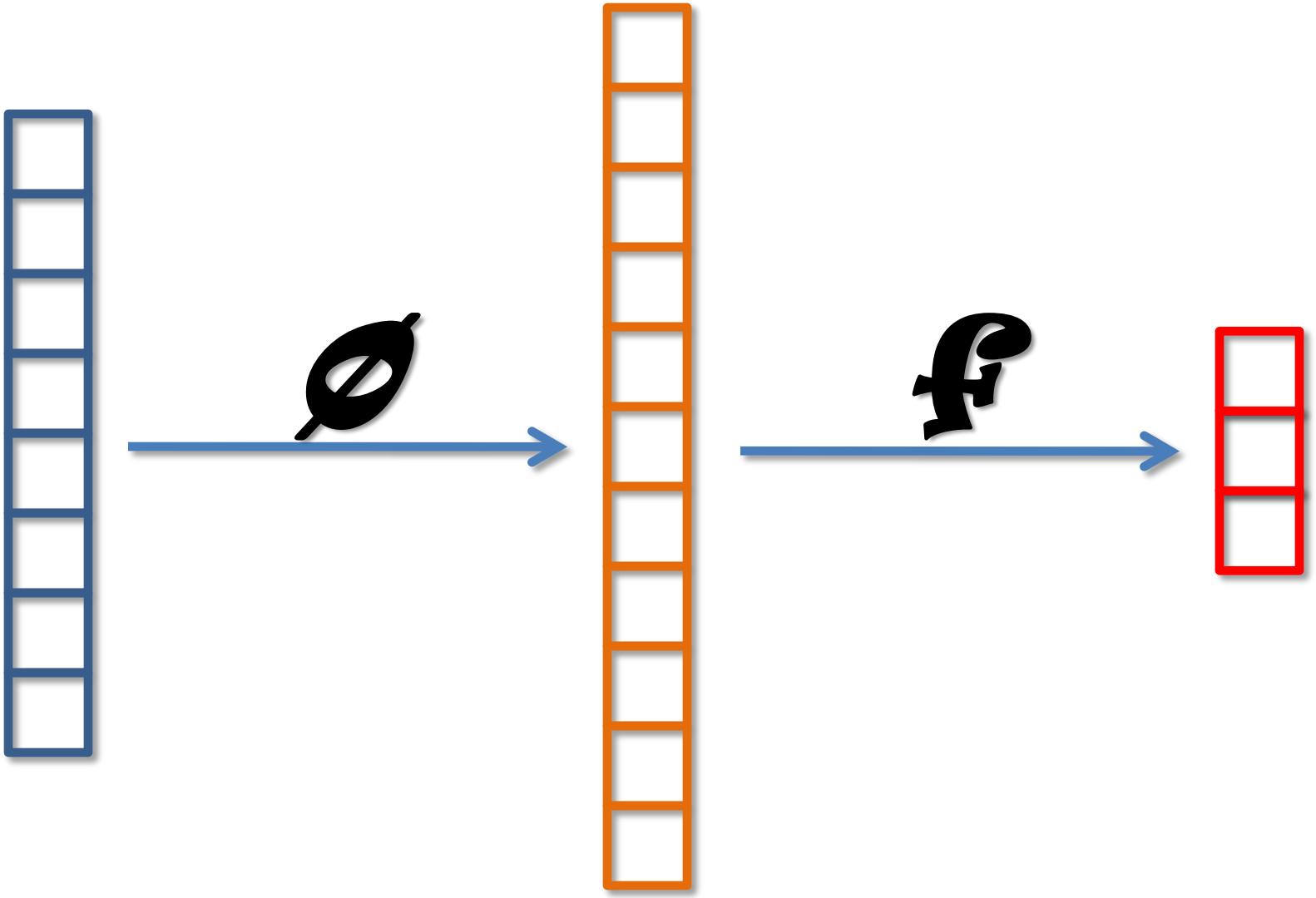
- Single-view learning
- Multi-view learning

Inference

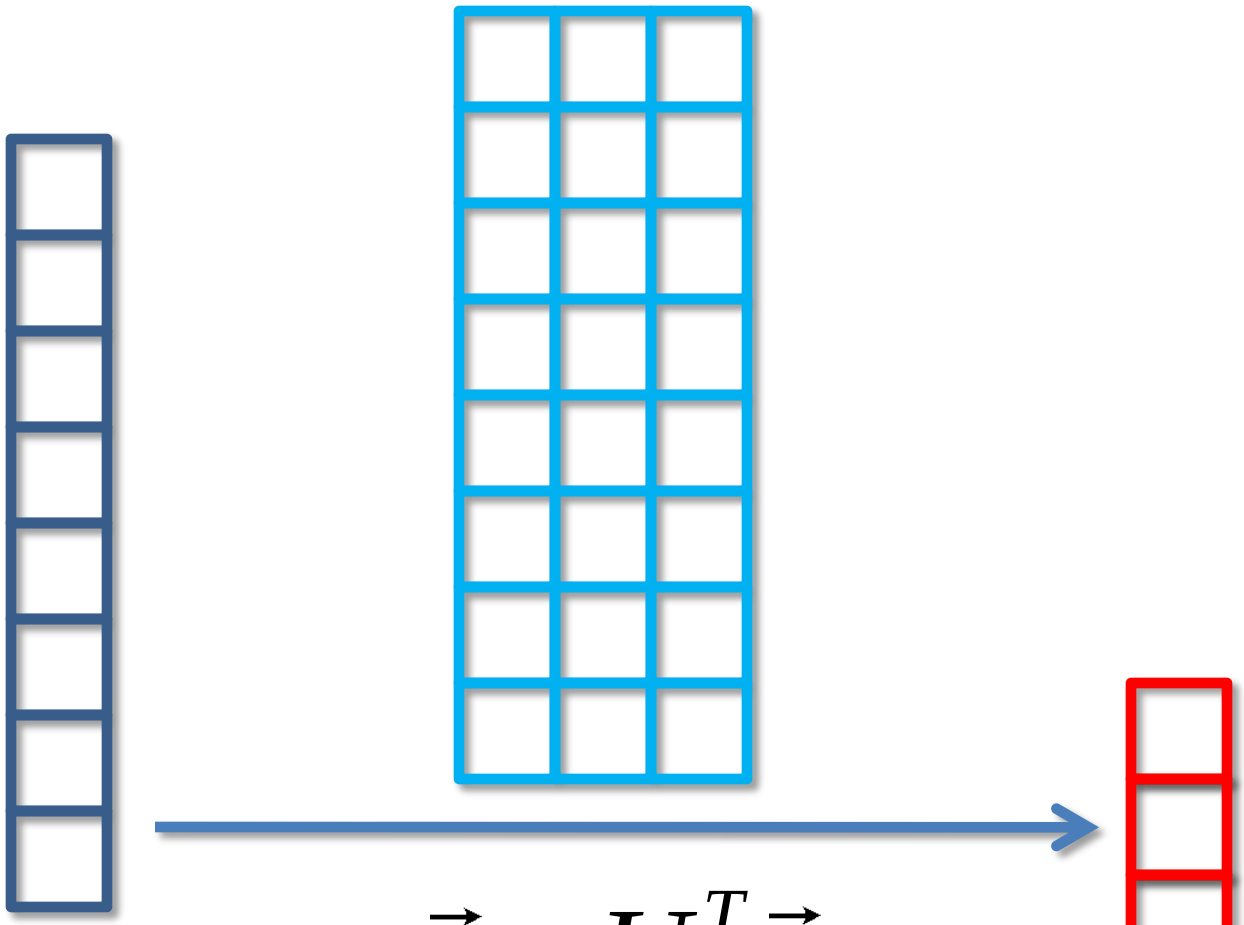
What Is Dimension Reduction?



What Is Dimension Reduction?

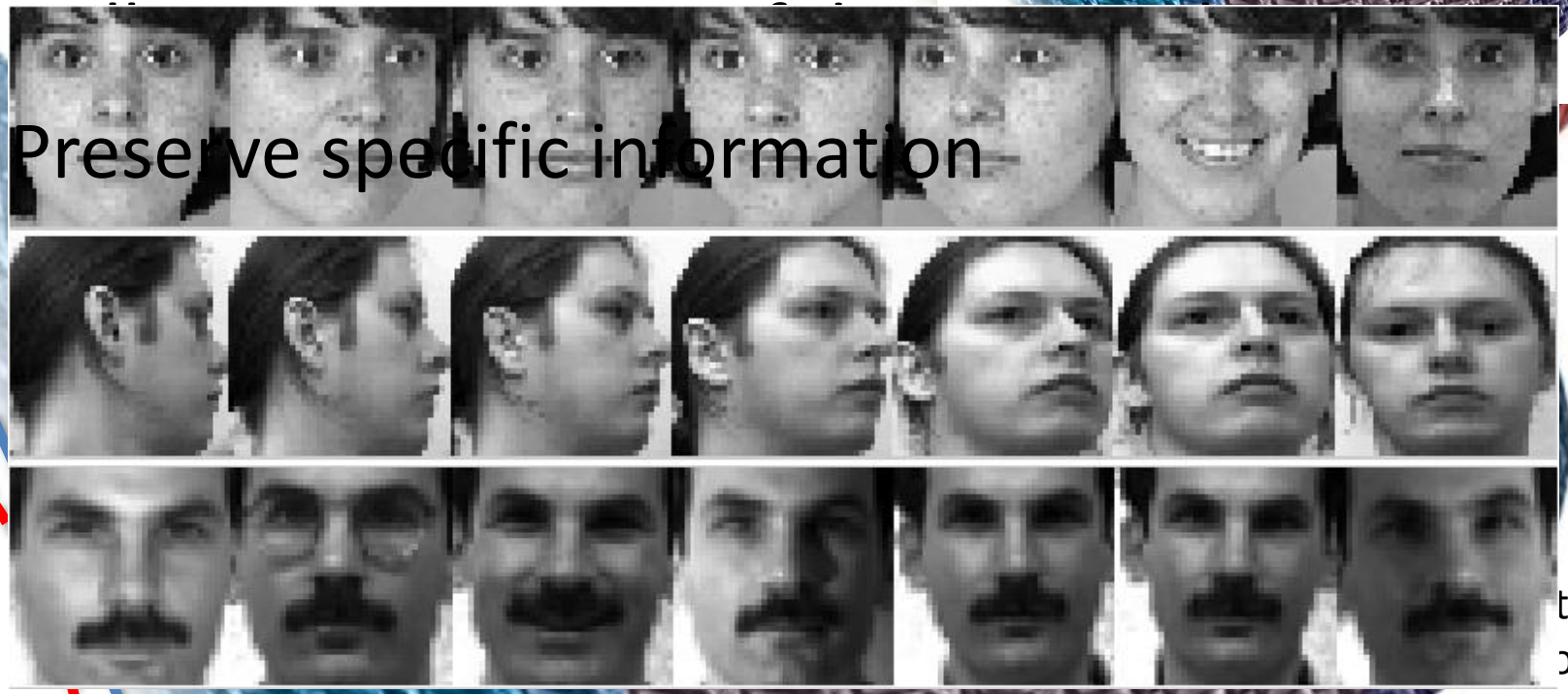


What Is Dimension Reduction?



Importance

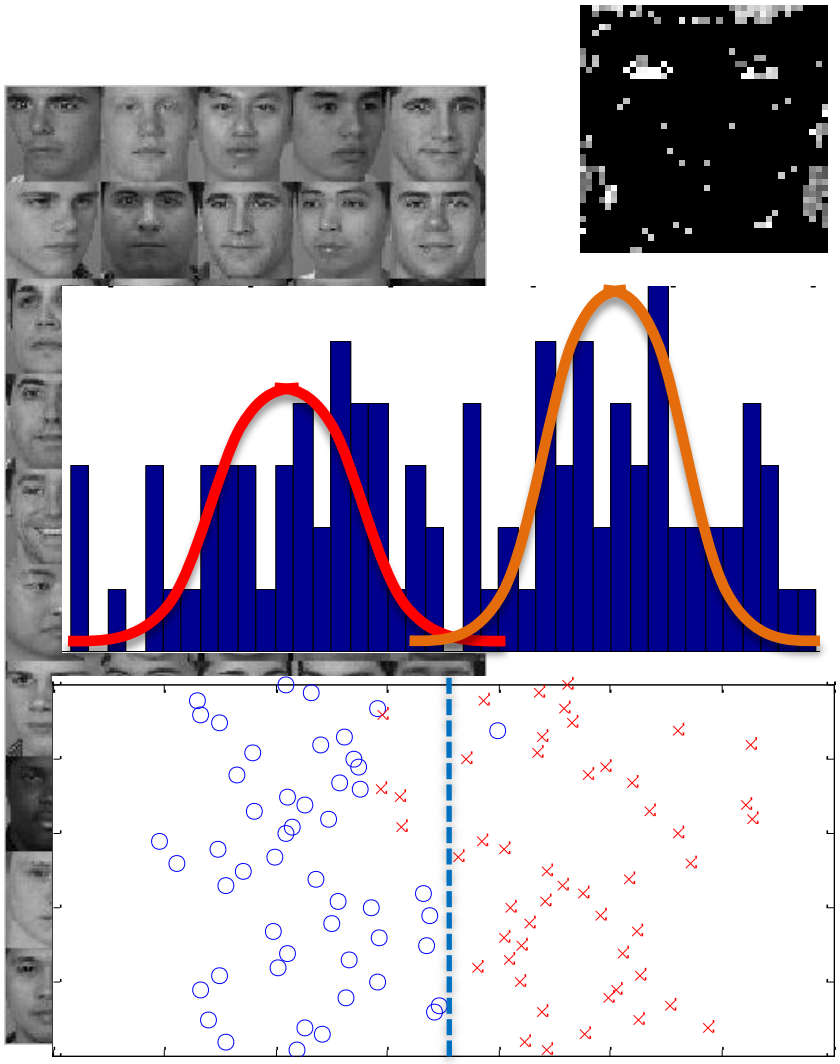
- Visualize spatial relationship of samples
-
- Preserve specific information
-



th of
on of

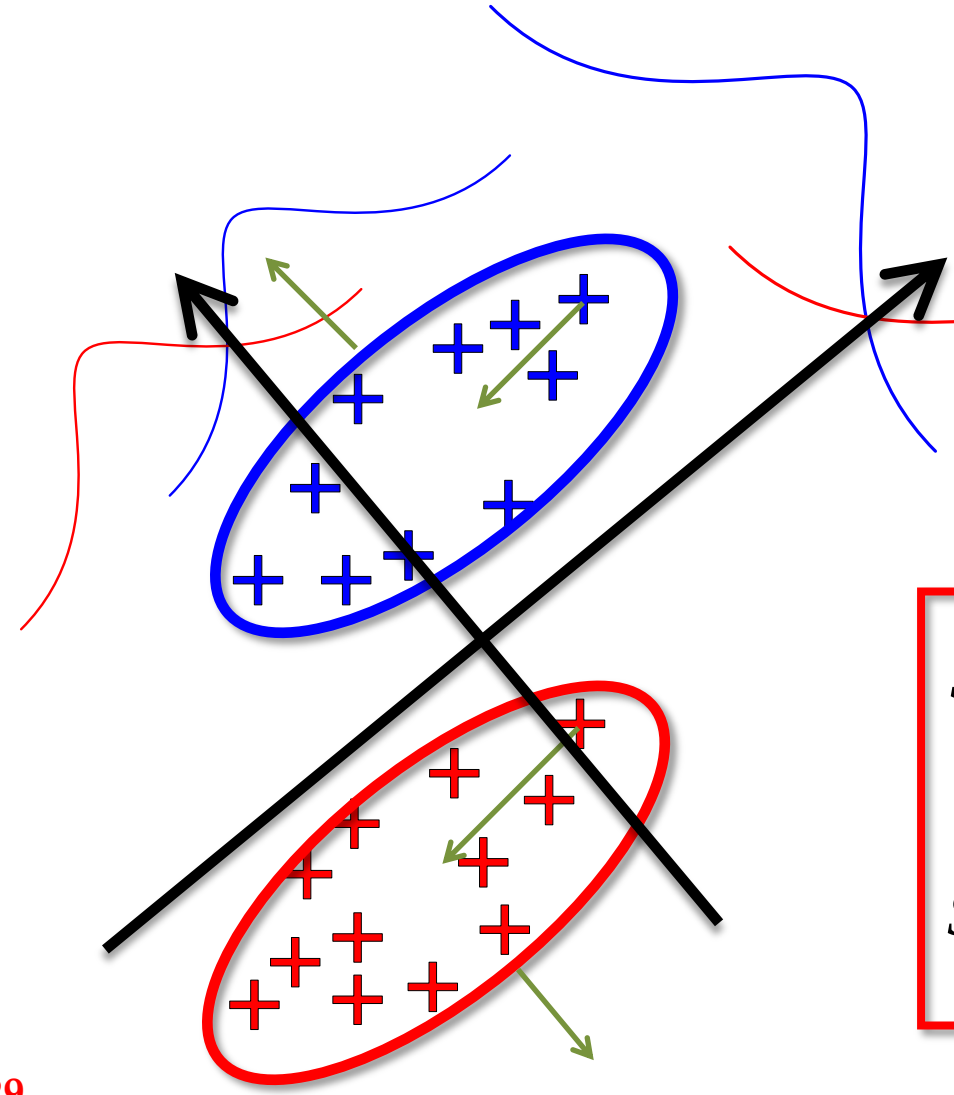
dimensionality

Importance



28

Fisher's Linear Discriminant Analysis



SVD

$$U^* = \arg\max_U \operatorname{tr}\left(\left(U^T S_w U\right)^{-1}\left(U^T S_b U\right)\right)$$

$$S_b = \frac{1}{n}\sum_{i=1}^{c} n_i \left(\vec{m}_i - \vec{m}\right)\left(\vec{m}_i - \vec{m}\right)^T$$

$$S_w = \frac{1}{n}\sum_{i=1}^{c}\sum_{j=1}^{n_i}\left(\vec{x}_{i;j} - \vec{m}_i\right)\left(\vec{x}_{i;j} - \vec{m}_i\right)^T$$

What is FLDA?

$$U^* = \arg \int \min_U \left[\pi_1 p_1 \left(U, \vec{x} \right), \pi_2 p_2 \left(U, \vec{x} \right) \right] d\vec{x}$$ Bayes error minimization

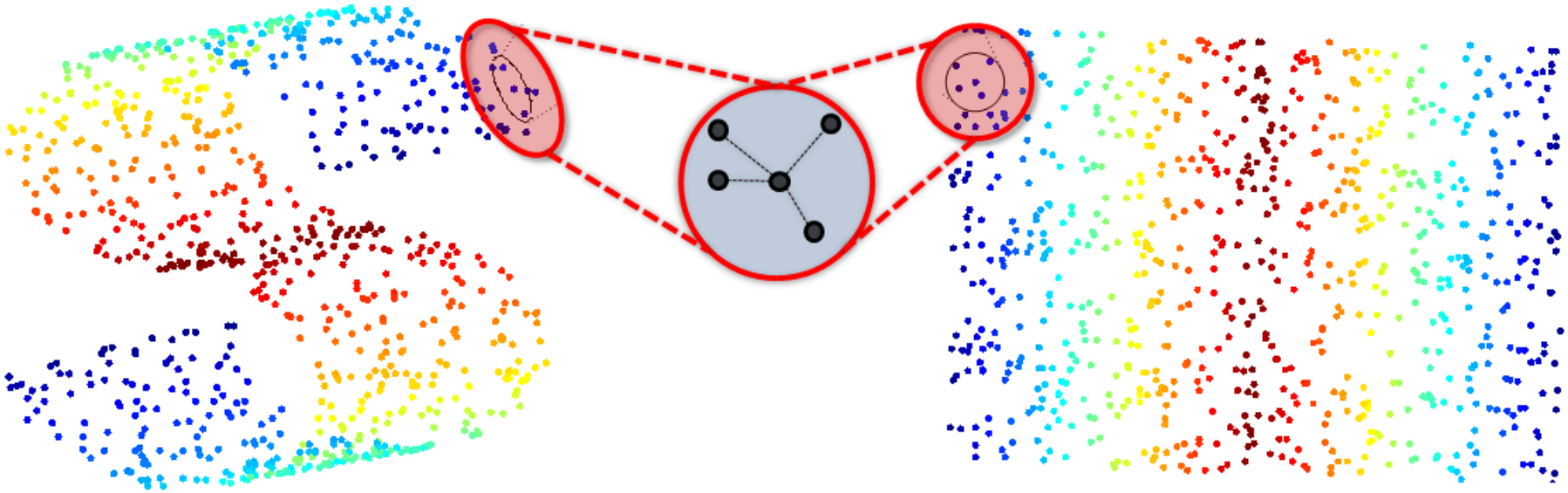
$$\leq \arg \min_U \log \int p_1^{1/2} \left(U, \vec{x} \right) p_2^{1/2} \left(U, \vec{x} \right) d\vec{x}$$ Bhattacharyya upper bound

$$= \arg \min_U \left[-\frac{1}{2} \mathrm{tr} \left(\left(U^T S_w U \right)^{-1} \left(U^T S_b U \right) \right) - \frac{1}{2} \log \frac{\left| U^T \left(\frac{\Sigma_1 + \Sigma_2}{2} \right) U \right|}{\left| U^T \Sigma_1 U \right|^{1/2} \left| U^T \Sigma_2 U \right|^{1/2}} \right]$$

What Can Dimension Reduction Do?

- Not only a pre-processing method
- Classification/regression
- Clustering (Spectral clustering)

- Distance metric learning
  - rank deficient distance metric learning

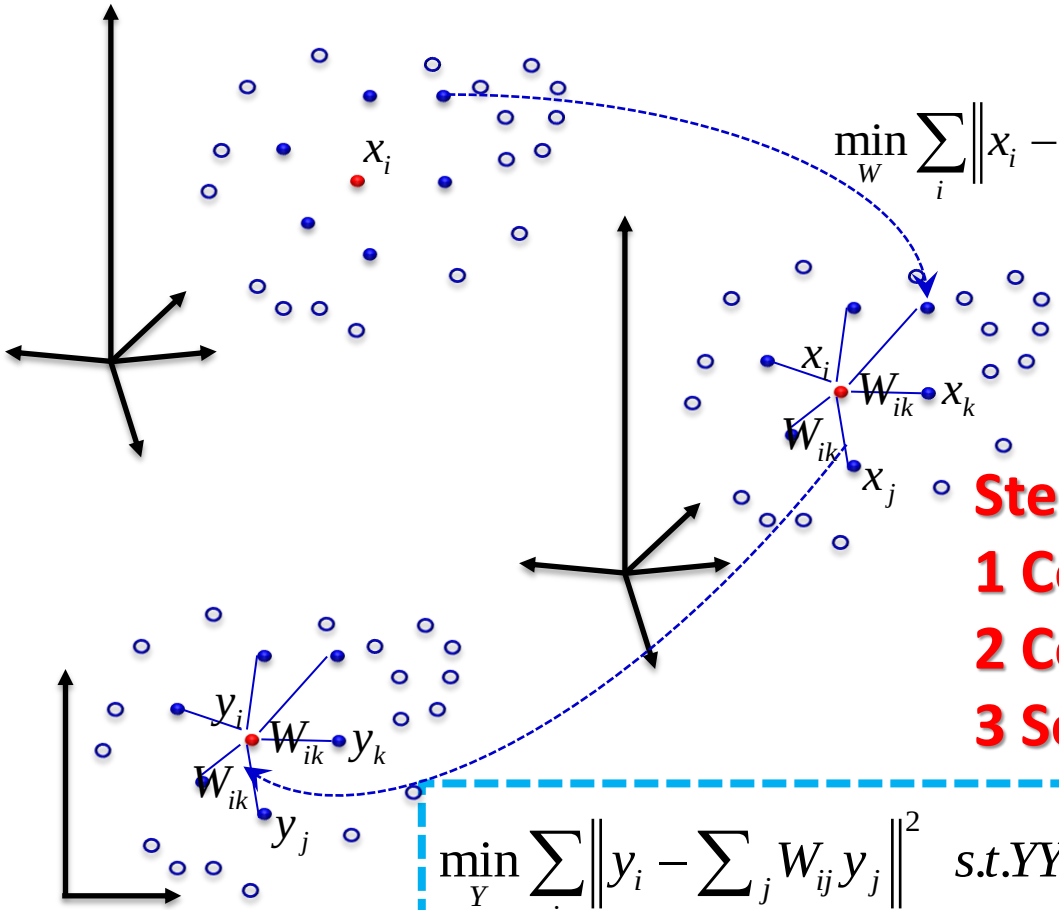


Patch Alignment Framework for Improving the Performance

PART THREE

UTS:QCIS
QUANTUM COMPUTATION & INTELLIGENT SYSTEMS

Locally linear embedding



$$\min_W \sum_i \left\| x_i - \sum_j W_{ij} x_j \right\|^2$$

Step:

1 Construct neighborhood graph

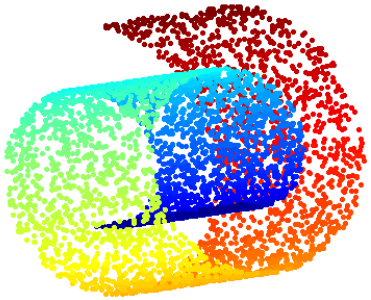
2 Compute reconstruction weights

3 Solve *d*-dimensional embedding

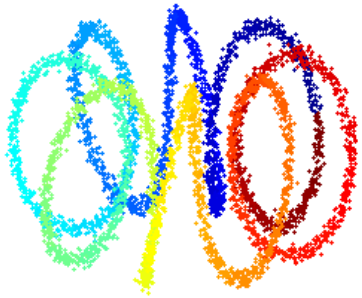
$$\min_Y \sum_i \left\| y_i - \sum_j W_{ij} y_j \right\|^2 \quad s.t. YY^T = I$$

Sam Roweis and Lawrence Saul, "Nonlinear Dimensionality Reduction by Locally Linear Embedding," ***Science***, v.290 no.5500 , Dec.22, 2000. pp.2323--2326.

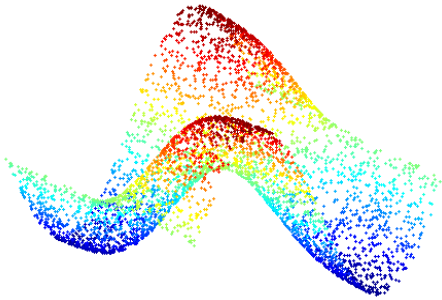
X



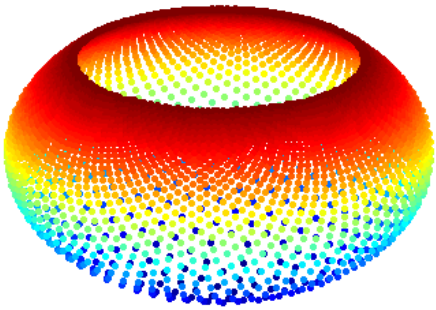
Swiss roll



Toroidal helix



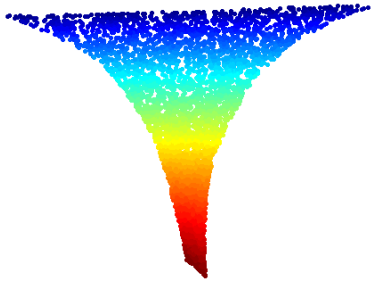
Two peaks

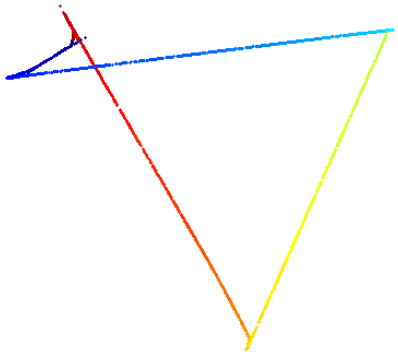


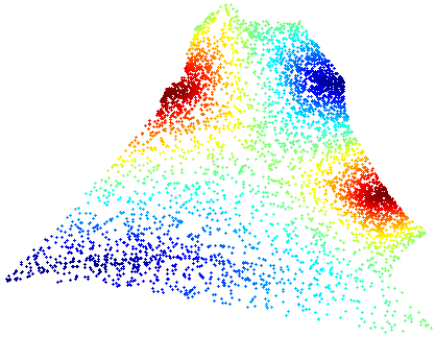
Punctured sphere

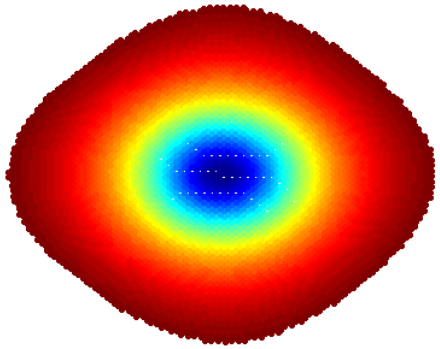
Dimension Reduction

Y

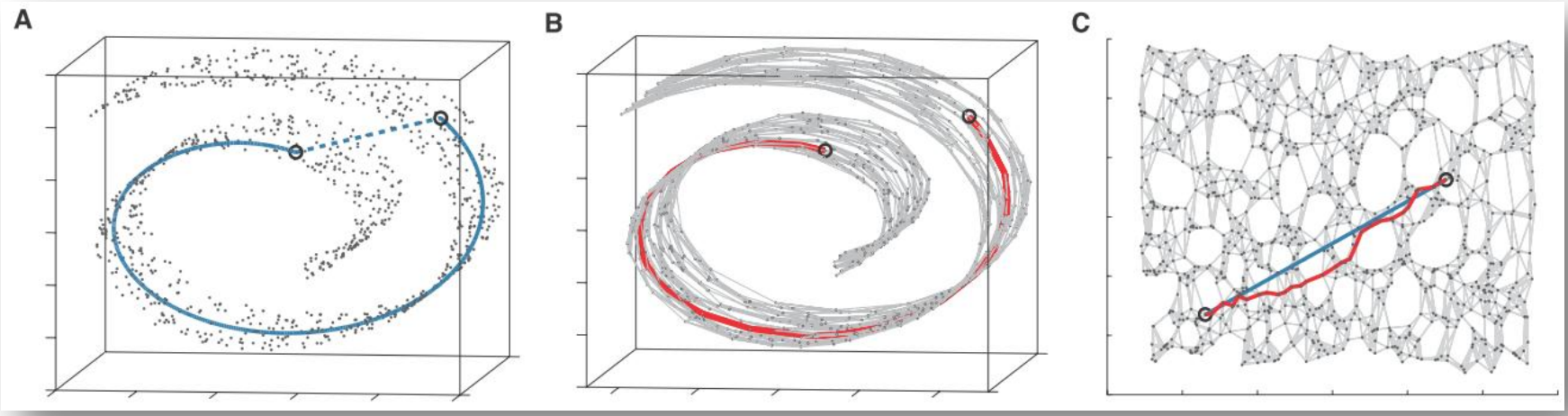








ISOMAP



Step:

1 Construct neighborhood graph

2 Compute shortest path matrix D_Y

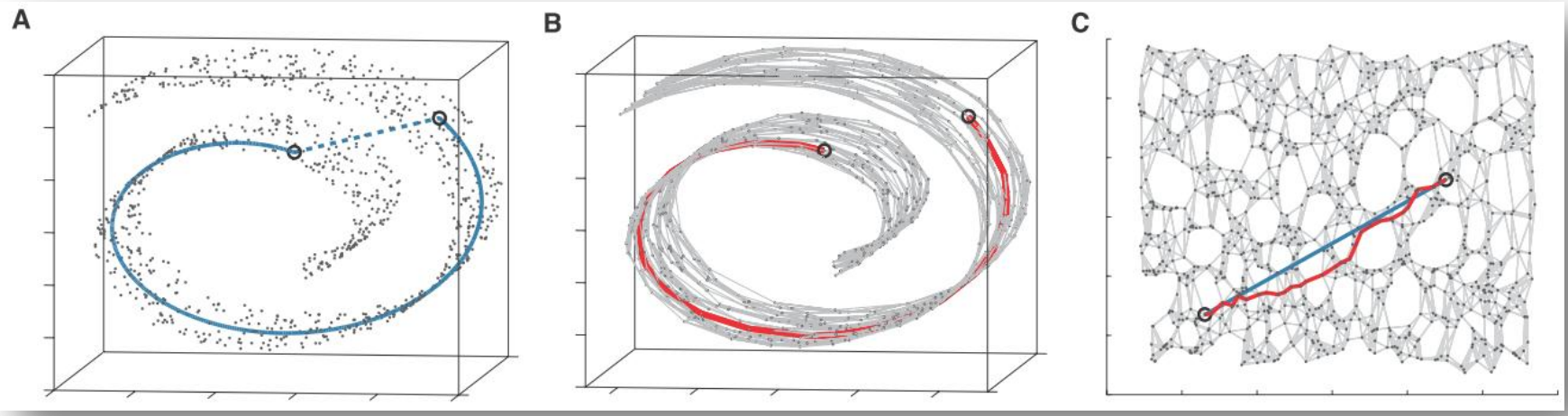
3 Construct *d*-dimensional embedding

$$\min \left\| \tau(D_G) - \tau(D_Y) \right\|^2$$

$$s.t.\ YY^T = diag(\lambda_1, \cdots, \lambda_d)$$

J. B. Tenenbaum, V. de Silva and J. C. Langford, "A Global Geometric Framework for Nonlinear Dimensionality Reduction," ***Science***, 290 (5500): 2319-2323, 22 December 2000

ISOMAP



$$\min \left\| \tau(D_G) - \tau(D_Y) \right\|^2$$
$$s.t.\ YY^T = diag(\lambda_1, \cdots, \lambda_d)$$

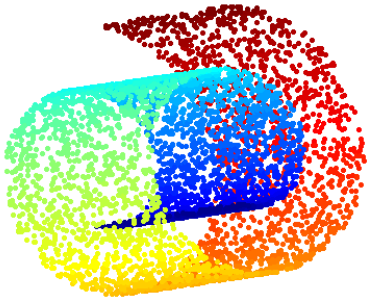
$$D_Y : \left\{ d_Y(i,j) = \left\| y_i - y_j \right\|^2 \right\}$$ **Euclidean distances between Embeddings**

$$D_G = \{ d_G(i,j) \}$$ **Geodesic distances between Original data**

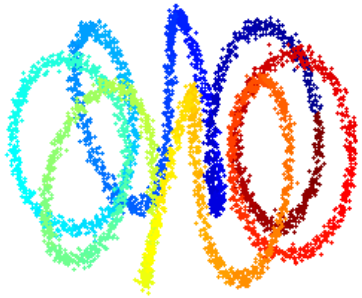
$$\tau(D) = -HSH/2 \quad \{S_{ij} = D_{ij}^2\} \quad \{H_{ij} = \delta_{ij} - 1/N\}$$

J. B. Tenenbaum, V. de Silva and J. C. Langford, "A Global Geometric Framework for Nonlinear Dimensionality Reduction," ***Science***, 290 (5500): 2319-2323, 22 December 2000

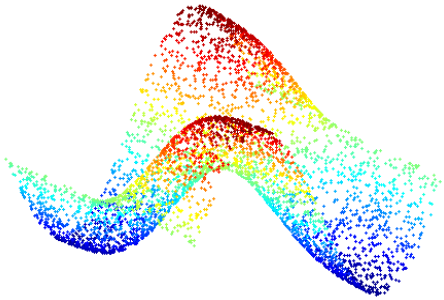
X



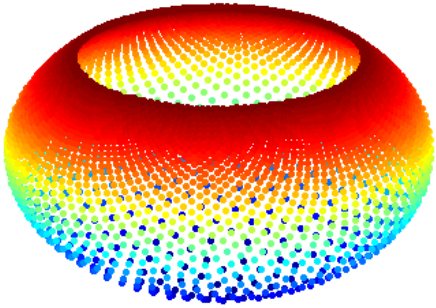
Swiss roll



Toroidal helix



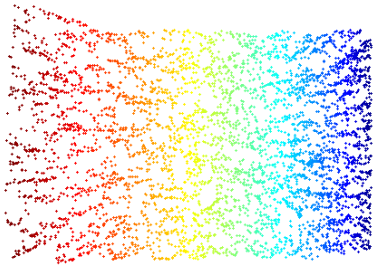
Two peaks

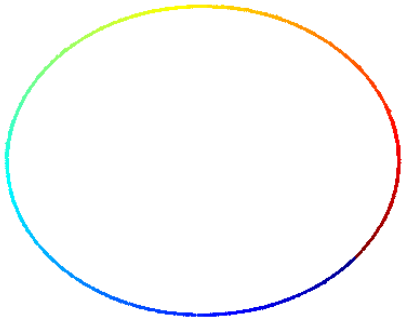


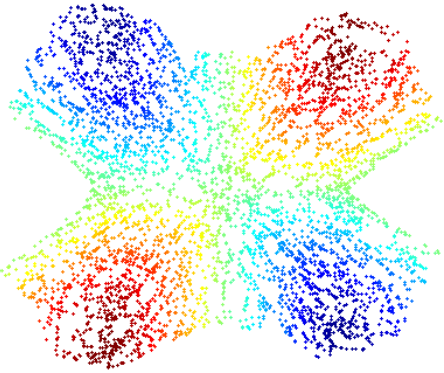
Punctured sphere

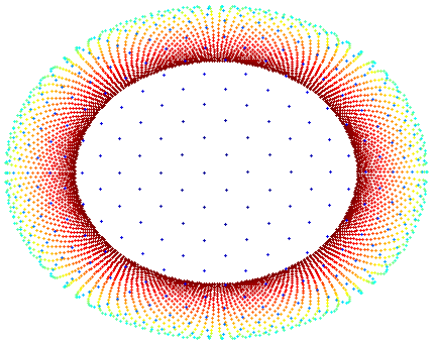
Dimension Reduction

Y

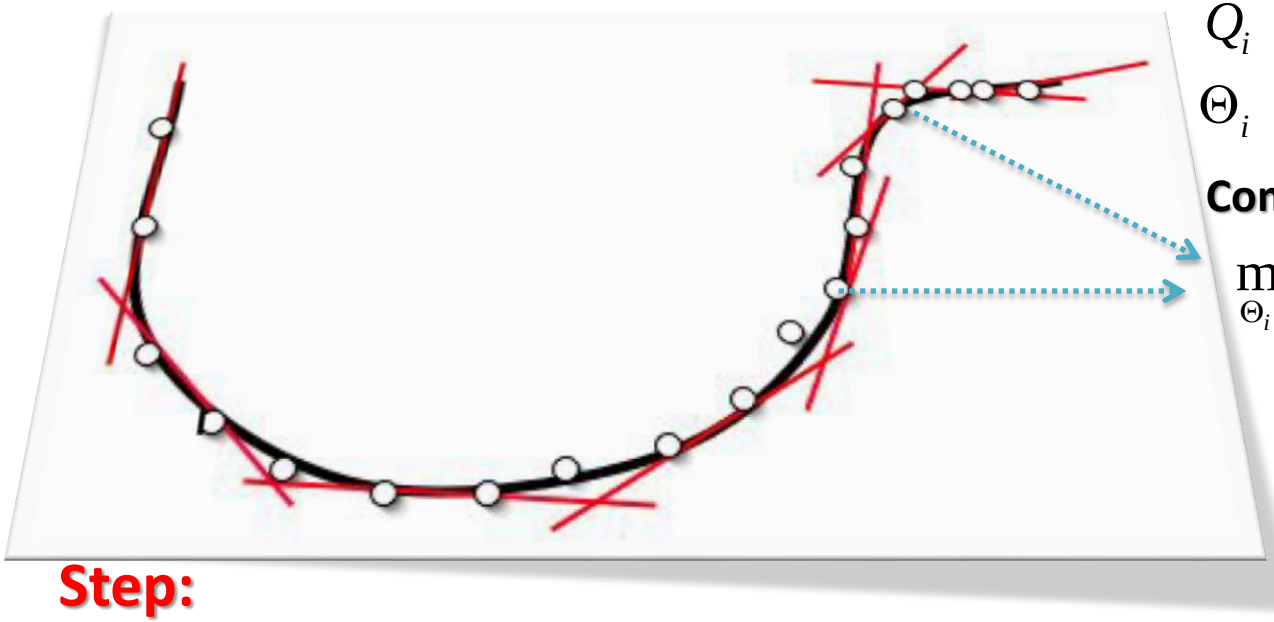








Local tangent space alignment



Q_i **Basis of local tangent plane**

Θ_i **Local tangent coordinates**

Compute tangent plane

$$\min_{\Theta_i, Q_i} \left\| X_i R_{k+1} - Q_i \Theta_i \right\|^2$$

$$R_{k+1} = I_{k+1} - e_{k+1} e_{k+1}^T / (k+1)$$

centralization matrix

Step:

1 Construct neighborhood graph

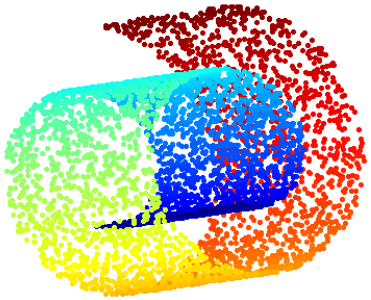
2 Compute tangent coordinates

3 Construct *d*-dimensional embedding

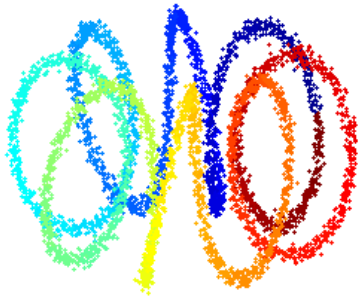
$$\min_{Y_i, T_i} \sum_i \left\| Y_i R_{k+1} - T_i \Theta_i \right\|^2$$

Zhenyue Zhang and Hongyuan Zha, "Principal Manifolds and Nonlinear Dimension Reduction via Local Tangent Space Alignment," ***SIAM Journal of Scientific Computing***, 2002

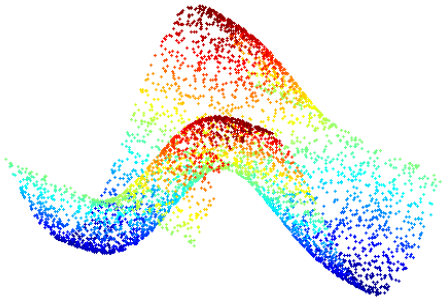
X



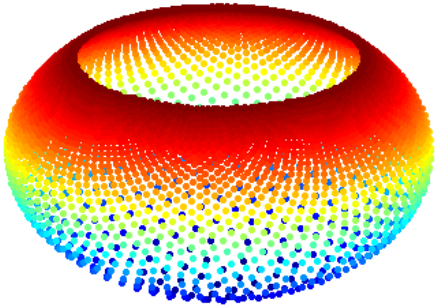
Swiss roll



Toroidal helix



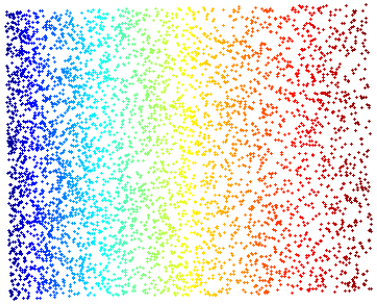
Two peaks

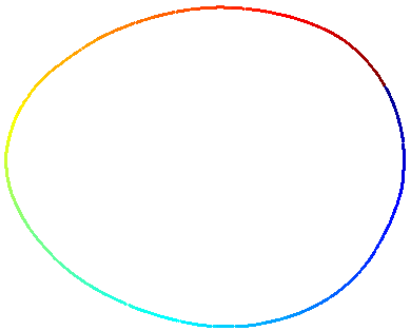


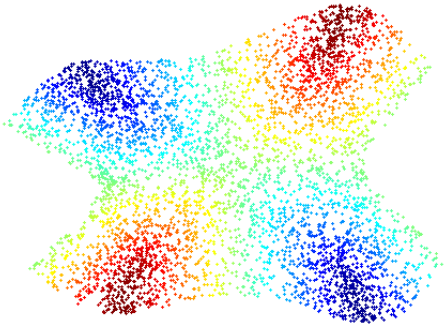
Punctured sphere

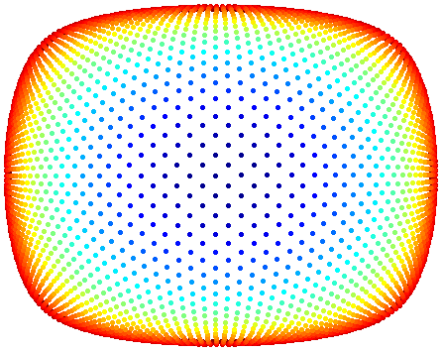
Dimension Reduction

Y

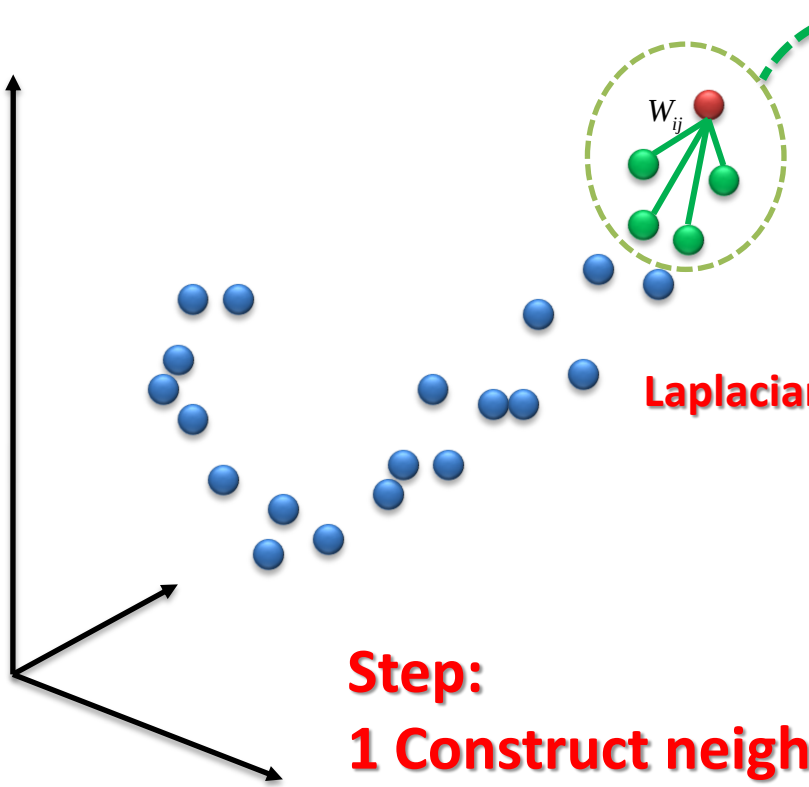








Laplacian eigenmaps



Graph

$$W_{ij} = \begin{cases} \exp(-\|x_i - x_j\|^2 / t), & i \in N(j) \text{ or } j \in N(i) \\ 0, & \text{others} \end{cases}$$

Laplacian matrix $L = D - W, D_{ii} = \sum_j W_{ij}$

$$\min_{Y^T DY = I} \sum_{i=1}^{N} \sum_{j=1}^{N} \| y_i - y_{i_j} \|^2 W_{ij}$$

Step:

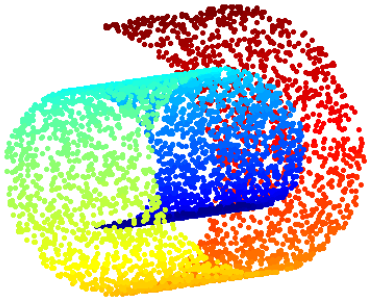
1 Construct neighborhood graph

2 Compute Laplacian matrix

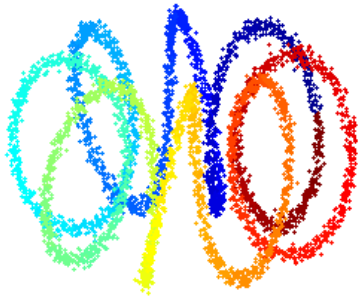
3 Solve *d*-dimensional embedding

Mikhail Belkin and Partha Niyogi, "Laplacian Eigenmaps for Dimensionality Reduction and Data Representation," ***Neural Computation***, 2003 15(6), 1373-1396

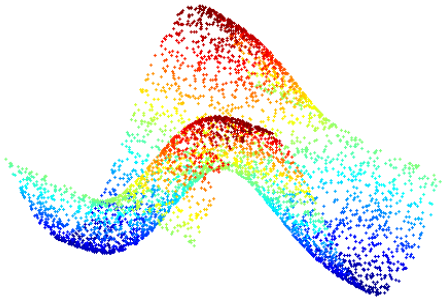
X



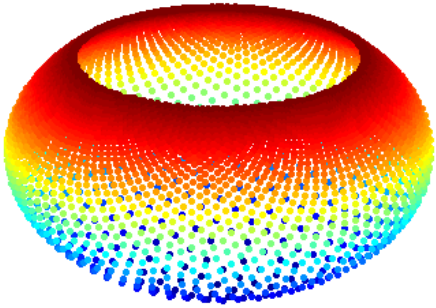
Swiss roll



Toroidal helix



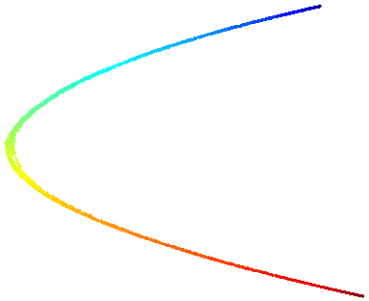
Two peaks

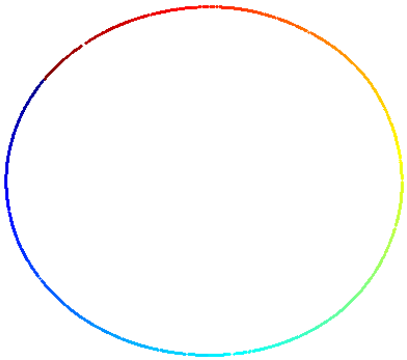


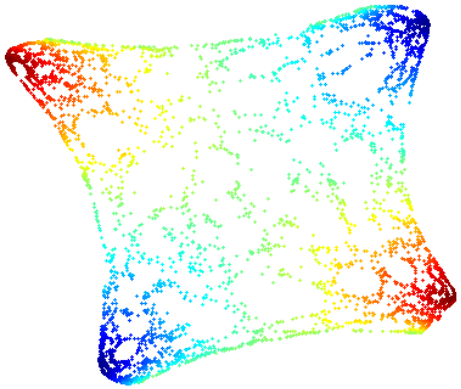
Punctured sphere

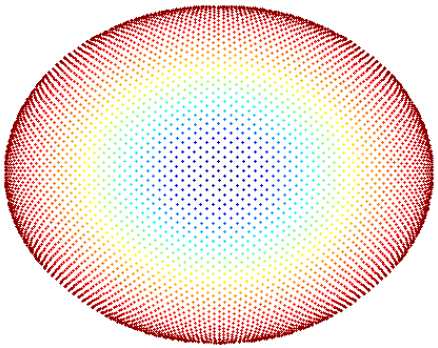
Dimension Reduction

Y

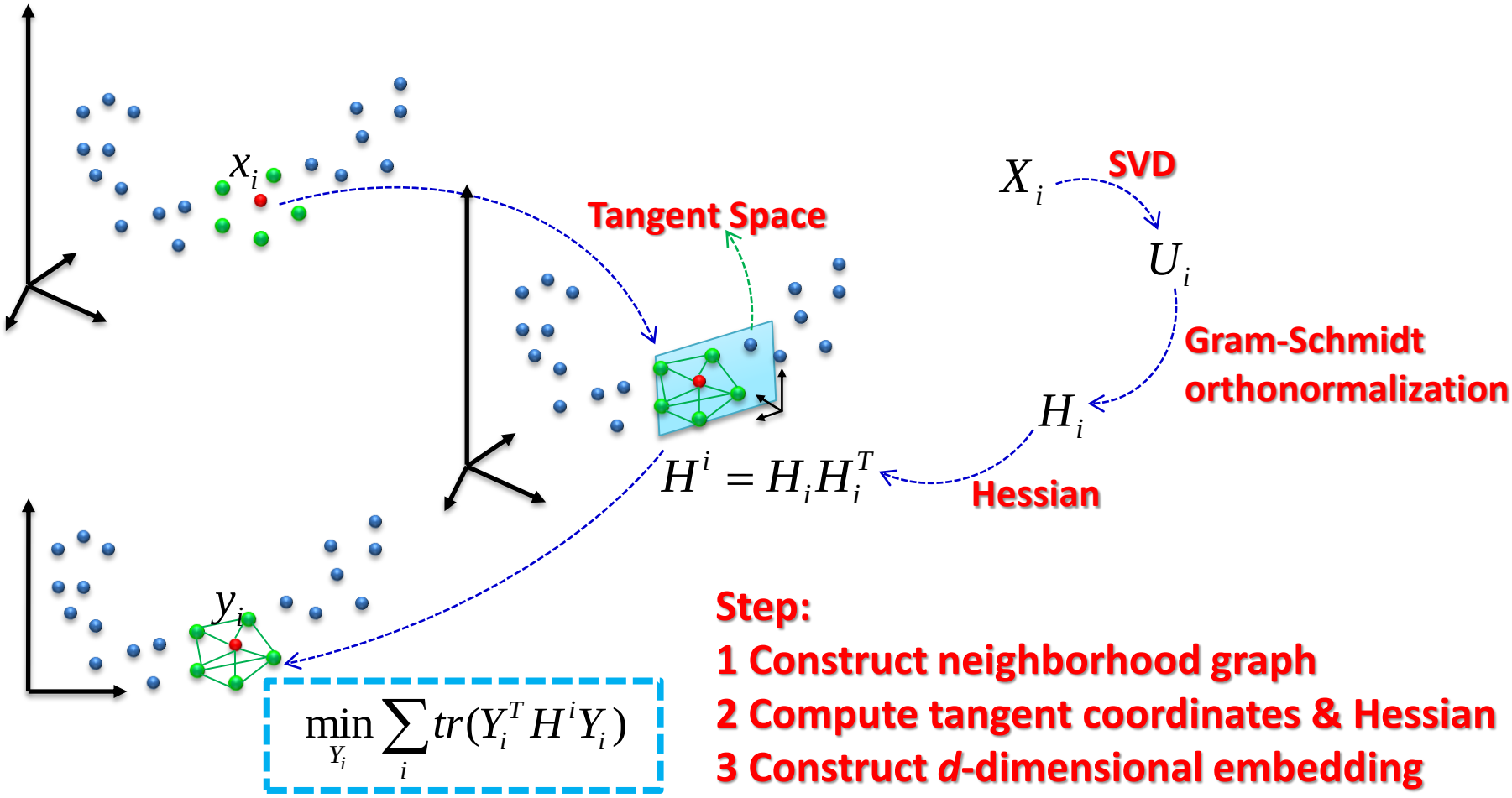








Hessian eigenmaps



Step:

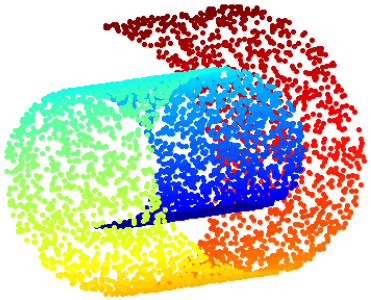
1 Construct neighborhood graph

2 Compute tangent coordinates & Hessian

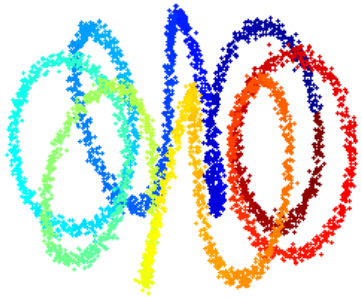
3 Construct *d*-dimensional embedding

D. L. Donoho and C. Grimes, "Hessian eigenmaps: New Locally Linear Embedding Techniques for High Dimensional Data**," Proceedings of the National Academy of Sciences of the United States of America**, vol.100, no. 10, pp. 5591-5596, 2003.

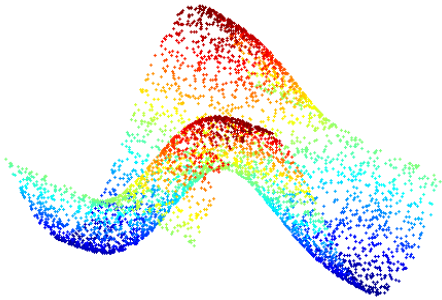
X



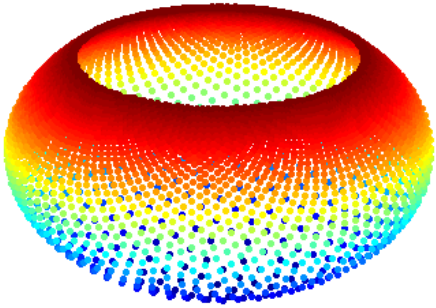
Swiss roll



Toroidal helix



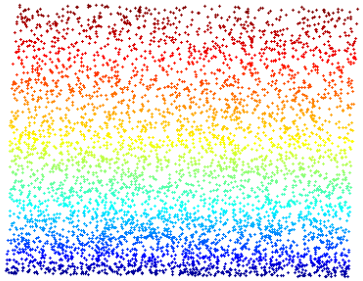
Two peaks

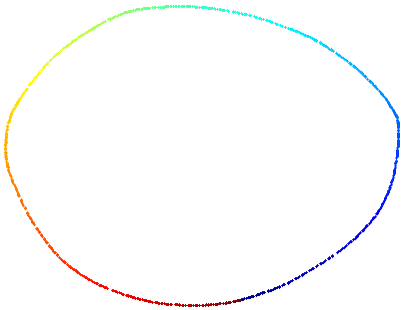


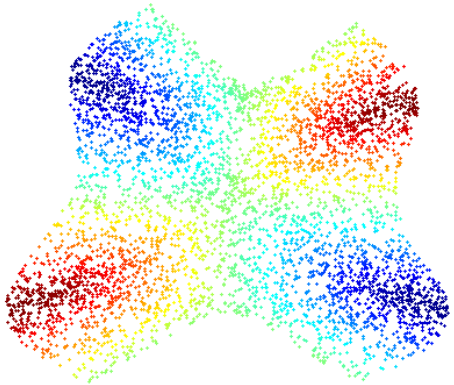
Punctured sphere

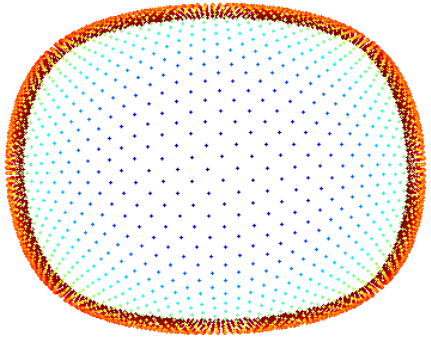
Dimension Reduction

Y

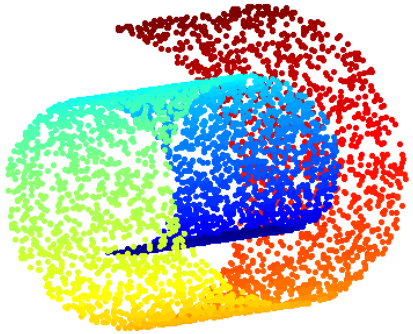




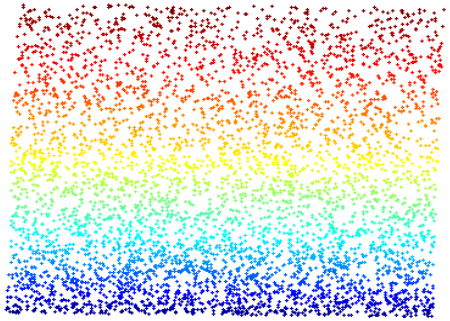




Manifold Learning



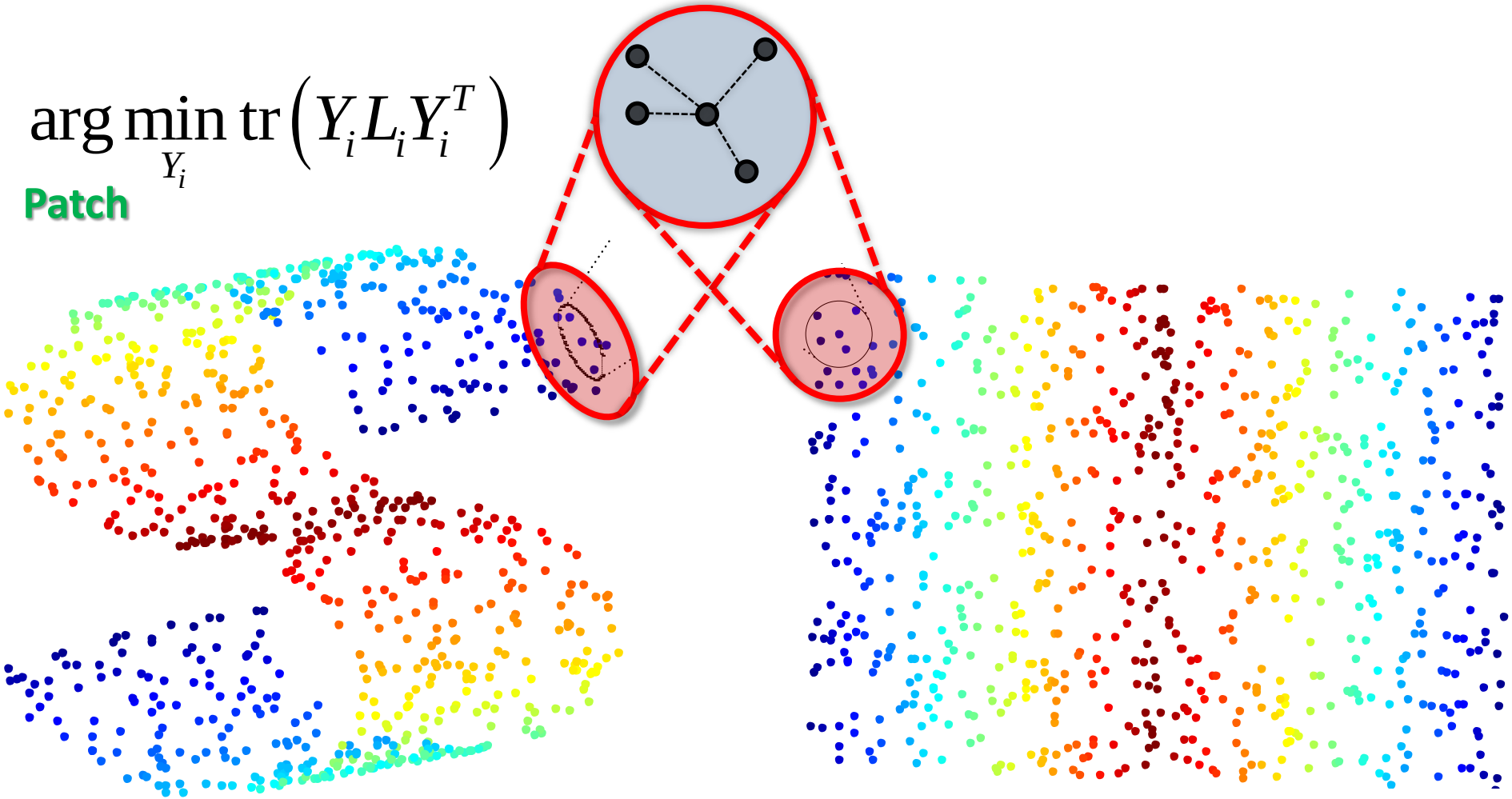
- Construct neighborhood graph
  - Why?
- Compute the characteristics through this neighborhood graph
  - What?
- Construct low-dimensional embedding
  - How?



Patch alignment framework

$$\arg\min_{Y_i} \operatorname{tr}\left(Y_i L_i Y_i^T\right)$$

Patch



Patch alignment formulation

- Part optimization

encodes the objective function

$$\arg\min_{Y_i} \mathrm{tr}\left(Y_i L_i Y_i^T\right)$$

low-dimensional patch $Y_i = \left[y_i, y_{i_1} \cdots, y_{i_k}\right]$

- Whole alignment

$$\arg\min_{Y} \sum_{i=1}^{N} tr\left(YS_i L_i S_i^T Y^T\right) = \arg\min_{Y} tr\left(Y\left(\sum_{i=1}^{N} S_i L_i S_i^T\right) Y^T\right)$$

$$= \arg\min_{Y} tr\left(YLY^T\right),$$

selection matrix: $YS_i = Y_i$

Patch alignment formulation

- Part optimization

encodes the objective function

$$\arg\min_{Y_i} \mathrm{tr}\left(Y_i L_i Y_i^T\right)$$

low-dimensional patch $Y_i = \left[y_i, y_{i_1} \cdots, y_{i_k}\right]$

- Whole alignment

$$\arg\min_{Y} \sum_{i=1}^{N} tr\left(Y S_i L_i S_i^T Y^T\right) = \arg\min_{Y} tr\left(Y\left(\sum_{i=1}^{N} S_i L_i S_i^T\right) Y^T\right)$$

$$= \arg\min_{Y} tr\left(Y L Y^T\right),$$

$$L = \sum_{i=1}^{N} S_i L_i S_i^T \qquad L(F_i, F_i) \leftarrow L(F_i, F_i) + L_i$$

F_i selects the indexes corresponding to patch i

Examples

- **LLE**: Locally linear embedding
- **ISOMAP**: Isometric mapping
- **LTSA**: Local tangent space alignment
- **LE**: Laplacian eigenmaps
- **HLLE**: Hessian eigenmaps
- …

LLE in PAF

reconstruction coefficients in LLE:

$$\arg\min_{W_{i1},\cdots,W_{ik}} \left\| x_i - \sum_{j\in N(i)} W_{ij} x_j \right\|^2$$

$$\arg\min_{Y_i} \operatorname{tr}\left(Y_i L_i Y_i^T\right) = \arg\min_{y_i, y_{i_1},\cdots y_{i_k}} \| y_i - \sum_{j\in N(i), j=1}^{k} W_{ij} y_{i_j} \|^2$$

part optimization representation given by:

$$L_i = \begin{bmatrix} 1 & -W_i^T \\ -W_i & W_i W_i^T \end{bmatrix}$$

$$W_i = [W_{i1}, \cdots, W_{ik}]^T$$

ISOMAP in PAF

inner product matrix of D_G

$$\arg\max_{Y_i} \sum_{i=1}^{N} \mathrm{tr}\left(Y_i L_i Y_i^T\right) \Leftrightarrow \arg\min_{Y} \| \tau(D_G) - Y^T Y \|^2$$

$$\Updownarrow$$

$$\arg\min \sum_{m,n} \left(d_G(m,n) - d'(m,n)\right)^2$$

part optimization representation given by:

$$L_i = \frac{1}{N} \tau\left(D_G^i\right)$$

LTSA in PAF

Compute tangent plane

- **Compute local tangent coordinates**

$$\arg\min_{\Theta_i, Q_i} \left\| X_i R_{k+1} - Q_i \Theta_i \right\|^2$$

$$\Theta_i = Q_i^T X_i R_{k+1}$$

Centralization matrix

$$R_{k+1} = I_{k+1} - e_{k+1} e_{k+1}^T / (k+1)$$

Q_i **Basis of local tangent plane**

Θ_i **Local tangent coordinates**

- Part Optimization

$$\arg\min_{Y_i, T_i} \left\| Y_i R_{k+1} - T_i \Theta_i \right\|^2 \Rightarrow T_i = Y_i R_{k+1} \Theta_i^+$$

part optimization representation given by:

$$L_i = R_{k+1} - V_i V_i^T$$

$$\arg\min_{Y_i} \left\| Y_i R_{k+1} \left(I_{k+1} - \Theta_i^+ \Theta_i \right) \right\|^2$$

***d* right singular vectors of** $X_i R_{k+1}$

$$\arg\min_{Y_i} tr\left(Y_i L_i Y_i^T \right) = \arg\min_{Y_i} \left\| Y_i R_{k+1} \left(I_{k+1} - V_i V_i^T \right) \right\|^2$$

LE in PAF

Heat kernel:

$$(\omega_i)_j = \exp\left(-\| x_i - x_{i_j} \|^2 / t\right)$$

$$\arg\min_{Y_i} \operatorname{tr}\left(Y_i L_i Y_i^T\right) = \arg\min_{y_i} \sum_{i=1}^{N}\sum_{j=1}^{l} \| y_i - y_{i_j} \|^2 (\omega_i)_j$$

$$\Uparrow$$

$$\arg\min_{Y} \sum_{i=1}^{N}\sum_{j=1}^{N} \| y_i - y_{i_j} \|^2 W_{ij}$$

$$W_{ij} = \begin{cases} \exp(-\|x_i - x_j\|^2 / t), & i \in N(j) \text{ or } j \in N(i) \\ 0, & \text{others} \end{cases}$$

part optimization representation given by:

$$L_i = \begin{bmatrix} \sum_{j=1}^{l} (\omega_i)_j & -\omega_i^T \\ -\omega_i & diag(\omega_i) \end{bmatrix}$$

HLLE in PAF

Hessian matrix

$$\arg\min_{Y_i} \operatorname{tr}\left(Y_i L_i Y_i^T\right) = \arg\min_{Y_i} \operatorname{tr}\left(Y_i H_i H_i^T Y_i^T\right)$$

part optimization representation given by:

$$L_i = H_i H_i^T$$

PAF for Understanding PCA

Objective function

$$\arg\max_{y_i} \mathrm{tr}\left(S_T\right) = \arg\max_{y_i} \mathrm{tr}\left(\sum_{i=1}^{N}\left(y_i - y^m\right)\left(y_i - y^m\right)^T\right)$$

Reformulation

$$\arg\max_{y_i} \sum_{i=1}^{N} \mathrm{tr}\left(\frac{1}{N^2}\left(\sum_{j=1}^{N-1}\left(y_i - y_{i_j}\right)\right)\left(\sum_{j=1}^{N-1}\left(y_i - y_{i_j}\right)\right)^T\right)$$

$$= \arg\max_{Y_i} \sum_{i=1}^{N} \mathrm{tr}\left(\frac{1}{N^2}\left(Y_i\begin{bmatrix} N\text{-}1 \\ -e_{N-1}\end{bmatrix}\right)\left(Y_i\begin{bmatrix} N\text{-}1 \\ -e_{N-1}\end{bmatrix}\right)^T\right)$$

$$= \arg\max_{Y_i} \sum_{i=1}^{N} \mathrm{tr}\left(Y_i L_i Y_i^T\right)$$

$$L_i = \frac{1}{N^2}\begin{bmatrix} N\text{-}1 \\ -e_{N-1}\end{bmatrix}\begin{bmatrix} N\text{-}1 & -e_{N-1}{}^T\end{bmatrix}$$

PAF for Understanding FLDA

Objective function

Projection matrix

$$\arg\max_{Y} \frac{\mathrm{tr}\left(S_B\right)}{\mathrm{tr}\left(S_W\right)} \quad s.t.\ \ U^T U = I$$

Scatter matrix within class

$$S_W = \sum_{i=1}^{C}\sum_{j=1}^{N_i}\left(y_i^{(j)} - y_i^m\right)\left(y_i^{(j)} - y_i^m\right)^T$$

Scatter matrix between classes

$$S_B = \sum_{i=1}^{C} N_i \left(y_i^m - y^m\right)\left(y_i^m - y^m\right)^T$$

$$y_i = U^T x_i$$

Objective Reformulation

$$\begin{cases} \arg\max_{Y} \mathrm{tr}\left(S_B\right) \\ \arg\min_{Y} \mathrm{tr}\left(S_W\right) \end{cases} \quad s.t.\ \ U^T U = I$$

PAF for Understanding FLDA

$$\arg\min_{Y} \operatorname{tr}\left(S_W\right) = \arg\min_{y_i^{(j)}} \sum_{i=1}^{C}\sum_{j=1}^{N_i}\left(y_i^{(j)} - y_i^m\right)\left(y_i^{(j)} - y_i^m\right)^T$$

$$= \arg\min_{y_i} \sum_{i=1}^{N} \frac{1}{N_i^2}\left(\sum_{j=1}^{N_i}\left(y_i - y_{i_j}\right)\left(y_i - y_{i_j}\right)^T\right)$$

$$= \arg\min_{Y_i} \sum_{i=1}^{N} \operatorname{tr}\left(\frac{1}{N_i^2}\left(Y_i\begin{bmatrix} N_i - 1 \\ -e_{N_i - 1}\end{bmatrix}\right)\left(Y_i\begin{bmatrix} N_i - 1 \\ -e_{N_i - 1}\end{bmatrix}\right)^T\right)$$

$$= \arg\min_{Y_i} \sum_{i=1}^{N} \operatorname{tr}\left(Y_i L_i^W Y_i^T\right) \qquad L_i^W = \frac{1}{N_i^2}\begin{bmatrix} N_i - 1 \\ -e_{N-1}\end{bmatrix}\begin{bmatrix} N_i - 1 & -e_{N_i - 1}{}^T\end{bmatrix}$$

$$L^W(F_i, F_i) \leftarrow L^W(F_i, F_i) + L_i^W$$

F_i **selects the indexes corresponding to patch** i

PAF for Understanding FLDA

$$\arg\max_{Y} \operatorname{tr}\left(S_B\right) = \arg\max_{y_i^m} \sum_{i=1}^{C} N_i \left(y_i^m - y^m\right)\left(y_i^m - y^m\right)^T$$

$$= \arg\max_{y_i} \left(\sum_{i=1}^{C} N_i \frac{1}{C^2} \left(\sum_{j=1}^{C\text{-}1} \left(y_i^m - y_{i_j}^m \right) \right) \left(\sum_{j=1}^{C\text{-}1} \left(y_i^m - y_{i_j}^m \right) \right)^T \right)$$

$$= \arg\max_{Y_i^m} \sum_{i=1}^{C} \operatorname{tr} \left(\frac{N_i}{C^2} \left(Y_i^m \begin{bmatrix} C-1 \\ -e_{C-1} \end{bmatrix} \right) \left(Y_i^m \begin{bmatrix} C-1 \\ -e_{C-1} \end{bmatrix} \right)^T \right)$$

$$= \arg\max_{Y_i^m} \sum_{i=1}^{C} \operatorname{tr} \left(Y_i^m L_i^B \left(Y_i^m \right)^T \right)$$

$$L_i^B = \frac{N_i}{C^2} \begin{bmatrix} C\text{-}1 \\ -e_{\text{C-1}} \end{bmatrix} \begin{bmatrix} C\text{-}1 & -e_{\text{C-1}} \end{bmatrix}$$

$$L^B(F_i, F_i) \leftarrow L^B(F_i, F_i) + L_i^B$$

F_i **selects the indexes corresponding to patch** i

PAF for Understanding FLDA

Objective function

$$\arg\max_{U} \frac{\mathrm{tr}\left(S_B\right)}{\mathrm{tr}\left(S_W\right)} \quad s.t.\ \ U^T U = I$$

$$\arg\max_{U} \frac{\mathrm{tr}\left(Y^m L^B \left(Y^m\right)^T\right)}{\mathrm{tr}\left(Y L^W Y^T\right)} \quad s.t.\ \ U^T U = I$$

Out-of-Sample Problem

- The above methods (LLE, ISOMAP, LTSA, LE, HLLE) cannot generalize to testing samples since there are **no explicit mapping relation** from *X* to *Y*
- **Recompute** the whole process if apply to new testing samples

Linearization for Out-of-Sample

- Part optimization

$$\arg\min_{Y_i} \operatorname{tr}\left(Y_i L_i Y_i^T\right)$$

Linearization

$$Y_i = U^T X_i$$

$$\arg\min_{U} \operatorname{tr}\left(U^T X_i L_i X_i^T U\right)$$

- Whole alignment

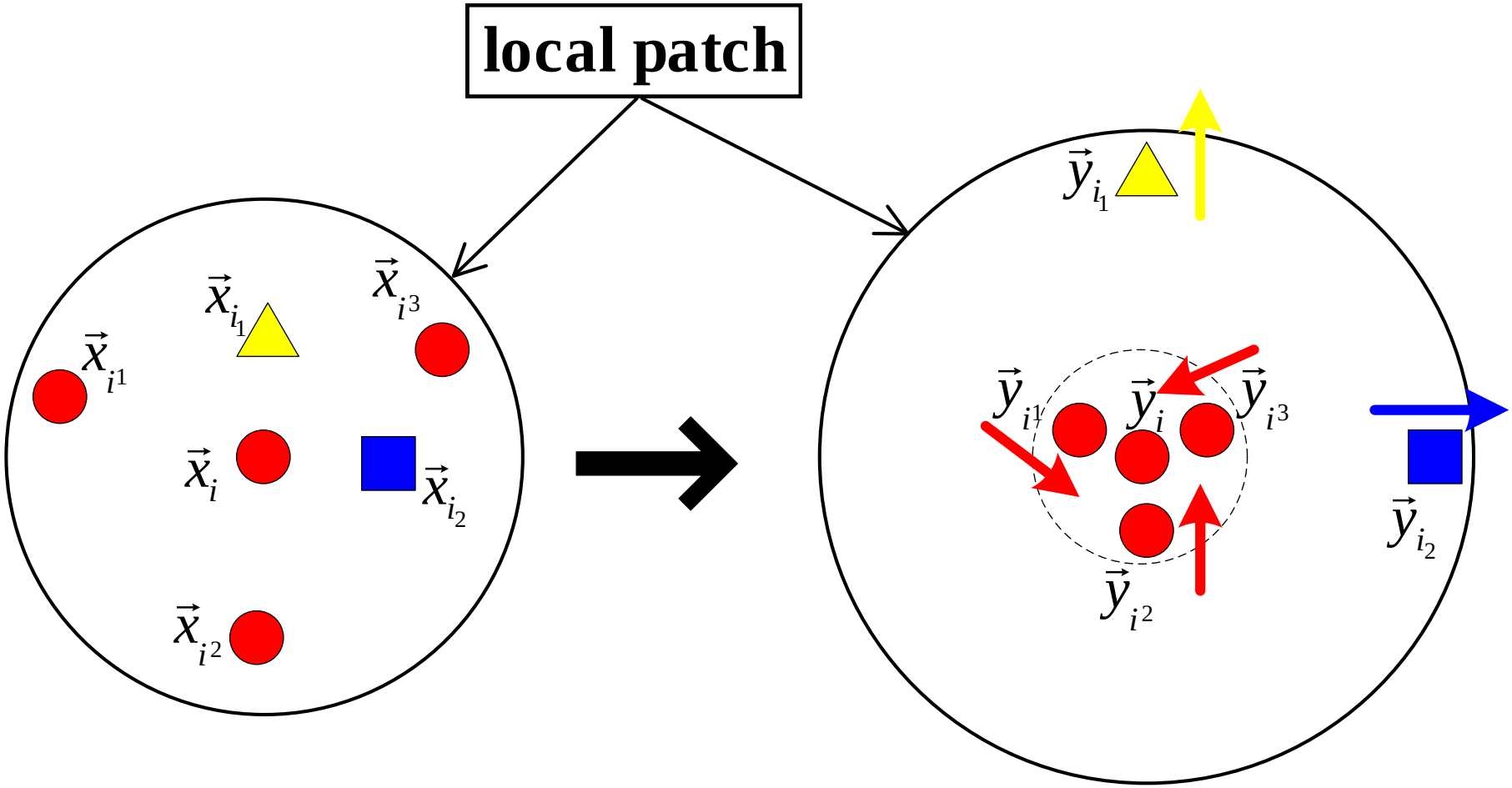
$$\arg\min_{Y} \sum_{i=1}^{N} \operatorname{tr}\left(Y S_i L_i S_i^T Y^T\right) = \arg\min_{Y} \operatorname{tr}\left(Y\left(\sum_{i=1}^{N} S_i L_i S_i^T\right) Y^T\right)$$

$$= \arg\min_{U} \operatorname{tr}\left(U^T X L X^T U\right)$$

Examples

- **NPE (ONPP)**: Neighbor preserving embedding
- **ISOMap ??**
- **LLTSA**: Linear local tangent space alignment
- **LPP**: Locality preserving projections
- **HLLE ??**
- …

Discriminative locality alignment



Discriminative locality alignment

Within Classes

$$\arg\min_{y_i} \sum_{j=1}^{k_1} \left\| y_i - y_{i^j} \right\|^2$$

Between Classes

$$\arg\max_{y_i} \sum_{p=1}^{k_2} \left\| y_i - y_{i_p} \right\|^2$$

Part Optimization

$$\arg\min_{y_i} \left(\sum_{j=1}^{k_1} \left\| y_i - y_{i^j} \right\|^2 - \beta \sum_{p=1}^{k_2} \left\| y_i - y_{i_p} \right\|^2 \right)$$

Discriminative locality alignment

Part Optimization Reformulation

$$\arg\min_{y_i}\left(\sum_{j=1}^{k_1}\left\|y_i - y_{i^j}\right\|^2 - \beta\sum_{p=1}^{k_2}\left\|y_i - y_{i_p}\right\|^2\right)$$

$$\omega_i = \left[\overbrace{1,\cdots,1}^{k_1}\ \overbrace{-\beta,\ldots,-\beta}^{k_2}\right]^T$$

$$\arg\min_{y_i}\left(\sum_{j=1}^{k_1}\left\|y_i - y_{i^j}\right\|^2 (\omega_i)_j + \sum_{p=1}^{k_2}\left\|y_i - y_{i_p}\right\|^2 (\omega_i)_{p+k_1}\right)$$

$$= \arg\min_{y_i}\left(\sum_{j=1}^{k_1+k_2}\left\|y_{F_i(1)} - y_{F_i(j+1)}\right\|^2 (\omega_i)_j\right)$$

$$L_i = \begin{bmatrix} \sum_{j=1}^{k_1+k_2}(\omega_i)_j & -\omega_i^T \\ -\omega_i & diag(\omega_i) \end{bmatrix}$$

$$= \arg\min_{Y_i} tr\left(Y_i \begin{bmatrix} -e_{k_1+k_2}^T \\ I_{k_1+k_2} \end{bmatrix} diag(\omega_i) \begin{bmatrix} -e_{k_1+k_2} & I_{k_1+k_2} \end{bmatrix} Y_i^T\right)$$

$$= \arg\min_{Y_i} tr(Y_i L_i Y_i^T)$$

Discriminative locality alignment

Whole Alignment

$$\arg\min_{y_1,\cdots,y_N} \sum_{i=1}^{N}\left(\sum_{j=1}^{k_1}\left\|y_i - y_{i^j}\right\|^2 - \beta\sum_{p=1}^{k_2}\left\|y_i - y_{i_p}\right\|^2\right)$$

$$\arg\min_{Y} \sum_{i=1}^{N} tr\left(YS_iL_iS_i^TY^T\right)$$

$$= \arg\min_{Y} tr\left(Y\left(\sum_{i=1}^{N} S_iL_iS_i^T\right)Y^T\right)$$

$$= \arg\min_{Y} tr\left(YLY^T\right)$$

Linearization

$$Y = U^TX$$

$$= \arg\min_{U} tr(U^TXLX^TU) \quad s.t.\ U^TU = I$$

Experimental settings

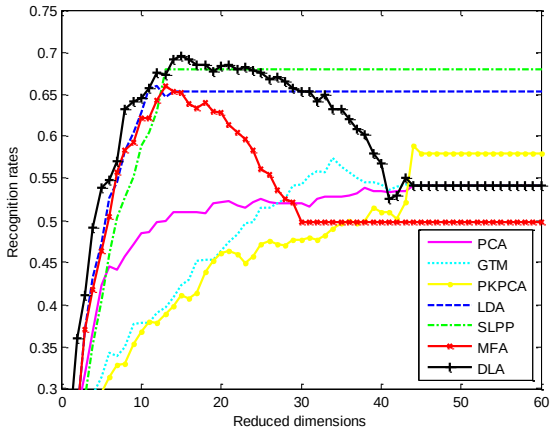
Datasets: FERET, 700 40X40 images, 100 individuals (7)
UMIST, 565 40X40 images, 20 individuals (28)
YALE, 165 40X40 images, 15 individuals (11)
Training: FERET(4,5), UMIST(5,7), YALE(5,7)

Me

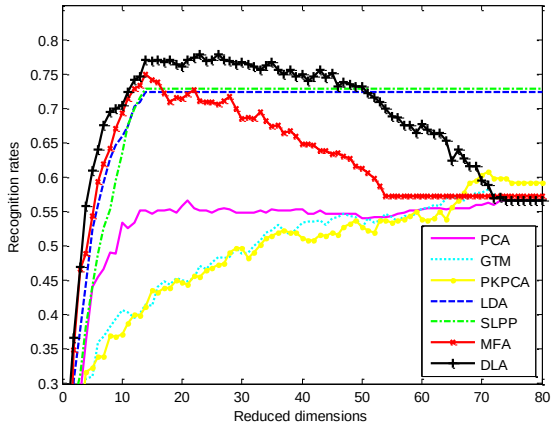
Set

M,

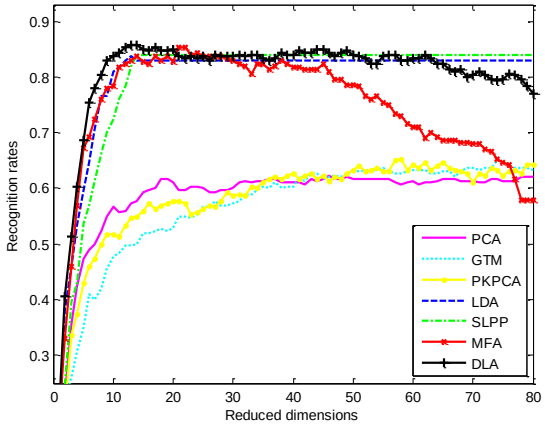
Recognition rate versus dimension



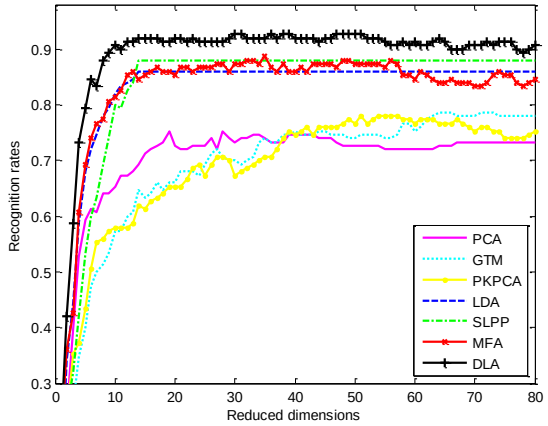
(a)



(b)



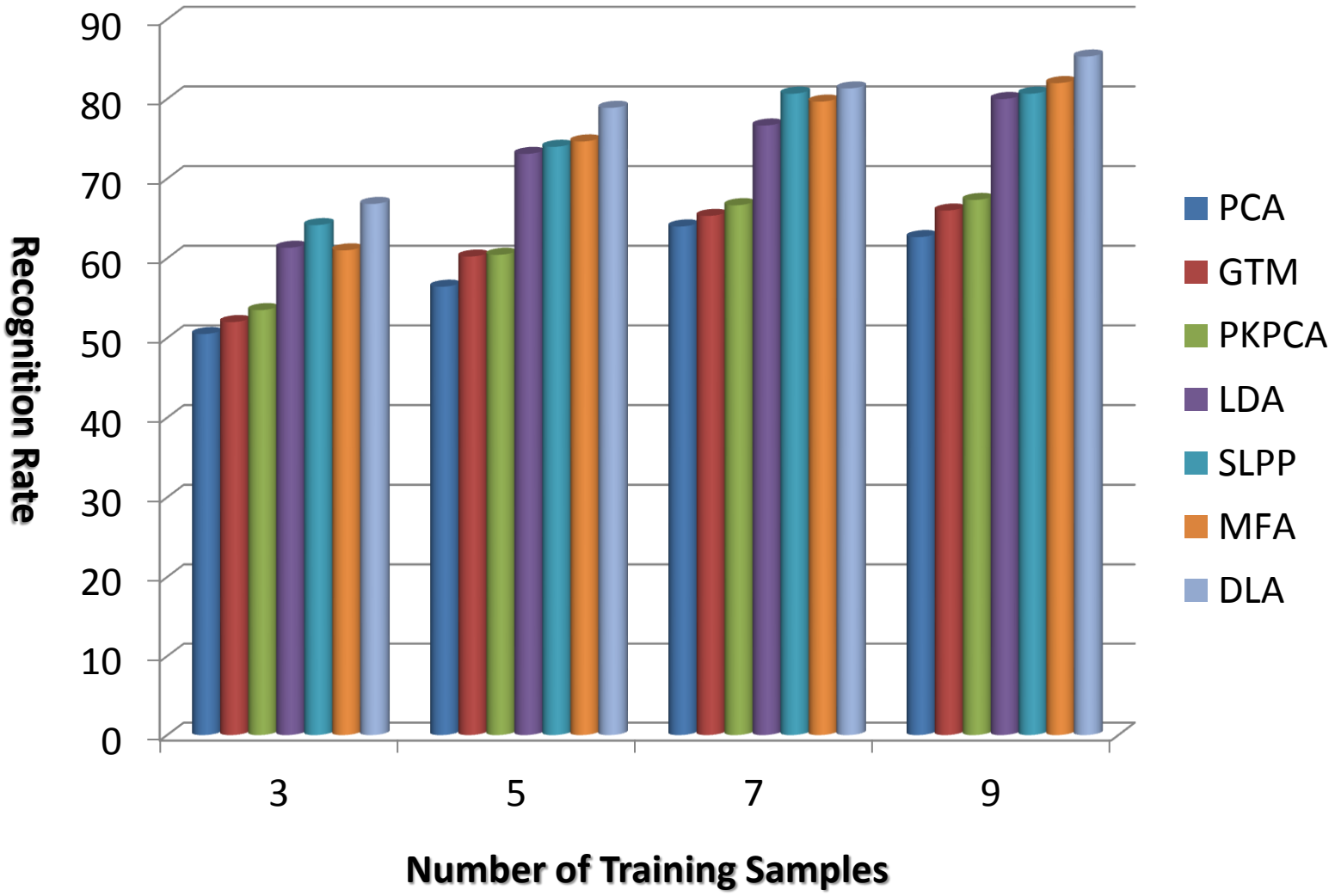
(c)



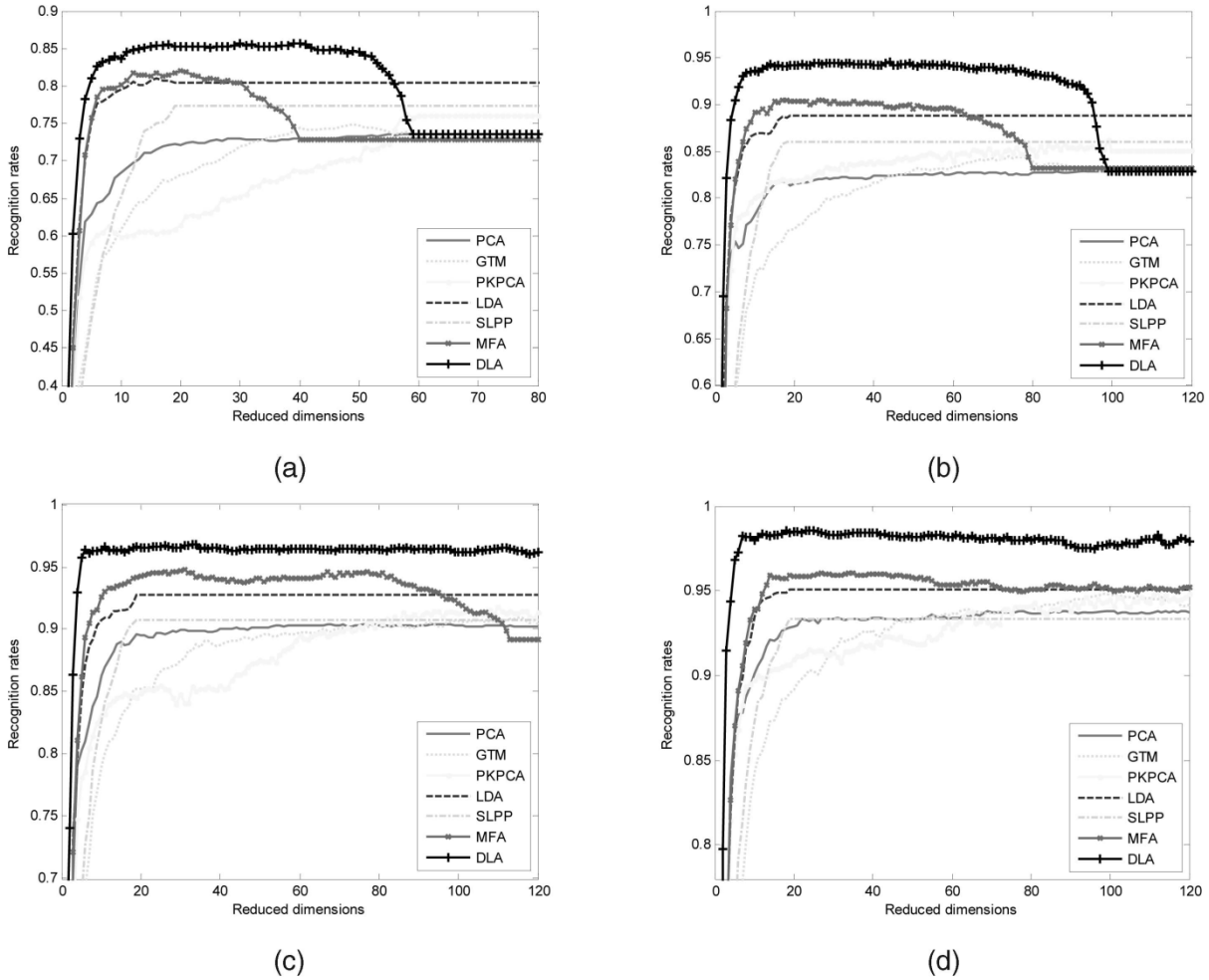
(d)

Recognition rate vs. subspace dimension on Yale dataset

Best Recognition Rates on Yale

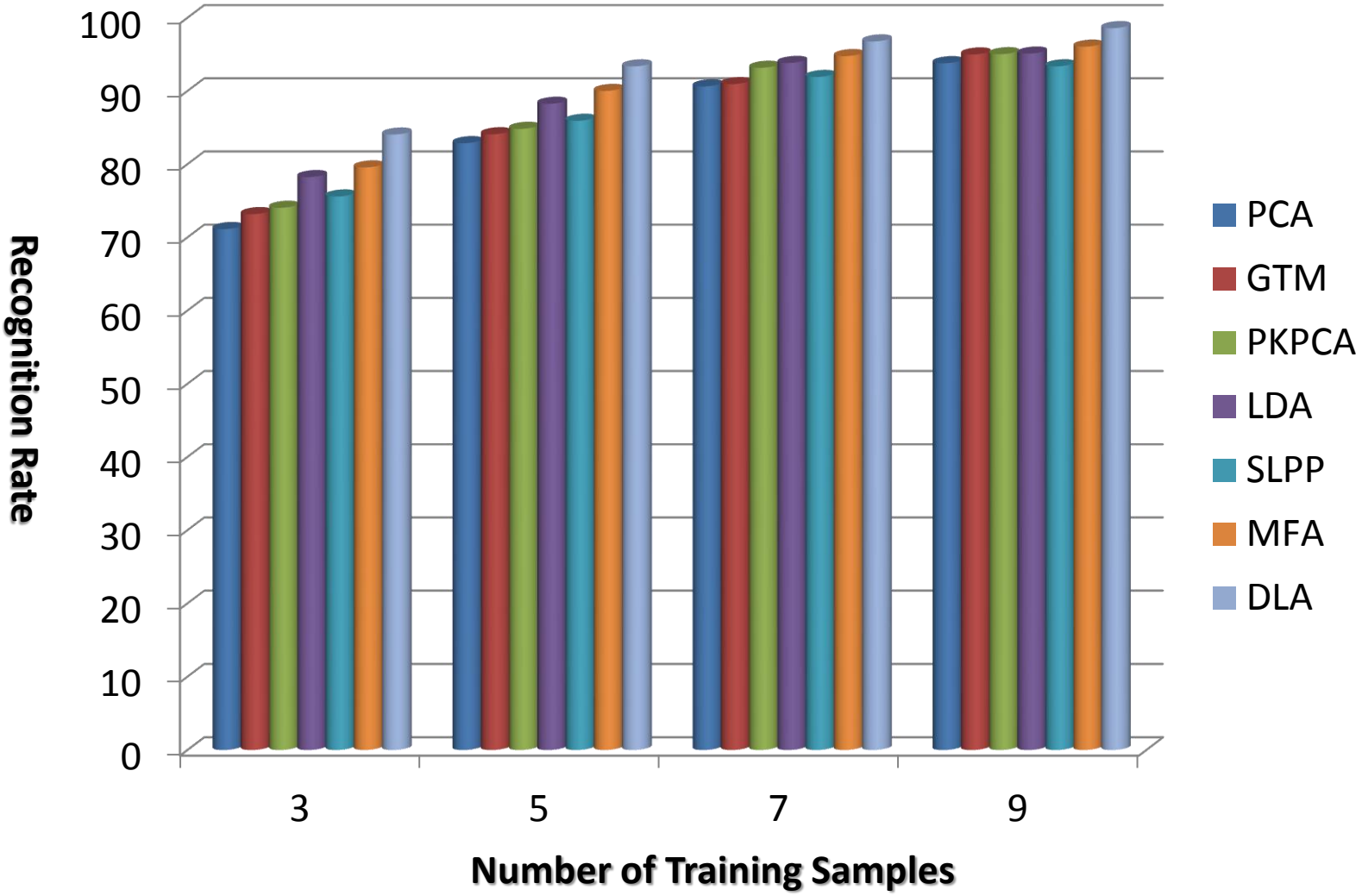


Recognition rate versus dimension

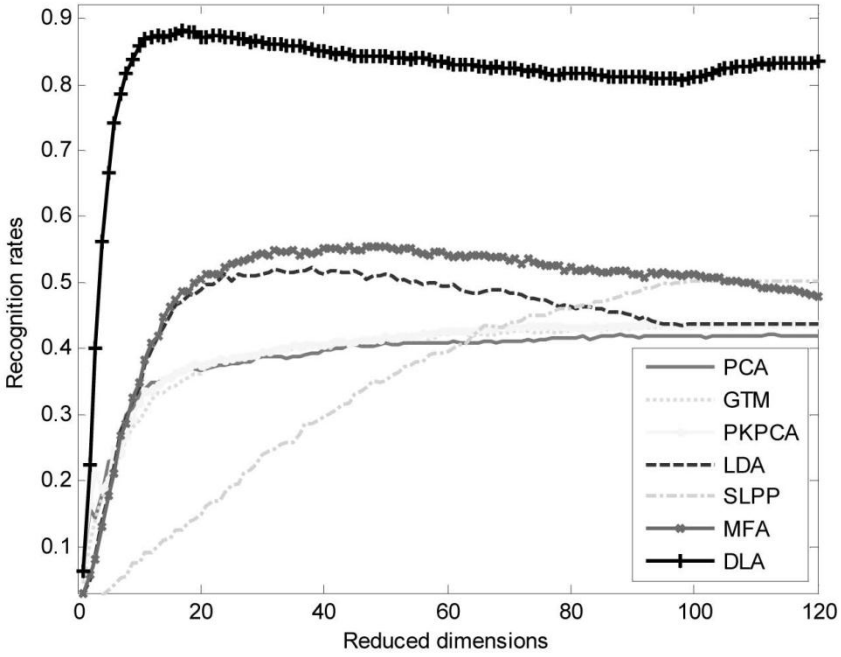


Recognition rate vs. subspace dimension on UMIST dataset

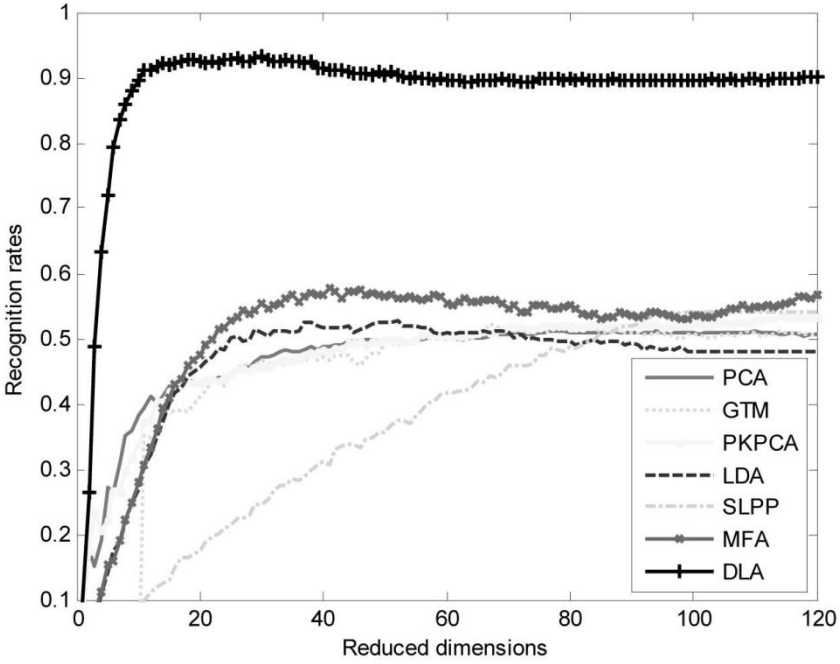
Best Recognition Rates on UMIST



Recognition rate versus dimension



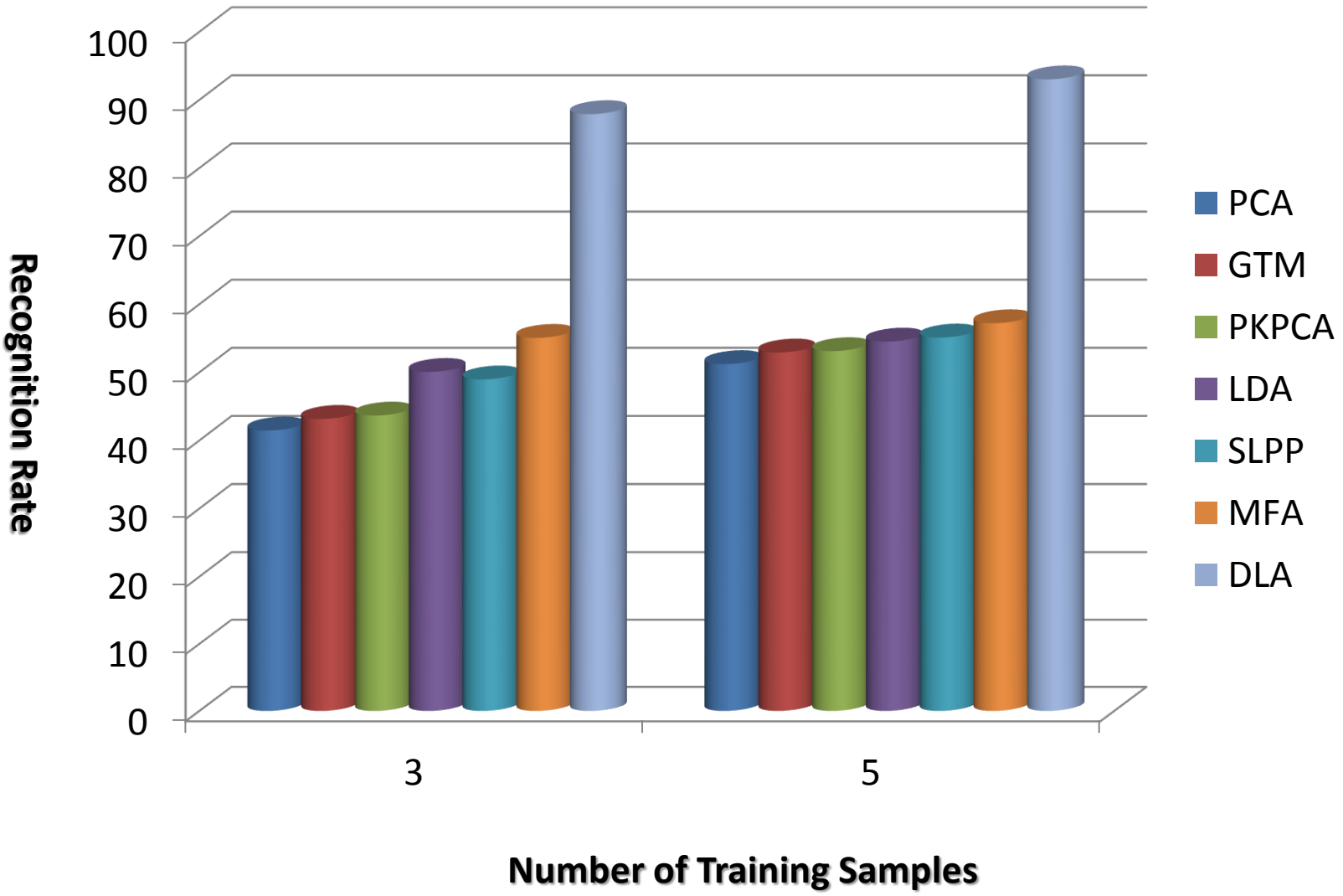
(a)



(b)

Recognition rate vs. subspace dimension on FERET dataset

Best Recognition Rates on FERET



Manifold Learning Key Problems

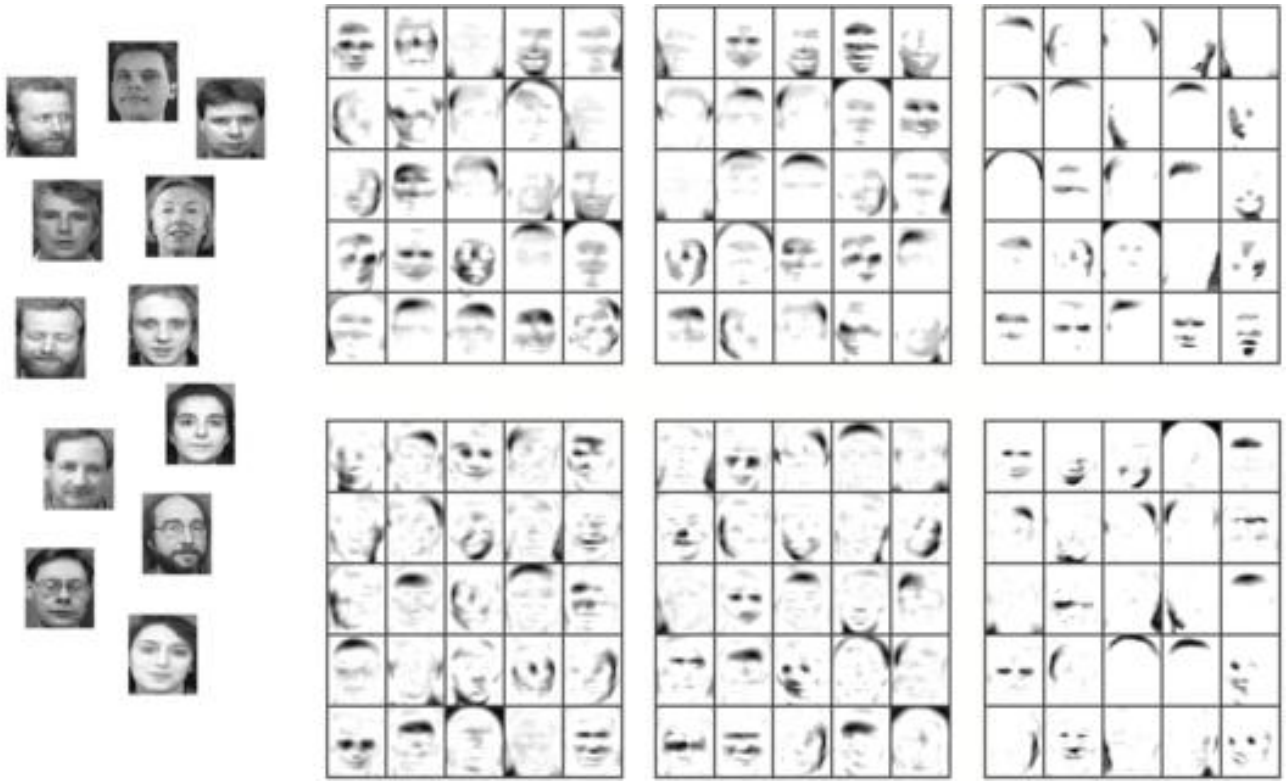
- Intrinsic dimensionality
- Noise
- Incremental data
- Random projection (fast computation)
- Parallel computation
- Approximation (Hashing)
- Specific problems (extensions of PAF)

Extensions of Patch Alignment Framework

- Nonnegative Patch Alignment Framework
- Sparse Patch Alignment Framework
- Transfer Patch Alignment Framework
- Multiview Patch Alignment Framework
- Active Patch Alignment Framework

PART FOUR

UTS:QCIS
QUANTUM COMPUTATION & INTELLIGENT SYSTEMS



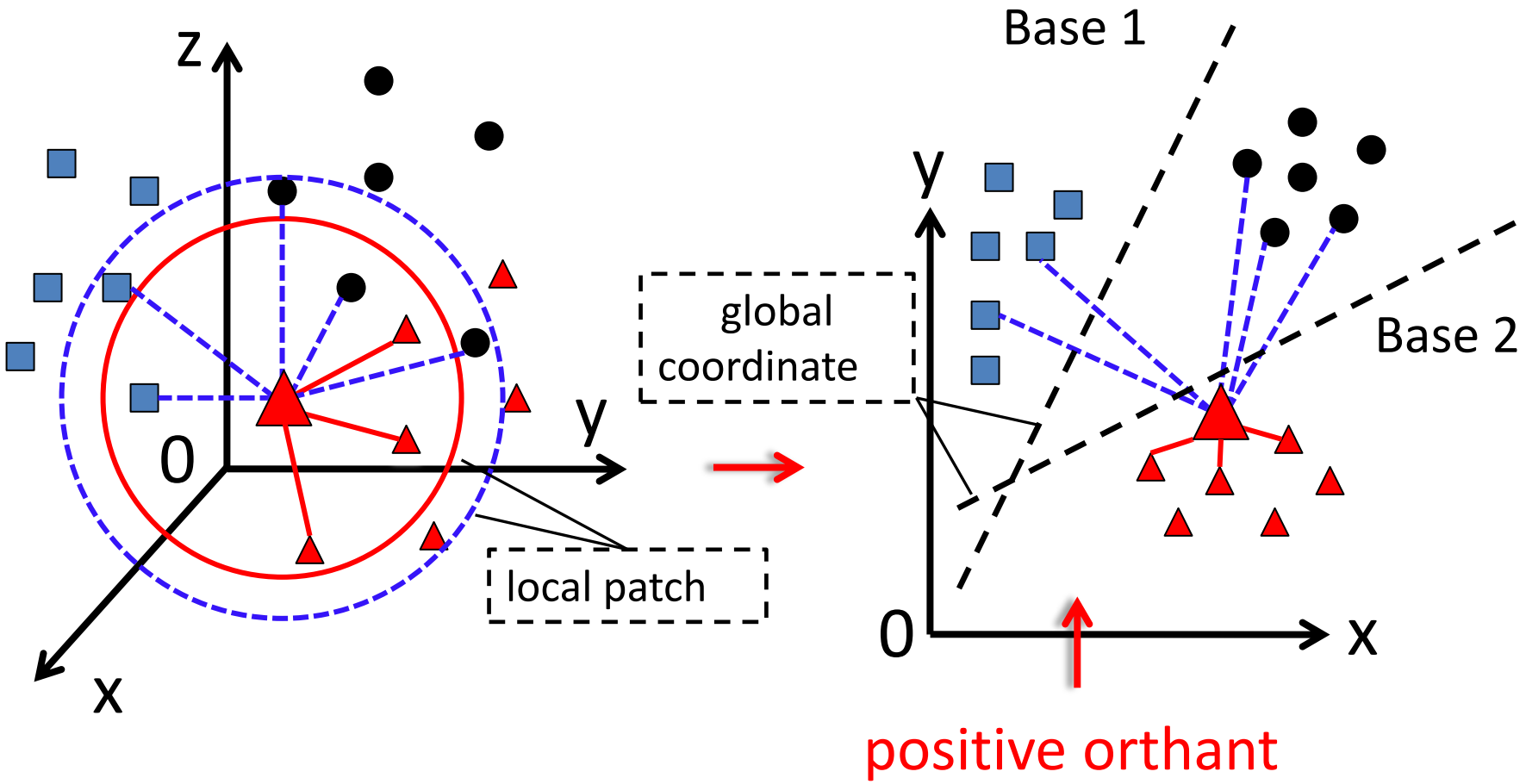
Nonnegative Patch Alignment Framework

PART FOUR – 1

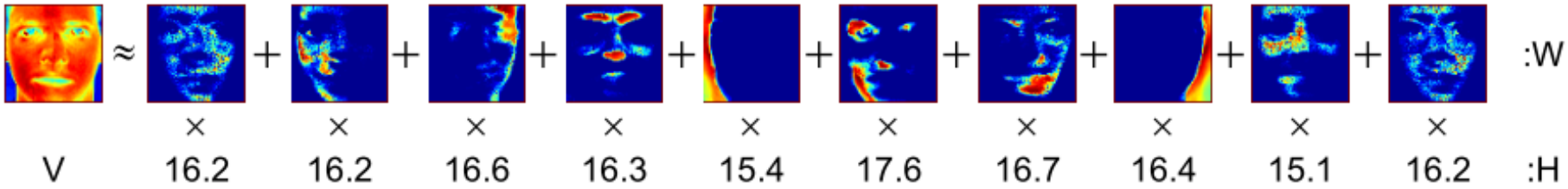
UTS:QCIS
QUANTUM COMPUTATION & INTELLIGENT SYSTEMS

Non-negative Patch Alignment

In practice, data are usually non-negative.



Non-negative Patch Alignment



non-negative part optimization

$$\arg\min_{H_i \geq 0} \operatorname{tr}\left(H_i L_i H_i^T\right)$$

non-negativity constraint

whole alignment

$$\arg\min_{H \geq 0} \sum_{i=1}^{N} \operatorname{tr}\left(H S_i L_i S_i^T H^T\right)$$

$$= \arg\min_{H \geq 0} \operatorname{tr}\left(H\left(\sum_{i=1}^{N} S_i L_i S_i^T\right) H^T\right)$$

$$= \arg\min_{H \geq 0} \operatorname{tr}\left(HLH^T\right)$$

Non-negative Patch Alignment

$$\arg\min_{W\geq 0, H\geq 0} F(W,H) = D(X, WH) + \frac{\gamma}{2}\mathrm{tr}\left(HLH^T\right)$$

Kullback-Leibler (KL) Divergence

jointly non-convex

$$D(X,Z) = \sum_{i,j}\left(X_{ij}\log\frac{X_{ij}}{Z_{ij}} - X_{ij} + Z_{ij}\right)$$

Frobenius Norm

$$D(X,Z) = \frac{1}{2}\left\|X - Z\right\|_F^2$$

Non-negative Patch Alignment

Fix W: $\arg\min_{H \geq 0} F(W,H) = KL(X, WH) + \frac{\gamma}{2}\mathrm{tr}\left(HLH^T\right)$

Gradient:

$$\frac{\partial F(W,H)}{\partial H_{ij}} = \gamma(HL)_{ij} - \sum_l \frac{X_{lj}W_{li}}{\sum_k W_{lk}H_{kj}} + \sum_k W_{ik}$$

Update Rule:

$$H_{ij} \leftarrow H_{ij} - \alpha_{ij}\frac{\partial F(W,H)}{\partial H_{ij}}$$

cannot guarantee non-negativity constraint.

step size

Non-negative Patch Alignment

Suppose: $L = L^+ - L^-, L^+ \geq 0, L^- \geq 0$, e.g.,

$$L^+ = (|L| + L)/2, L^- = (|L| - L)/2$$

Splitting Gradient into Positive and Negative Parts:

$$\frac{\partial F(W,H)}{\partial H_{ij}} = \nabla^+_{H_{ij}} F(W,H) - \nabla^-_H F(W,H)$$

$$\nabla^+_{H_{ij}} F(W,H) = \gamma (HL^+)_{ij} + \sum_k W_{ik} \geq 0$$

$$\nabla^-_{H_{ij}} F(W,H) = \gamma (HL^-)_{ij} + \sum_l \frac{X_{lj} W_{li}}{\sum_k W_{lk} H_{kj}} \geq 0$$

Non-negative Patch Alignment

Gradient Rescaling: $\alpha_{ij} = \theta_{ij} \frac{H_{ij}}{\nabla^{+}_{H_{ij}} F(W,H)}$

New Update Rule:

$$H_{ij} \leftarrow H_{ij} - \theta_{ij} \frac{H_{ij}}{\nabla^{+}_{H_{ij}} F(W,H)} (\nabla^{+}_{H_{ij}} F(W,H) - \nabla^{-}_{H_{ij}} F(W,H))$$

Let $\theta_{ij} = 1$, Multiplicative Update Rule (MUR):

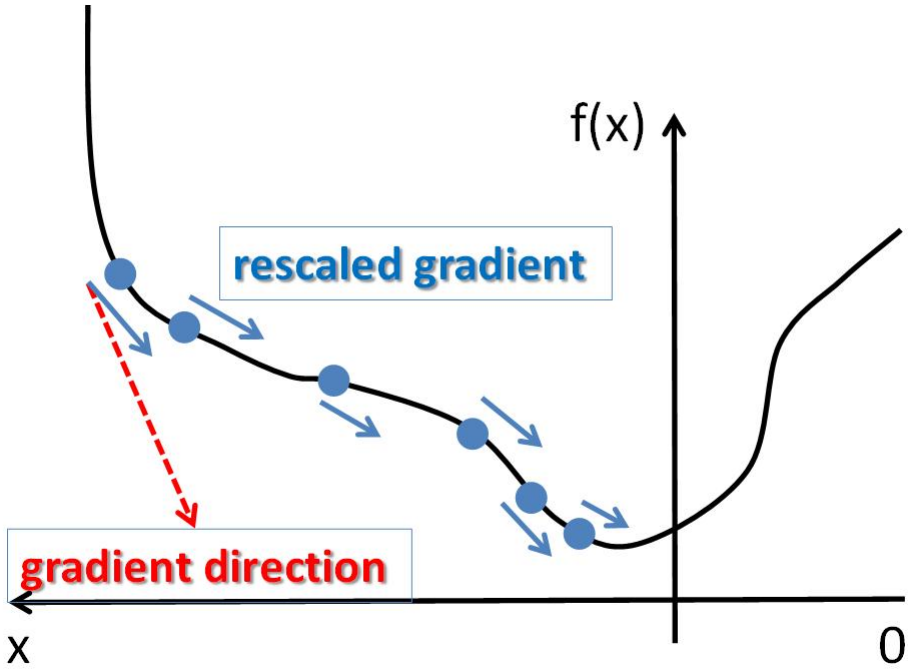
$$H_{ij} \leftarrow H_{ij} \frac{\nabla^{-}_{H_{ij}} F(W,H)}{\nabla^{+}_{H_{ij}} F(W,H)}$$

non-negativity constraint is guaranteed.

Non-negative Patch Alignment

Fix H: W can be updated similarly.

$$H \leftarrow H \circ \frac{\gamma HL^{-} + W^{T}\frac{X}{WH}}{\gamma HL^{+} + W^{T}1_{m\times n}} \qquad W \leftarrow W \circ \frac{\frac{X}{WH}H^{T}}{1_{m\times n}H^{T}}$$



MUR decreases the objective function (provable).

MUR converges slowly because $\theta_{ij} = 1$ is not optimal.

Non-negative Patch Alignment

THREE Fast Optimization Methods:

- Fast gradient descent (**FGD**)
- Optimal gradient method (**NeNMF**)
- Online RSA method (**OR-NMF**)

N. Guan, D. Tao, Z. Luo, and B. Yuan., "Nonnegative Patch Alignment Framework," ***IEEE Transactions on Neural Networks***, vol. 22, no. 8, pp. 1218-1230, 2011.
N. Guan, D. Tao, Z. Luo, and B. Yuan., "NeNMF: An Optimal Gradient Method for Non-negative Matrix Factorization," ***IEEE Transactions on Signal Processing***, vol. 60, no. 6, pp. 2882-2898, 2012.
N. Guan, D. Tao, Z. Luo, and B. Yuan., "Online Non-negative Matrix Factorization with Robust Stochastic Approximation," ***IEEE Transactions on Neural Networks and Learning Systems***, vol. 23, no. 7, pp. 1087-1099, 2012.

Non-negative Patch Alignment-FGD

Let $\theta_{ij} = \theta$, Rescaled Gradient Descent:

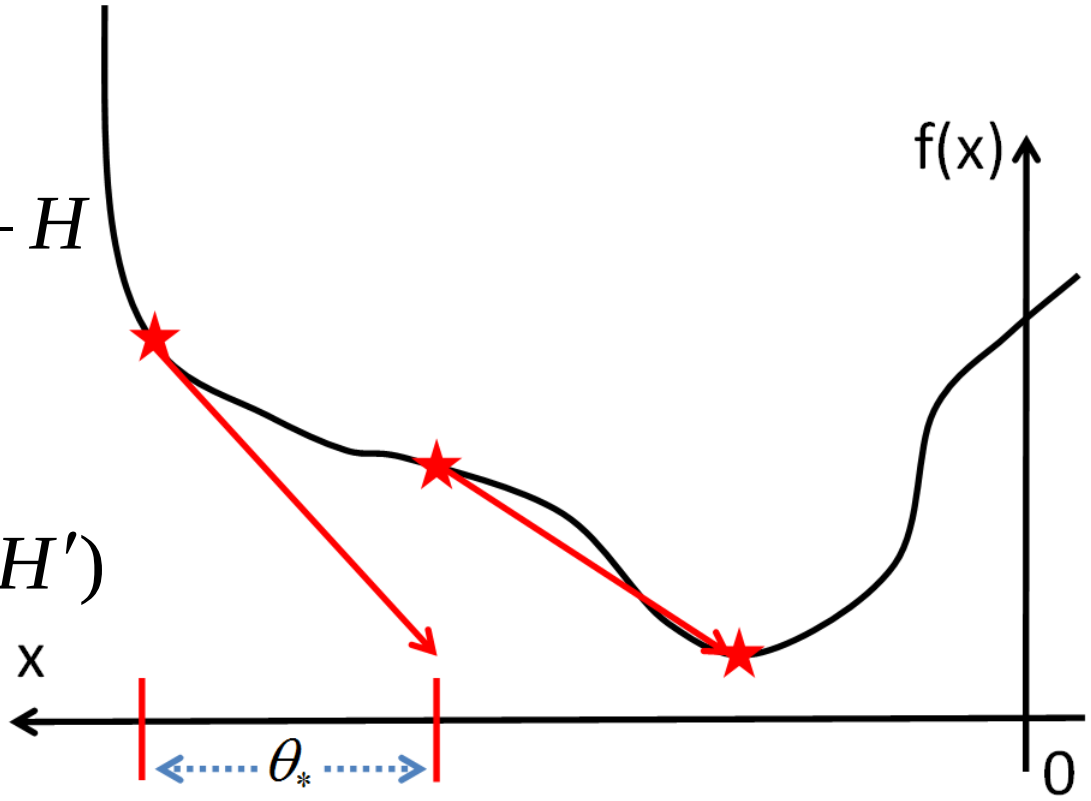
$$H' = H - \theta \frac{H}{\nabla_H^+ F(W,H)} \circ (\nabla_H^+ F(W,H) - \nabla_H^- F(W,H))$$

$$= H + \theta\Delta$$

$$\Delta = H \circ \frac{\nabla_H^- F(W,H)}{\nabla_H^+ F(W,H)} - H$$

Line Search:

$$\arg\min_{\theta, H' \geq 0} \varphi(\theta) = F(W,H')$$



Non-negative Patch Alignment-FGD

Solving Line Search Problem:

$$\varphi'(\theta) = \gamma tr(\Delta L \Delta^T)\theta + \gamma tr(HL\Delta^T) - \sum_{ij} \frac{X_{ij}(W\Delta)_{ij}}{(WH)_{ij} + (W\Delta)_{ij}\theta} + \sum_{ij}(W\Delta)_{ij}$$

$$\varphi''(\theta) = \gamma tr(\Delta L \Delta^T) + \sum_{ij} \frac{X_{ij}(W\Delta)_{ij}^2}{((WH)_{ij} + (W\Delta)_{ij}\theta)^2} \geq 0$$

Newton Update Rule:

$$\theta_{k+1} = \theta_k - \frac{\varphi'(\theta_k)}{\varphi''(\theta_k)} \Leftarrow \theta_0 = 1$$

$$\Rightarrow \theta_{k+1}$$

Final Step Size:

boundary

$$\theta_* = \min(\theta_{k+1}, \theta')$$

$$\theta' = \max\{H_{ij} / |\Delta_{ij}| \mid \Delta_{ij} < 0\}$$

Non-negative Patch Alignment-FGD

Algorithm 1: Fast Gradient Descent **(FGD)**

Input: X, W^t, H^t, ξ

Output: H^{t+1}

1: Initialize $\theta_0 = 1, k = 0$

2: Calculate $\Delta = H \circ \frac{\nabla_H^- F(W,H)}{\nabla_H^+ F(W,H)} - H$

3: Calculate $\theta' = \max\{H_{ij} / |\Delta_{ij}| \mid \Delta_{ij} < 0\}$

Repeat

4: Update $\theta_{k+1} = \theta_k - \frac{\varphi'(\theta_k)}{\varphi''(\theta_k)}$

5: $k \leftarrow k+1$

Until $|\theta_{k+1} - \theta_k| \le \xi$

6: Set $\theta_* = \min(\theta_{k+1}, \theta')$

7: $H^{t+1} = H^t + \theta_* \Delta$

1) FGD decreases the objective function (provable).

2) FGD accelerates MUR without increasing time cost.

3) FGD reduces to MUR ($\theta_* = 1$) in some cases.

Non-negative Patch Alignment-MFGD

Multiple step sizes: set a step size for each column of H.

$$\vec{\theta} = [\theta_1, \ldots, \theta_n]^T$$

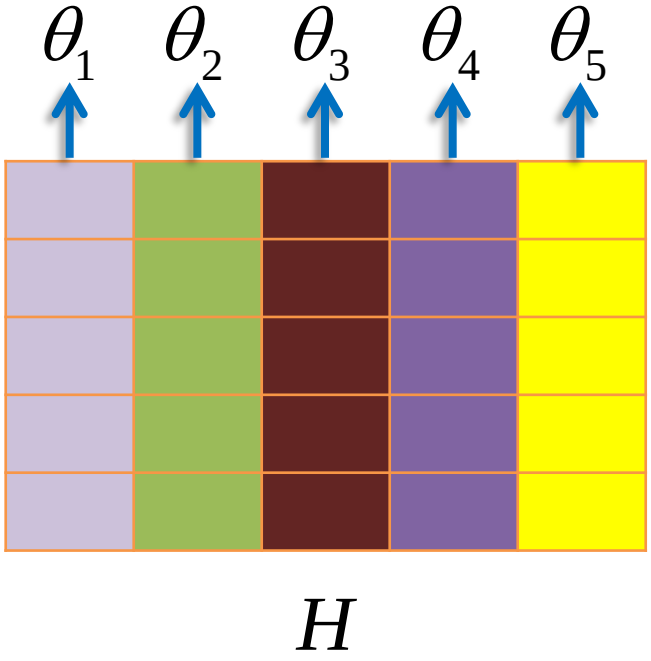
Line Search:

$$\arg\min_{\vec{\theta}, H' \geq 0} \phi(\vec{\theta}) = F(W, H') \qquad H' = H + \Delta \times diag(\vec{\theta})$$

MFGD is convex: **diagonal matrix**

$$Hessian(\phi) = A + \gamma L \circ (\Delta^T \Delta) \succ 0$$

$$A_{jj} = \sum_l \frac{X_{lj} (W\Delta)_{lj}^2}{((WH)_{lj} + \theta_j (W\Delta)_{lj})^2}$$



Non-negative Patch Alignment-MFGD

$$\nabla\phi(\vec{\theta}) = \frac{\partial\phi(\vec{\theta})}{\partial\theta_j} = \sum_l (W\Delta)_{lj} - \sum_l \frac{X_{lj}(W\Delta)_{lj}}{(WH)_{lj} + \theta_j (W\Delta)_{lj}} + \gamma(LH^T\Delta)_{jj} + \gamma\sum_i \theta_j L_{ij}(\Delta^T\Delta)_{ij}$$

Newton Update Rule:

$$\vec{\theta}^0 = \vec{1} \Rightarrow \vec{\theta}^{k+1} = \vec{\theta}^k - Hessian^{-1}(\phi)\nabla\phi(\vec{\theta}^k) \Rightarrow \vec{\theta}^{k+1}$$

Non-negative Patch Alignment-MFGD

The time complexity of Hessian inverse is: $O(n^3)$.

Sherman-Morrison-Woodbury Formula:

$$\begin{aligned} Hessian(\phi)^{-1} &= (A+\gamma L\circ(\Delta^T\Delta))^{-1} = (A+U\sum U^T)^{-1} \\ &\approx (A+U'\sum' U'^T)^{-1} \\ &= A^{-1} - A^{-1}U'(\sum'^{-1} + U'^T A^{-1}U')^{-1}U'^T A^{-1} \end{aligned}$$

The SVD of $\gamma L\circ(\Delta^T\Delta)) = U\sum U^T$ can be calculated previously, thus it reduces complexity of Hessian inverse:

$$O(n^3)\rightarrow O(p^3),\ p<<n.$$

Non-negative Patch Alignment-MFGD

Multiple Step Sizes Fast Gradient Descent:

$$H \leftarrow H + \Delta \times diag(\vec{\theta}_*)$$

MFGD decreases the objective function (proved).

Time Complexity:

MFGD: $O(mnr + n^2 r + n^3)$ MUR: $O(mnr + n^2 r)$

The time costs of one iteration MUR and MFGD are comparable when $(m \leq n)$, but MFGD further reduces the objective function.

Non-negative Patch Alignment-MFGD

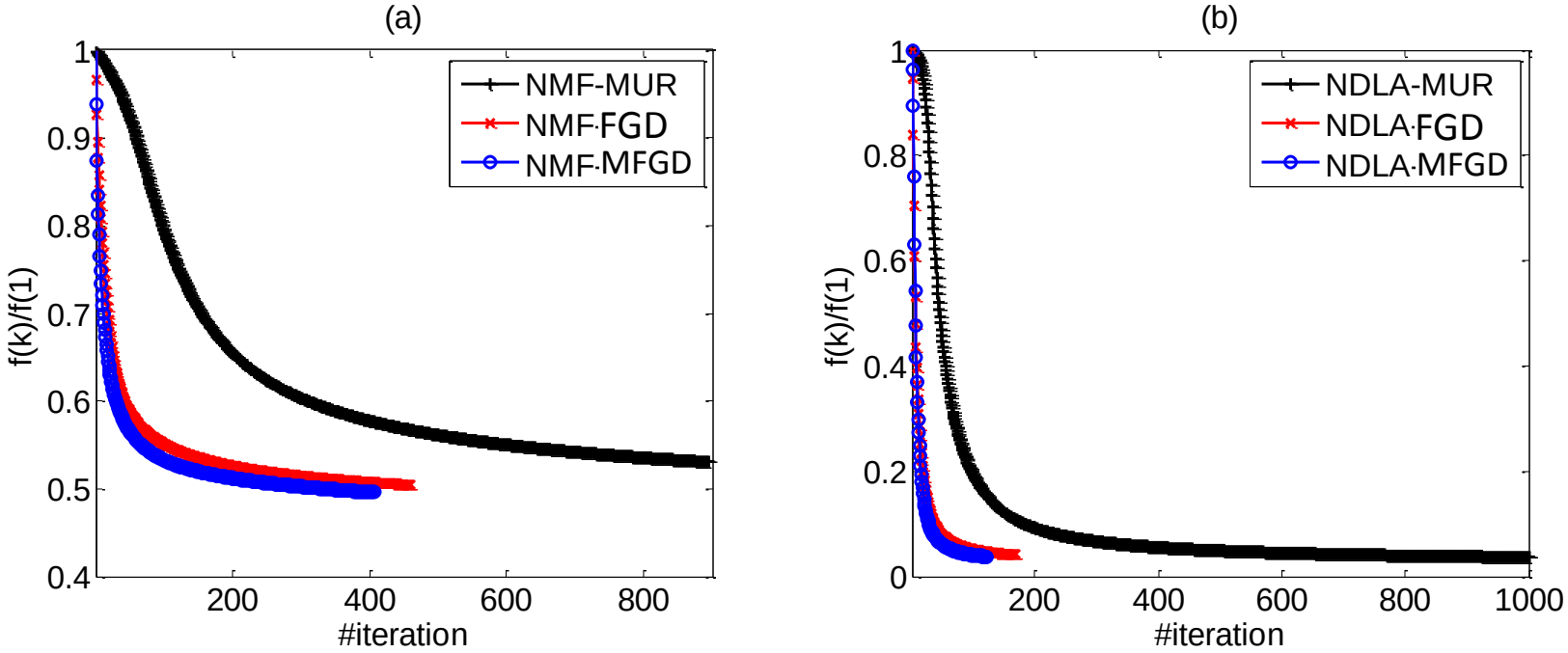


Fig. 2. Objective values versus iteration numbers for MUR, FGD, and MFGD when solving (a) NMF on a 2048x256-D matrix and (b) NDLA on a 1600x300-D matrix.

Both FGD and MFGD converge much faster than MUR.

Non-negative Patch Alignment-MFGD

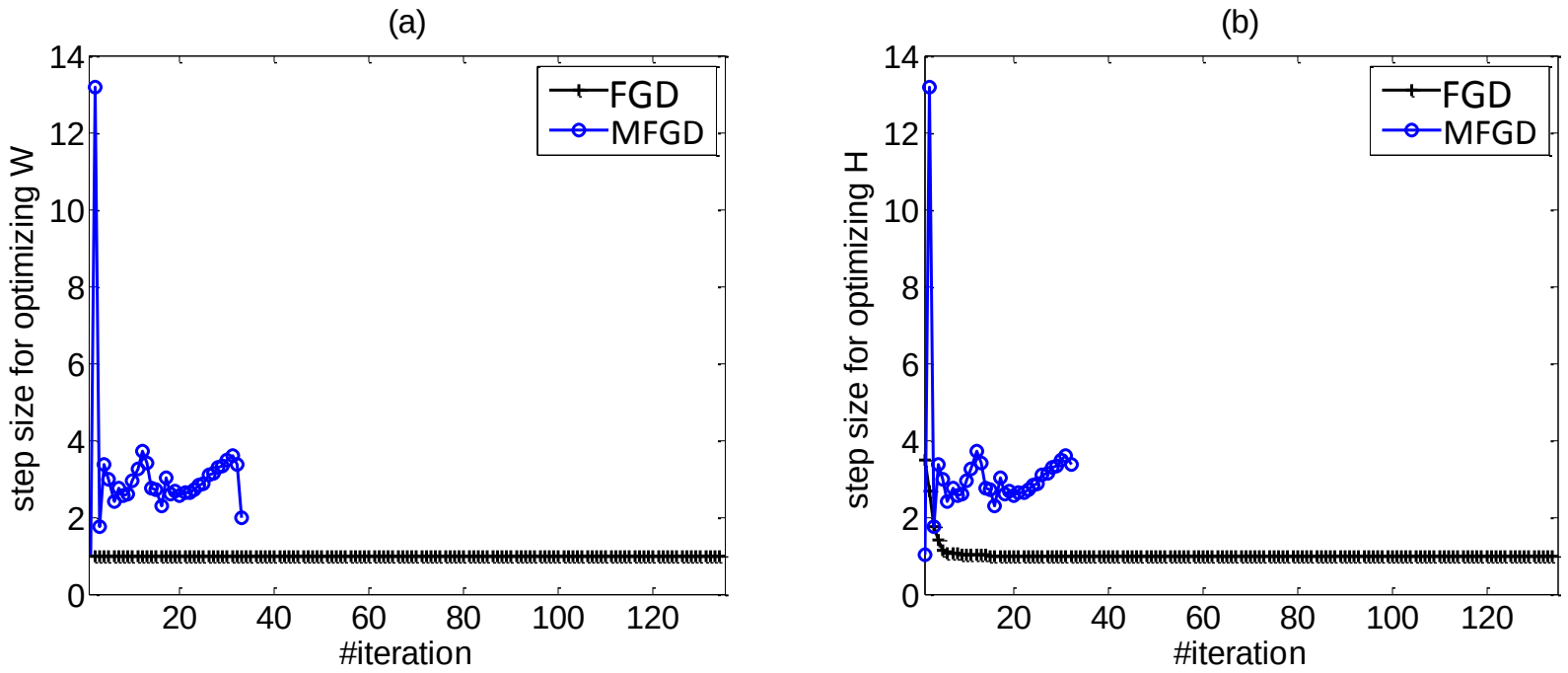


Fig. 3. Step size for optimizing (a) W and (b) H versus iteration number on the MNIST dataset with the dimensionality set to 300. For FGD, the average step sizes for rows of W and H columns of H are given.

MFGD consistently improves FGD.

Non-negative Patch Alignment

NMF: $$\min_{W\geq 0, H\geq 0} D(X, WH)$$

Derive from NPAF: setting $\gamma = 0$

MUR: $$H \leftarrow H \circ \frac{W^T \frac{X}{WH}}{W^T 1_{m\times n}} \quad W \leftarrow W \circ \frac{\frac{X}{WH} H^T}{1_{m\times n} H^T}$$

MFGD can be applied to optimize NMF.

$$\vec{\theta}^0 = \vec{1} \Rightarrow \vec{\theta}^{k+1} = \vec{\theta}^k - A^{-1}\nabla_\phi(\vec{\theta}^k) \Rightarrow \vec{\theta}_*$$

$$H \leftarrow H + \Delta \times diag(\vec{\theta}_*)$$

Non-negative Patch Alignment

LNMF: $$\min_{W\geq 0, H\geq 0} D(X, WH) + \alpha \sum_{ij} (W^T W)_{ij} - \beta \sum_{i} (HH^T)_{ii}$$

Three regularizations:

1) $\min \sum_{i\neq j} (W^T W)_{ij}$ makes bases approximately orthogonal
2) $\min \sum_{i} (W^T W)_{ii}$ suppresses over-decomposition of bases
3) $\max \sum_{i} (HH^T)_{ii}$ retains important components

Derive from NPAF:

$$\min_{W\geq 0, H\geq 0} D(X, WH) + \alpha tr(W 1_{r\times r} W^T) - \beta tr(H^T H)$$

Non-negative Patch Alignment

Understanding:

Fix W: $\arg\min_{H\geq 0} \beta tr(H^T(-I_{r\times r})H) + D(X, WH)$

$$\arg\max_{H\geq 0} tr(H^T H) = \arg\max_{H_i\geq 0} \sum_i tr(H_i^T H_i), H_i = H_{i:}^T$$

local patch H_i is built on each row of H, part optimization encourages retaining important components on each dimension.

Non-negative Patch Alignment

Fix H: $\arg\min_{W\geq 0} \alpha tr(W1_{r\times r}W^T) + D(X, WH)$

$$\arg\max_{W\geq 0} tr(W1_{r\times r}W^T) = \arg\max_{W_i\geq 0} \sum_i tr(W_i L_i W_i^T)$$

$$L_i = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} [1\cdots 1], W_i = [\vec{w}_i, \vec{w}_1, \vec{w}_2, ..., \vec{w}_r]$$

local patch W_i is built on each column of W, part optimization over each patch suppresses inner products between W_i and all the columns.

Non-negative Patch Alignment

Optimization under NPAF:

MUR:

$$H \leftarrow H \circ \frac{\beta H + W^T \frac{X}{WH}}{W^T 1_{m\times n}} \qquad W \leftarrow W \circ \frac{\frac{X}{WH} H^T}{\alpha W 1_{r\times r} + 1_{m\times n} H^T}$$

MUR and MFGD:

parameter β should be set sufficiently small to guarantee decreasing the objective function during updating H .

Non-negative Patch Alignment

GNMF: $\min_{W\geq 0, H\geq 0} \lambda tr(HLH^T) + D(X, WH)$

Understanding:

$$\arg\min_{H_i\geq 0} tr(H_i L_i H_i^T) = \arg\min_{H_i\geq 0} \sum_{j=1}^{k} S_{i,i^j} \left\| \vec{h}_i - \vec{h}_{i^j} \right\|^2$$

$$H_i = [\vec{h}_i, \vec{h}_{i^1}, ..., \vec{h}_{i^k}] \quad L_i = \begin{bmatrix} -\vec{1}_k^T \\ I_k \end{bmatrix} diag(\vec{1}_k) \begin{bmatrix} -\vec{1}_k^T, I_k \end{bmatrix}$$

S_{i,i^j} measures the similarity between sample i and i^j.

Non-negative Patch Alignment

local patch H_i is built on each sample and itself and its k nearest neighbors, part optimization minimizes distances between any two samples in H_i.

Optimization under NPAF:

MUR: $$H \leftarrow H \circ \frac{\lambda HL^- + W^T \frac{X}{WH}}{\lambda HL^+ + W^T 1_{m\times n}} \quad W \leftarrow W \circ \frac{\frac{X}{WH} H^T}{1_{m\times n} H^T}$$

Both MUR and MFGD can be applied to optimize GNMF.

Non-negative Patch Alignment

DNMF: $$\min_{W\geq 0, H\geq 0} D(X, WH) + \gamma S_W - \delta S_B$$

Fisher's criterion

Two types of non-negative patches:

within-class patch on each sample: itself and the rest in the same class.

$$H_i^W = [\vec{h}_i, \vec{h}_{i^1}, \ldots, \vec{h}_{i^{C_i}}]$$

centroid patch on each centroid: itself and the rest centroids of different classes.

$$H_c^B = [\bar{h}_c, \bar{h}_1, \ldots, \bar{h}_C]$$

Non-negative Patch Alignment

Part optimization:

For P_i^W: $$\arg\min_{H_i^W \geq 0} \sum_{j=1}^{C_i} \left\| \vec{h}_i - \vec{h}_i{}^j \right\|^2 = \arg\min_{H_i^B \geq 0} tr(H_i^W L_i^W (H_i^W)^T)$$

For P_c^B: $$\arg\max_{H_c^B \geq 0} \sum_{j \neq c} \left\| \bar{h}_c - \bar{h}_j \right\|^2 = \arg\max_{H_c^B \geq 0} tr(H_c^B L_c^B (H_c^B)^T)$$

Alignment:

$$\arg\min_{H_i^W \geq 0} \sum_{i=1}^{n} tr(H_i^W L_i^W (H_i^W)^T) = \arg\min_{H \geq 0} tr(HL^W H^T)$$

$$\arg\max_{H_c^B \geq 0} \sum_{c=1}^{C} tr(H_c^B L_c^B (H_c^B)^T) = \arg\min_{H \geq 0} tr(HL^B H^T)$$

Non-negative Patch Alignment

Unify under NPAF:

$$\min_{W\geq 0, H\geq 0} D(X, WH) + \gamma tr(H(L^W - \frac{\delta}{\gamma} L^B)H^T)$$

Optimization Methods:

This problem can be optimized by either MUR or MFGD.

The parameter δ should be set relatively small to guarantee the convexity.

Non-negative Patch Alignment

***Remark 1*:** NMF based dimension reduction algorithms can be unified under NPAF.

***Remark 2*:** differences of NMF based algorithms are fully characterized by the part optimization.

***Remark 3*:** all NMF related algorithms can be optimized by using MUR or MFGD.

***Remark 4*:** NPAF is a platform for developing new NMF based dimension reduction algorithms.

Non-negative DLA

Motivation: preserve local geometry and maximum-margin based discriminative information.

Two types of non-negative patches:

within-class patch on each sample: itself and the k_1 nearest neighbors in the same class.

$$H_i^W = [\vec{h}_i, \vec{h}_{i^1}, ..., \vec{h}_{i^{k_1}}] \geq 0$$

between-class patch on each sample: itself and the k_2 nearest neighbors in different classes.

$$H_i^B = [\vec{h}_i, \vec{h}_{i^1}, ..., \vec{h}_{i^{k_2}}] \geq 0$$

Non-negative DLA

Part optimization on both patches:

$$\arg\min_{H_i^W \geq 0} \sum_{j=1}^{k_1} S_{i,i^j}^W \left\| \vec{h}_i - \vec{h}_{i^j} \right\|^2 = \arg\min_{H_i^W \geq 0} tr(H_i^W L_i^W (H_i^W)^T)$$

$$\arg\max_{H_i^B \geq 0} \sum_{j=1}^{k_2} S_{i,i^j}^B \left\| \vec{h}_i - \vec{h}_{i^j} \right\|^2 = \arg\max_{H_i^B \geq 0} tr(H_i^B L_i^B (H_i^B)^T)$$

Alignment :

$$\arg\min_{H_i^W \geq 0} \sum_{i=1}^{n} tr(H_i^W L_i^W (H_i^W)^T) = \arg\min_{H \geq 0} tr(H L^W H^T)$$

$$\arg\min_{H_i^B \geq 0} \sum_{i=1}^{n} tr(H_i^B L_i^B (H_i^B)^T) = \arg\min_{H \geq 0} tr(H L^B H^T)$$

Non-negative DLA

Combined Objective function:

$$\arg\min_{H\geq 0} \frac{\gamma}{2} tr(H(L^B)^{-\frac{1}{2}} L^W (L^B)^{-\frac{1}{2}} H^T) + D(X, WH)$$

Alignment Matrix Splitting:

$$(L^B)^{-\frac{1}{2}} L^W (L^B)^{-\frac{1}{2}} = D - S$$

$$D = (L^B)^{-\frac{1}{2}} D^W (L^B)^{-\frac{1}{2}}, S = (L^B)^{-\frac{1}{2}} S^W (L^B)^{-\frac{1}{2}}$$

Both D and S are non-negative symmetric matrices (proved).

Non-negative DLA

Experiment Settings:

Dataset	#Image	#Sample	#Training	Occlusion Size	Dimensionality
ORL	112x92	400	8/person	20x20,25x25,30x30,35x35	10-120
UMIST	40x40	575	10/person	12x12,14x14,16x16,18x18	10-80
MINIST	28x28	3000	60/digit	6x6,8x8,10x10,12x12	10-120

Fig. 1. ORL (a) and UMIST (b) face image datasets and MNIST (c) handwritten dataset.

Performance Evaluation Criterion

Aim: evaluate whether the difference between the error rates of two classification algorithms is statistically significant.

"5x2 cv F-test": randomly divide total datasets into **two** folds for **five** times.

For each fold $(j = 1, 2)$ and each trail $(i = 1, \ldots, 5)$, compute

$$p_i^j = \left| Err_1(i, j) - Err_2(i, j) \right|$$

Define $\overline{p}_i = \frac{p_i^1 + p_i^2}{5}$ and $s_i^2 = (p_i^1 - \overline{p}_i)^2 + (p_i^2 - \overline{p}_i)^2$

Performance Evaluation Criterion

Then $F = \sum_{i=1}^{5}\sum_{j=1}^{2} \frac{(p_i^j)^2}{2\sum_{i=1}^{5} s_i^2}$ is an approximate F-distribution with 10 and 5 degrees of freedom.

Evaluation: Reject the hypothesis that the algorithms have statistically identical error rate with 95% confidence if the ***F*-statistics** is larger than 4.74.

Non-negative DLA

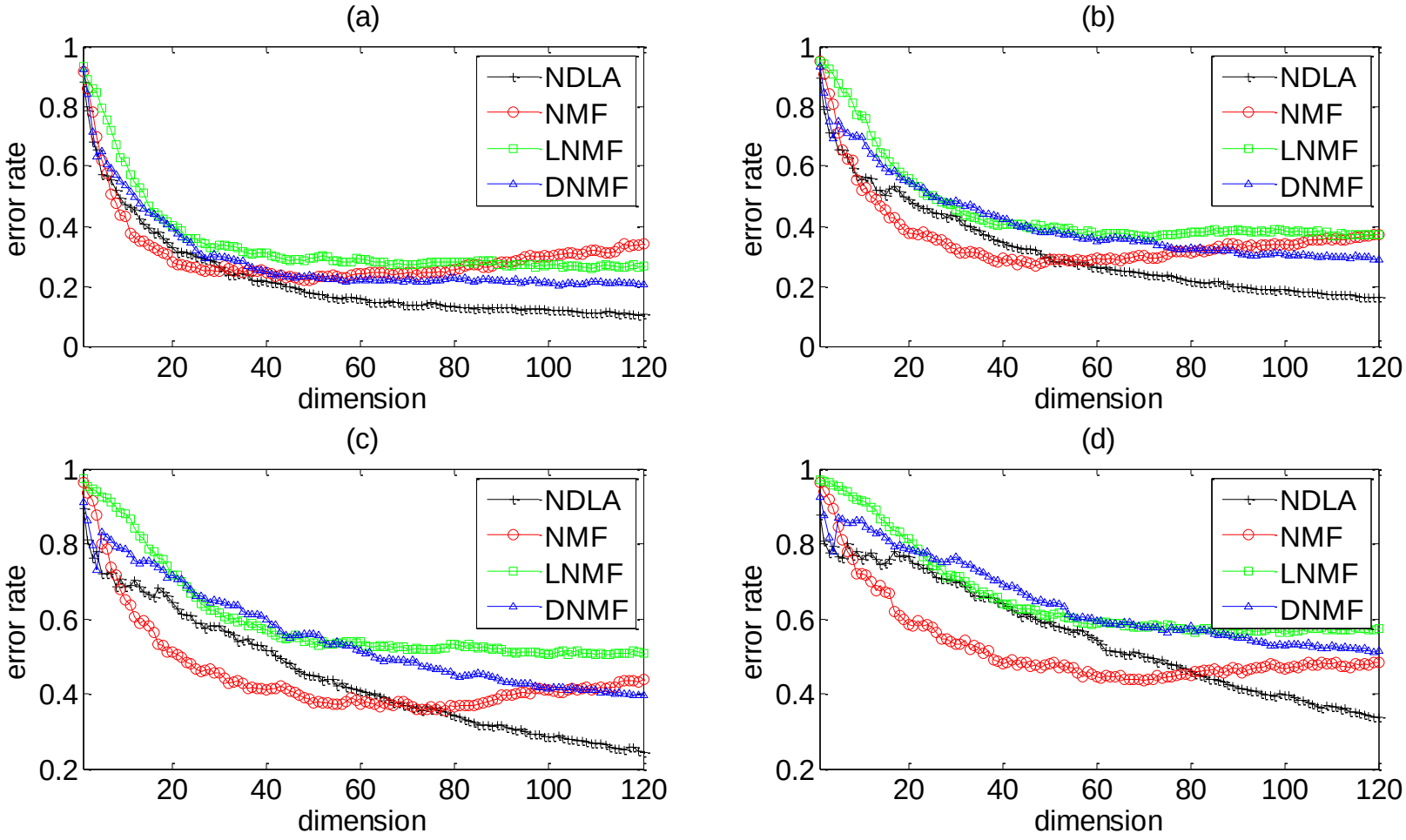


Fig. 5. Average error rate on *test set* when the partial occlusions size are 20x20, 25x25, 30x30, and 35x35 on the ORL dataset.

Non-negative DLA

Table III. F-test statistic value of NDLA versus other algorithms when the partial occlusions size are 20x20, 25x25, 30x30, and 35x35 on the ORL dataset.

Occlusion	PCA	FLDA	DLA	NMF	LNMF	DNMF
20 × 20	2.294	48.000 (√)	2.163	10.827 (√)	31.318 (√)	35.508 (√)
25 × 25	11.234 (√)	6.599 (√)	4.243	39.219 (√)	33.557 (√)	27.492 (√)
30 × 30	14.928 (√)	24.136 (√)	5.587 (√)	14.099 (√)	3.496	28.899 (√)
35 × 35	22.284 (√)	2.480	19.657 (√)	6.076 (√)	40.630 (√)	18.916 (√)

Table IV. Minimum average error rate on *test set* when the partial occlusions size are 20x20, 25x25, 30x30, and 35x35 on the ORL dataset.

Occlusion	PCA	FLDA	DLA	NMF	LNMF	DNMF	NDLA
20 × 20	12.8 (119)	16.6 (39)	11.7 (43)	22.1 (49)	25.3 (116)	20.0 (120)	**10.8 (120)**
25 × 25	20.4 (117)	24.8 (39)	18.1 (68)	28.3 (73)	37.1 (111)	28.2 (120)	**14.1 (119)**
30 × 30	33.7 (111)	33.4 (39)	31.3 (120)	36.1 (82)	26.3 (120)	38.8 (120)	**23.3 (120)**
35 × 35	45.8 (113)	39.9 (39)	45.6 (106)	42.9 (75)	57.3 (90)	50.2 (118)	**33.3 (120)**

Non-negative DLA

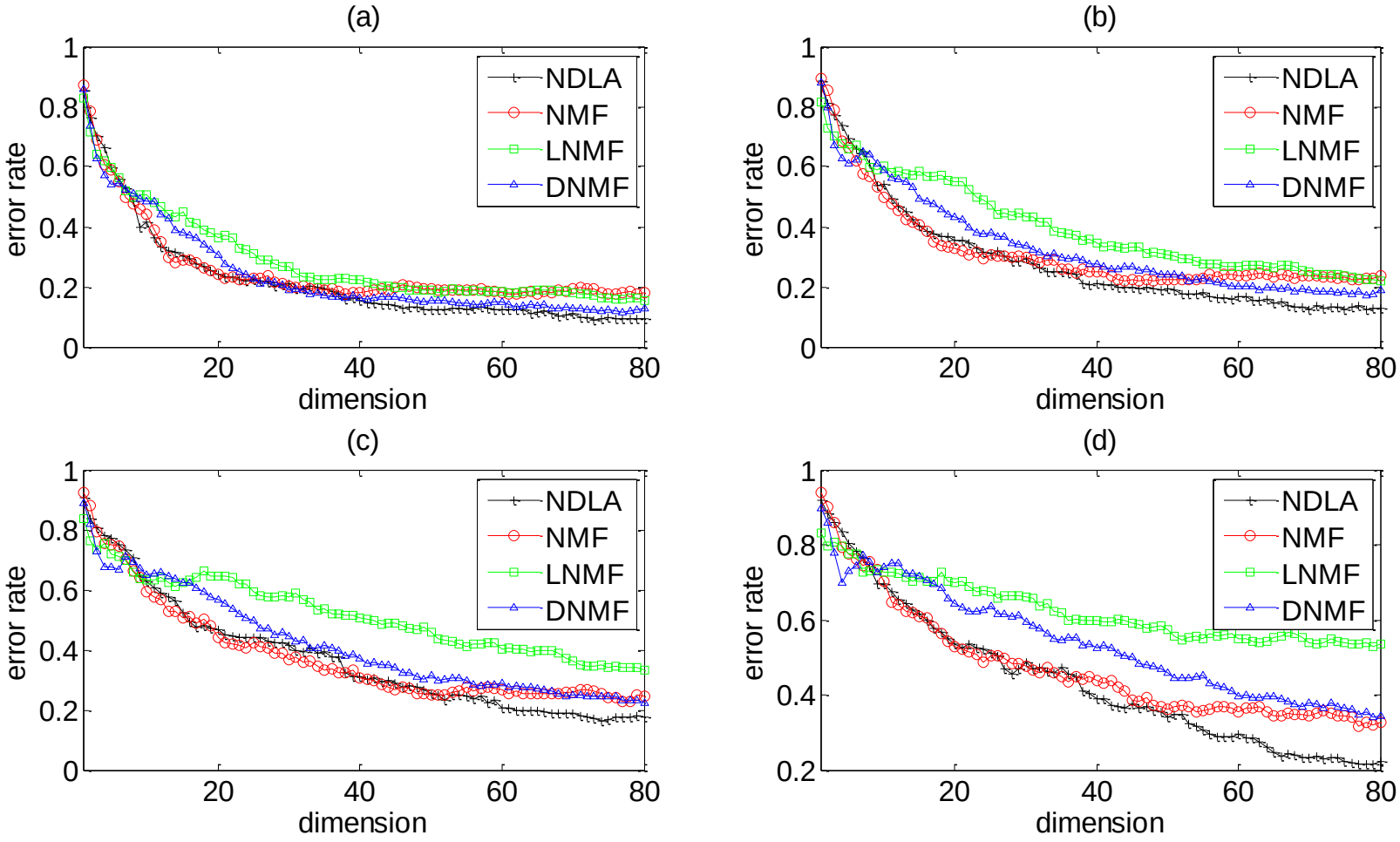


Fig. 6. Average error rate on *test set* when the partial occlusions size are 12x12, 14x14, 16x16, and 18x18 on the UMIST dataset.

Non-negative DLA

Table V. F-test statistic value of NDLA versus other algorithms when the partial occlusions size are 12x12, 14x14, 16x16, and 18x18 on the UMIST dataset.

Occlusion	PCA	FLDA	DLA	NMF	LNMF	DNMF
12×12	4.220	3.378	4.333	15.500 (√)	11.837 (√)	3.324
14×14	7.934 (√)	2.937	3.157	3.380	5.359 (√)	35.667 (√)
16×16	64.941 (√)	11.371 (√)	18.209 (√)	3.159	57.341 (√)	3.262
18×18	21.922 (√)	6.672 (√)	27.471 (√)	7.256 (√)	23.866 (√)	3.747

Table VI. Minimum average error rate on *test set* when the partial occlusions size are 12x12, 14x14, 16x16, and 18x18 on the UMIST dataset.

Occlusion	PCA	FLDA	DLA	NMF	LNMF	DNMF	NDLA
12×12	14.6 (59)	14.3 (19)	10.8 (74)	18.6 (28)	16.1 (80)	13.6 (78)	**09.6 (80)**
14×14	22.1 (63)	19.2 (19)	18.4 (72)	21.2 (49)	23.1 (80)	18.0 (79)	**13.1 (79)**
16×16	32.0 (78)	29.4 (19)	29.9 (70)	24.6 (77)	32.5 (80)	24.5 (77)	**17.6 (79)**
18×18	50.1 (71)	35.3 (19)	43.0 (73)	31.4 (79)	51.0 (80)	33.4 (80)	**22.5 (79)**

Non-negative DLA

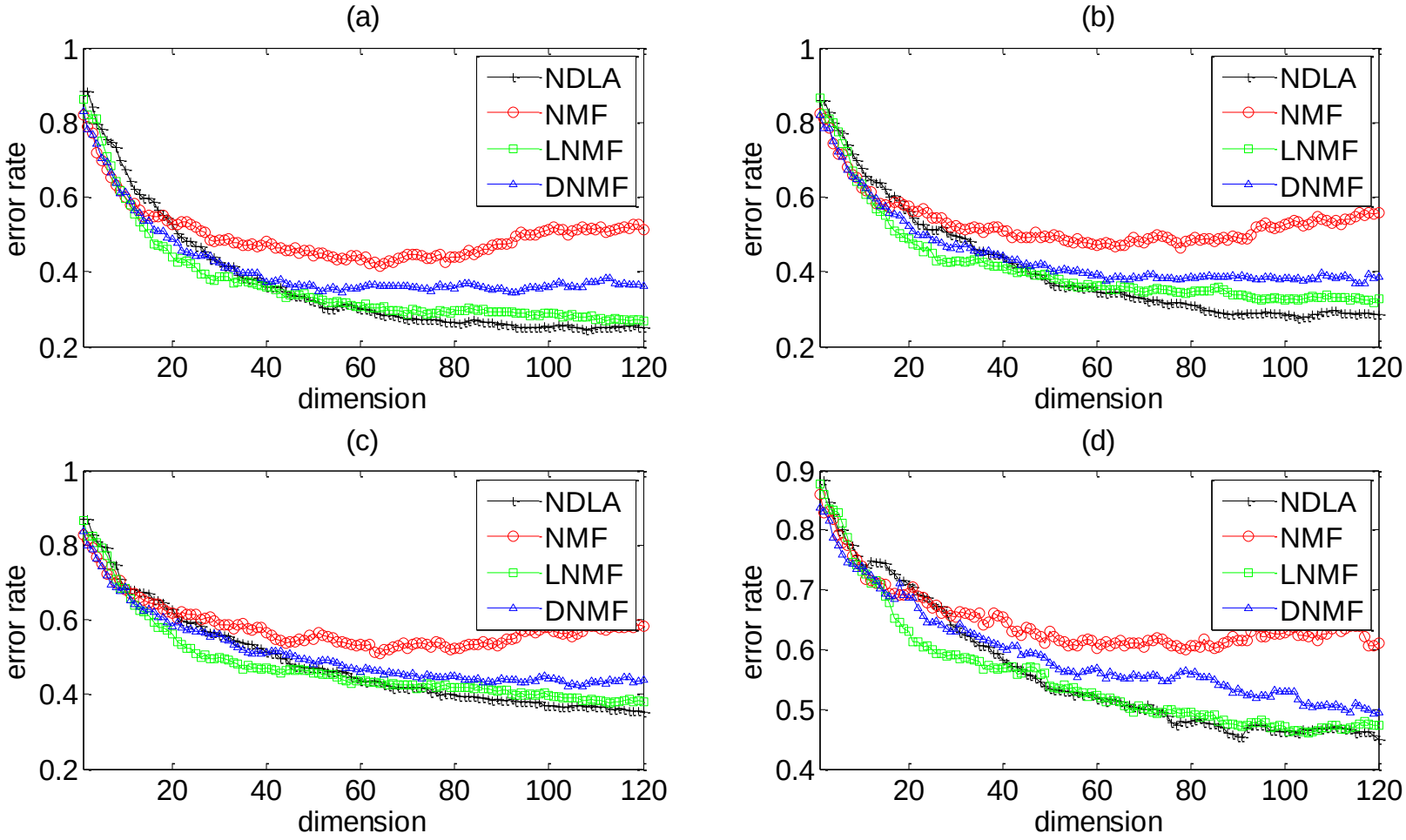


Fig. 7. Average error rate on *test set* when the partial occlusions size are 6x6, 8x8, 10x10, and 12x12 on the MNIST dataset.

Non-negative DLA

Table VII. F-test statistic value of NDLA versus other algorithms when the partial occlusions size are 20x20, 25x25, 30x30, and 35x35 on the MNIST dataset.

Occlusion	PCA	FLDA	DLA	NMF	LNMF	DNMF
6×6	3.253	92.959 (√)	8.498 (√)	61.269 (√)	1.922	1.521
8×8	2.718	93.080 (√)	4.527	85.937 (√)	2.053	1.601
10×10	1.998	27.497 (√)	3.546	26.737 (√)	3.611	4.795 (√)
12×12	1.526	51.420 (√)	2.362	15.375 (√)	3.833	39.361 (√)

Table VIII. Minimum average error rate on *test set* when the partial occlusions size are 20x20, 25x25, 30x30, and 35x35 on the MNIST dataset.

Occlusion	PCA	FLDA	DLA	NMF	LNMF	DNMF	NDLA
6×6	25.4 (50)	49.5 (9)	24.8 (50)	42.8 (55)	26.1 (112)	30.3 (92)	**24.1** (109)
8×8	30.6 (50)	52.9 (9)	29.5 (36)	47.5 (62)	29.3 (103)	34.0 (94)	**28.3** (119)
10×10	38.0 (61)	56.4 (9)	37.1 (63)	52.7 (64)	36.6 (120)	40.5 (104)	**35.0** (120)
12×12	48.3 (29)	64.2 (9)	47.7 (61)	60.0 (48)	45.3 (103)	49.4 (114)	**44.9** (120)

Optimization

NeNMF: Optimal Gradient Method

Table I. Existing NMF and NPAF Optimization Methods.

Distance $D(X, WH)$	KL divergence		Euclidean	
Algorithms\Models	NMF	NPAF	NMF	NPAF
Multiplicative Update Rule (MUR)	√	√	√	√
Fast Gradient Descent (FGD/MFGD)	√	√	√	√
Projected Non-negative Least Squares (PNLS)			√	
Projected Gradient (PG)			√	√
Projected Quasi-Newton (QN*)			√	
Broyden Fetcher Goldfarb Shanno (BFGS*)			√	
Projected Barzilai Borwein (PBB*)			√	
Cyclic Block Coordinate Gradient Projection (CBGP)			√	
Active Set (AS)			√	
Block Principal Pivoting (BPP)			√	
NeNMF*			√	√

* Converges at the rate of $O(1/k^2)$ for optimizing one factor matrix with another fixed.

NeNMF: Optimal Gradient Method

Euclidean distance based NMF: further study.

$$\arg\min_{W\geq 0, H\geq 0} F(W,H) = \frac{1}{2}\|X - WH\|_F^2$$

jointly non-convex

Alternating optimization:

$$H^{t+1} = \arg\min_{H\geq 0} F(W^t, H) = \frac{1}{2}\|X - W^t H\|_F^2$$

$$W^{t+1} = \arg\min_{W\geq 0} F(W, H^{t+1}) = \frac{1}{2}\|X - WH^{t+1}\|_F^2$$

Each sub-problem is non-negative least squares (**NNLS**).

NeNMF: Optimal Gradient Method

NNLS problem: revisit.

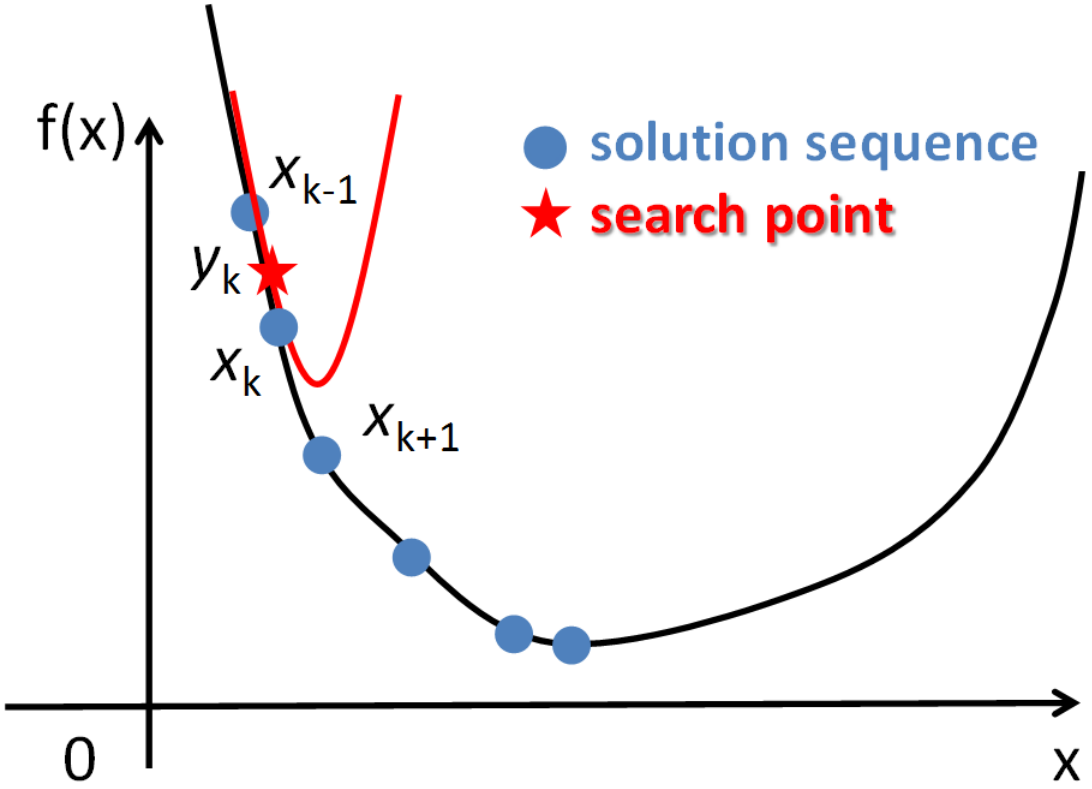
***Lemma 1*:** The objective function $F(W^t, H)$ is convex.

***Lemma 2*:** The gradient of the objective function $F(W^t, H)$ is Lipschitz continuous and the Lipschitz constant is $L = \left\| W^{tT} W^t \right\|_2$.

Lipschitz Continuity:

$$\forall H, H', \quad \frac{\left\| \nabla_H F(W^t, H) - \nabla_H F(W^t, H') \right\|_F}{\left\| H - H' \right\|_F} \leq L$$

NeNMF: Optimal Gradient Method



Step 1: Construct a search point:

$$y_k = \lambda_k x_{k-1} + (1-\lambda_k)x_k$$

$$0 < \lambda_k < 1$$

Step 2: Optimize the second-order approximation at the search point.

$$x_{k+1} = \arg\min_x f(y_k) + < \nabla(y_k), x - y_k > + \frac{L}{2}\|x - y_k\|^2$$

NeNMF: Optimal Gradient Method

Optimal Gradient Method (OGM) for NNLS:

$$H_k = \arg\min_{H \geq 0} \{\phi(Y_k, H) = F(W^t, Y_k) + < \nabla_H F(W^t, Y_k), H - Y_k > + \frac{L}{2} \|H - Y_k\|_F^2 \} \quad \text{(P)}$$

$$Y_{k+1} = H_k + \frac{\alpha_k - 1}{\alpha_{k+1}} (H_k - H_{k-1})$$

Smartly choosing coefficients:

$$\alpha_{k+1} = \frac{1 + \sqrt{4\alpha_k^2 + 1}}{2} \qquad \alpha_0 = 1$$

NeNMF: Optimal Gradient Method

Multiplier for $H \geq 0$

Lagrange Multipliers Method:

$$\arg\min_{H} \mathcal{L}(H) = \phi(Y_k, H) + \langle \lambda, H \rangle$$

Karush-Kuhn-Tucher (K.K.T.) Conditions of problem (P):

Stationarity ⇒ $\frac{\partial \mathcal{L}}{\partial H} = \nabla_H \phi(Y_k, H_k) + \lambda = 0$

Primal Feasibility ⇒ $H_k \geq 0$

Complementary slackness ⇒ $\lambda \circ H_k = 0$

NeNMF: Optimal Gradient Method

K.K.T. conditions

Closed form solution: $H_k = P(Y_k - \frac{1}{L}\nabla_H F(W^t, Y_k))$

Projection onto Non-negative Orthant:

$$P(X) = \begin{cases} X_{ij}, X_{ij} \geq 0 \\ 0, X_{ij} < 0 \end{cases}$$

NeNMF: Optimal Gradient Method

K.K.T. Based Stopping Criterion:

$$\left.\begin{array}{c} \nabla_H F(W^t, H_k) \geq 0 \\ H_k \geq 0 \\ \nabla_H F(W^t, H_k) \circ H_k = 0 \end{array}\right\} \Leftrightarrow \nabla_H^P F(W^t, H_k) = 0$$

projected gradient (Lin, 2007)

Practical Stopping Condition:

$$\left\| \nabla_H^P F(W^t, H_k) \right\|_F \leq \varepsilon_H$$

how close to the optima

$$\varepsilon_H = \max(10^{-3}, \varepsilon) \times \left\| \nabla_H^P F(W^1, H^1), \nabla_W^P F(W^1, H^1)^T \right\|_F$$

NeNMF: Optimal Gradient Method

Algorithm 1: Optimal gradient method **(OGM)**

Input: W^t, H^t

Output: H^{t+1}

1: Initialize $Y_0 = H^t, \alpha_0 = 1, L = \left\| W^{tT} W^t \right\|_2, k = 0$

Repeat

2: Update H_k, α_{k+1}, and Y_{k+1} with

2.1: $H_k = P(Y_k - \frac{1}{L} \nabla_H F(W^t, Y_k))$

2.2: $\alpha_{k+1} = \frac{1 + \sqrt{4\alpha_k^2 + 1}}{2}$

2.3: $Y_{k+1} = H_k + \frac{\alpha_k - 1}{\alpha_{k+1}} (H_k - H_{k-1})$

3: $k \leftarrow k+1$

Until K.K.T. based condition is satisfied

4: $H^{t+1} = H_K$

Algorithm 2: NeNMF

Input: $X \in R_+^{m \times n}, 1 \le r \le \min\{m, n\}$

Output: $W \in R_+^{m \times r}, H \in R_+^{r \times n}$

1: Initialize $W^1 \ge 0, H^1 \ge 0, t = 1$

Repeat

2: Update H^{t+1} and W^{t+1} with

2.1: $H^{t+1} = OGM(W^t, H^t)$

2.2: $(W^{t+1})^T = OGM((H^{t+1})^T, (W^t)^T)$

3: $t \leftarrow t+1$

Until K.K.T. based condition is satisfied

4: $W = W^t, H = H^t$

NeNMF: Optimal Gradient Method

Proposition 1*:** Given sequences $\{H_k\}_{k=0}^{\infty}$ and $\{Y_k\}_{k=0}^{\infty}$ generated by ***Algorithm 1, we have

$$F(W^t, H_k) - F(W^t, H_*) \leq \frac{2L\left\|H^t - H_*\right\|_F^2}{(k+1)^2}$$

where H_* is the optimal solution of NNLS.

***Remark 1*:** NeNMF converges to a stationary point of the NMF problem. (cf. Grippo & Sciandrone, 2000)

***Remark 2*:** NeNMF optimizes each factor matrix at the convergence rate of $O(\frac{1}{k^2})$ without any other overhead.

NeNMF: Optimal Gradient Method

Efficiency Evaluation:

Dataset	m	n	r
Synthetic 1	500	100	50
Synthetic 2	5,000	1,000	100
Reuters-21578	1,893	829	50
TDT-2	3,677	939	100

Effectiveness Evaluation:

Document Corpus	Document Number	Categories Number
Reuter-21578	8,292	30
TDT-2	9,394	30

NeNMF: Optimal Gradient Method

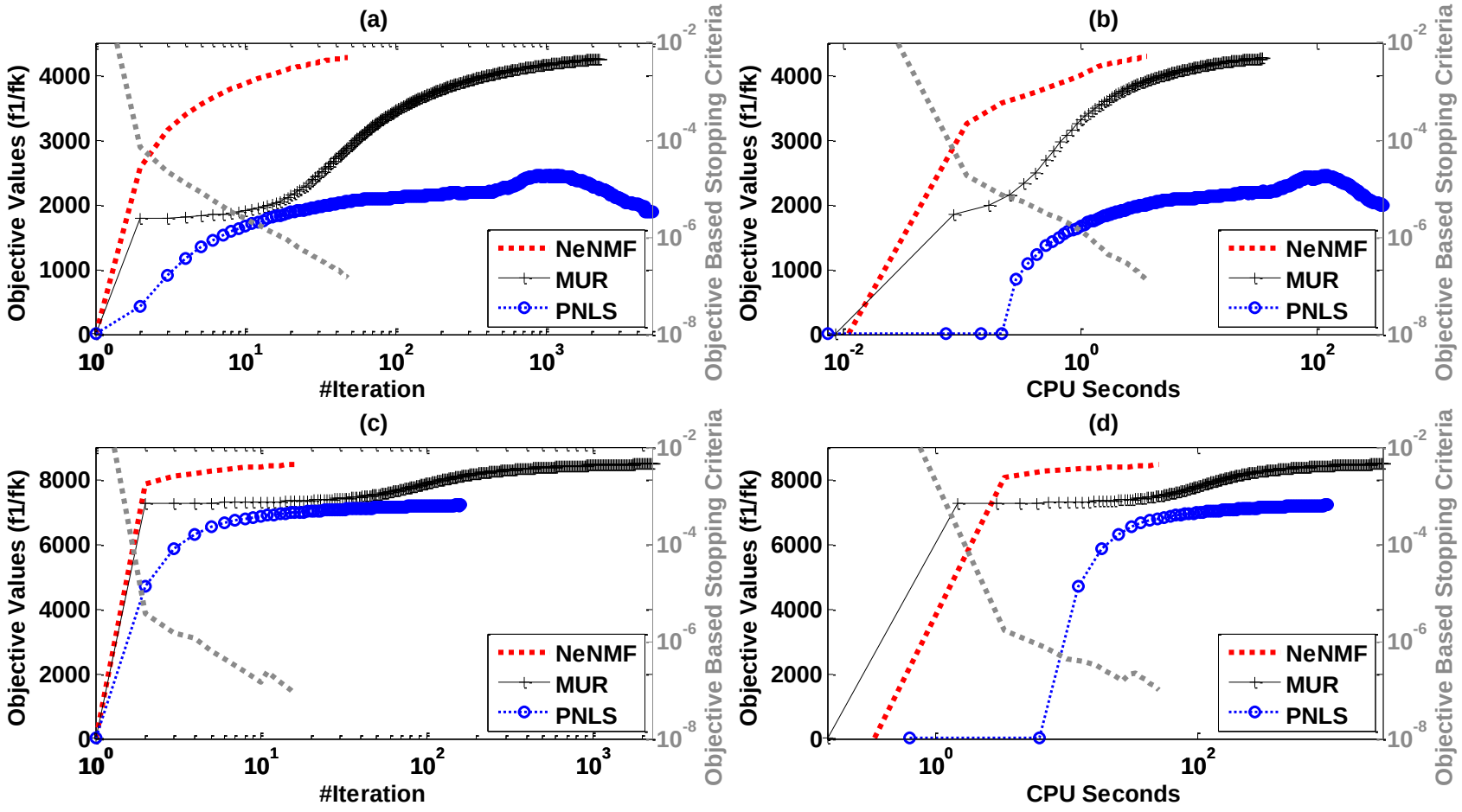


Fig. 1. Average objective values versus iteration numbers and CPU seconds of NeNMF, MUR, PNLS on the Synthetic 1 (a and) and Synthetic 2 (c and d) datasets.

NeNMF: Optimal Gradient Method

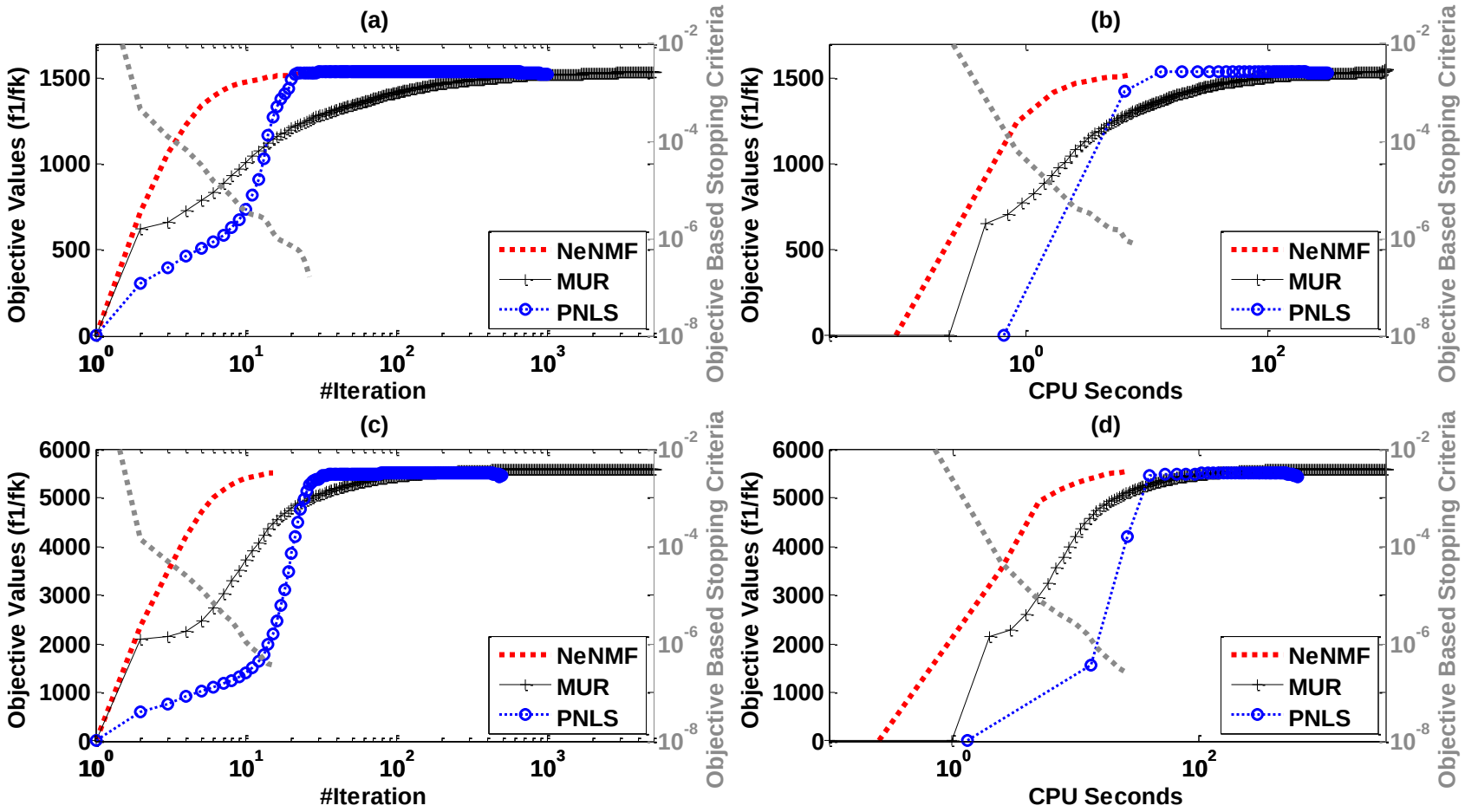


Fig. 2. Average objective values versus iteration numbers and CPU seconds of NeNMF, MUR, PNLS on the Reuters-21578 (a and) and TDT-2 (c and d) datasets.

NeNMF: Optimal Gradient Method

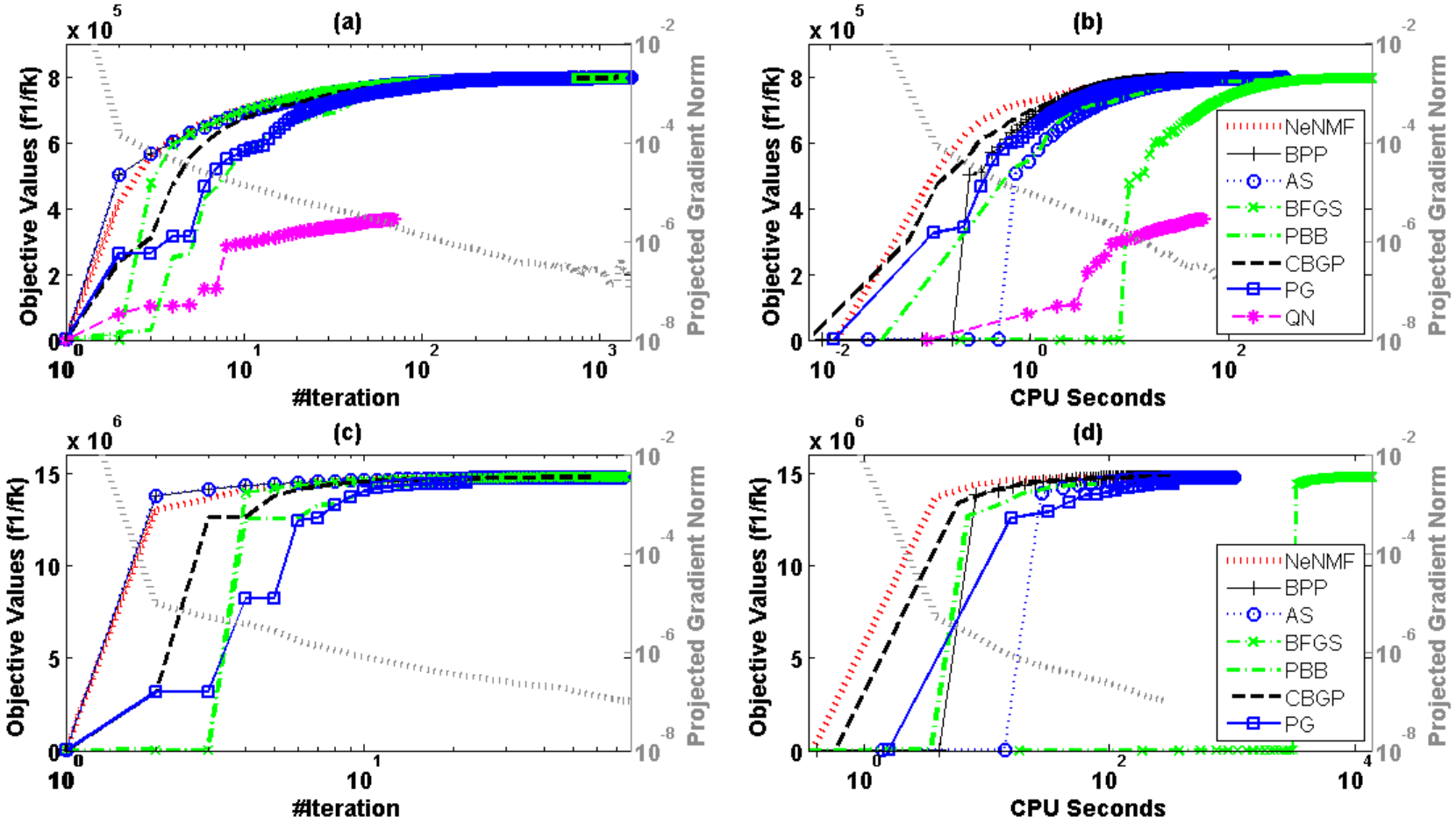


Fig. 4. Average objective values versus iteration numbers and CPU seconds of NeNMF, PG-, AS-based NMF solvers on the Synthetic 1 (a and) and Synthetic 2 (c and d) datasets.

NeNMF: Optimal Gradient Method

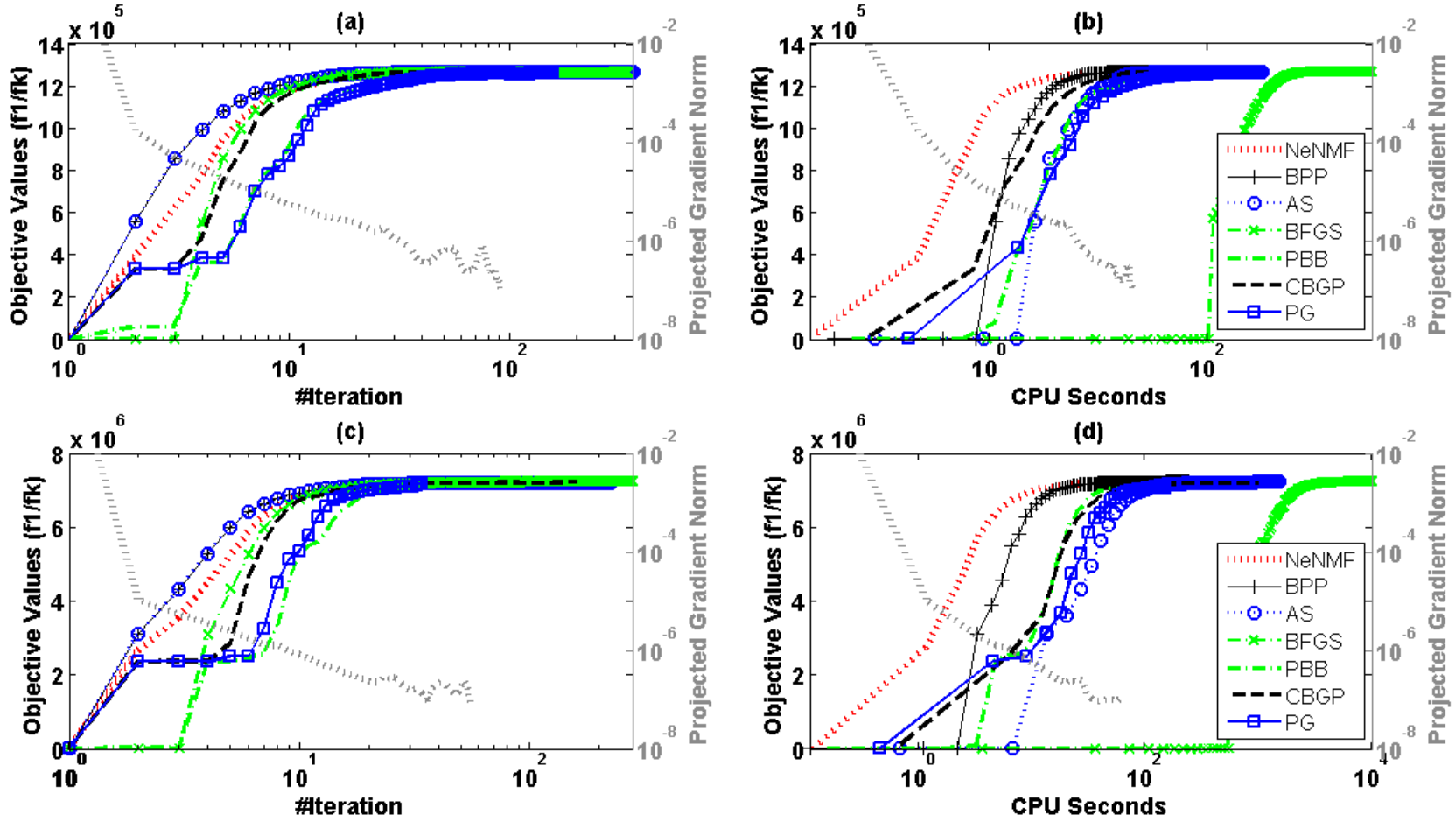


Fig. 5. Average objective values versus iteration numbers and CPU seconds of NeNMF, PG-, AS-based NMF solvers on the Reuters-215781 (a and) and TDT-2 (c and d) datasets.

NeNMF: Optimal Gradient Method

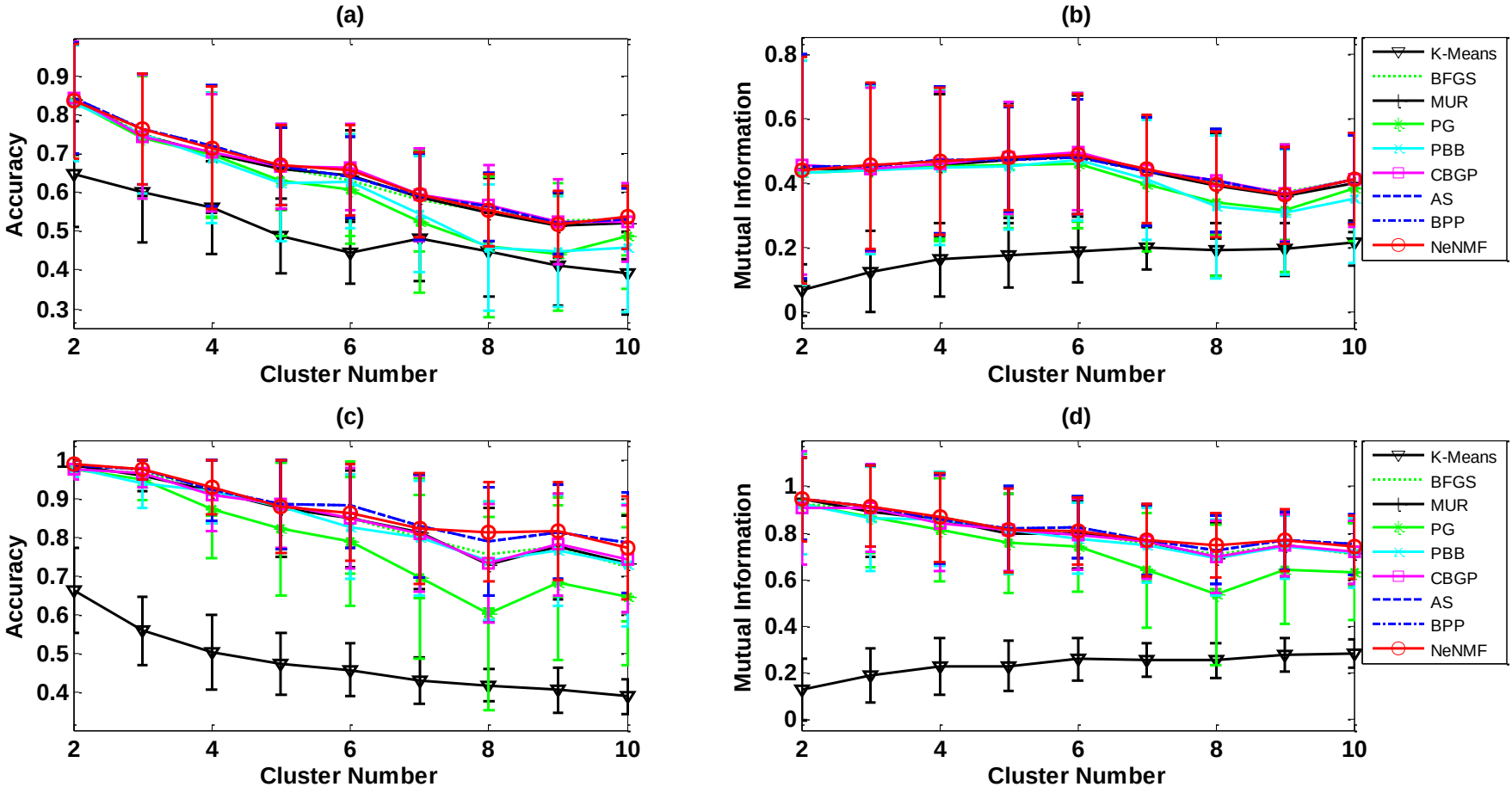


Fig. 7. The average accuracy and mutual information versus cluster number obtained NeNMF or other NMF solvers on Reuters-215781 (a and) and TDT-2 (c and d) datasets. All solvers start from the same initial point and stop when **the same stopping criterion** is met.

NeNMF: Optimal Gradient Method

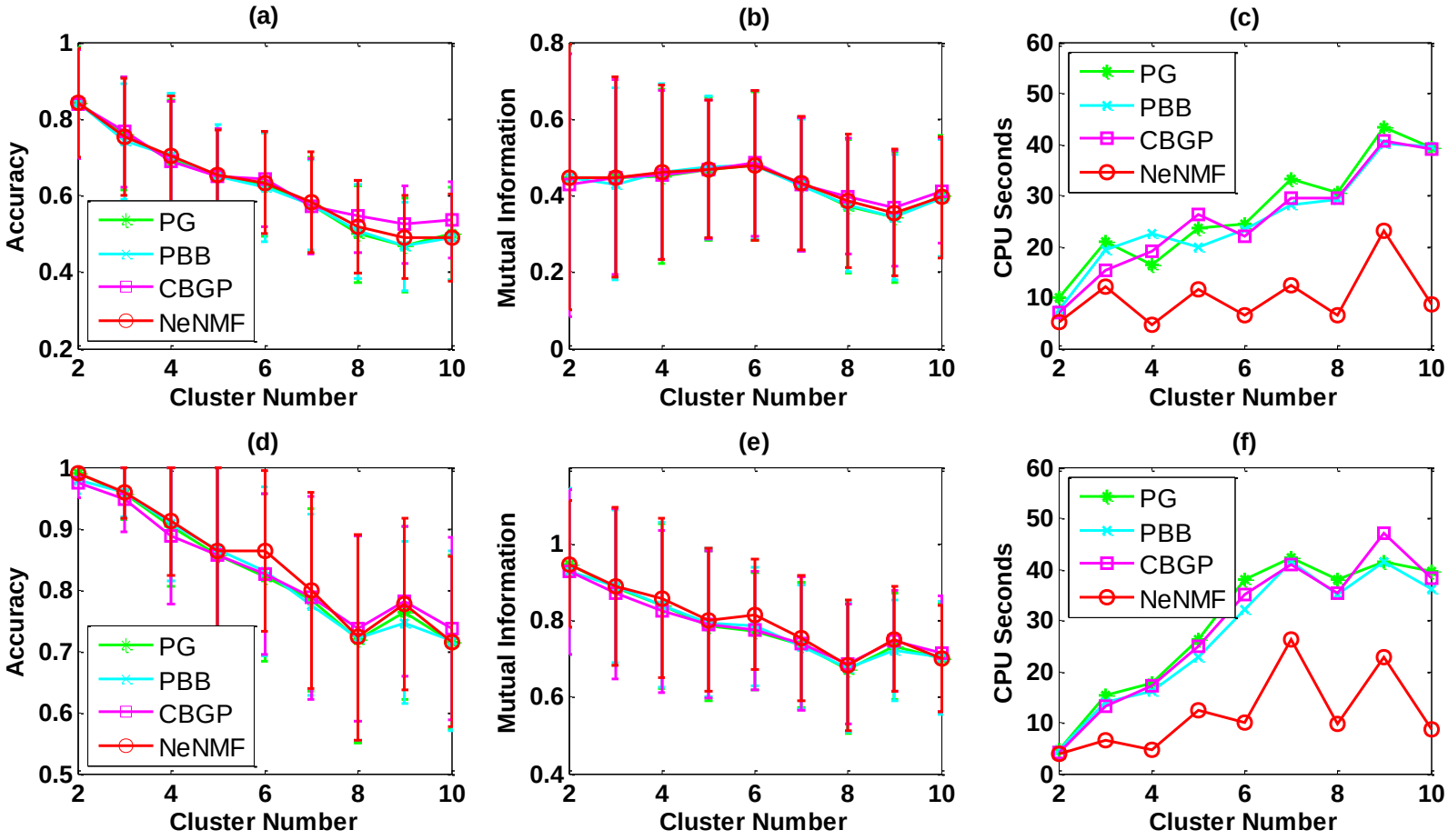


Fig. 8. The average accuracy and mutual information versus cluster number obtained NeNMF or other NMF solvers on Reuters-215781 (a and) and TDT-2 (c and d) datasets. All solvers start from the same initial point and stop with **identical objective value**.

NeNMF: Optimal Gradient Method

***Remark 1*:** NeNMF converges much faster than the existing methods.

***Remark 2*:** Under the same stopping condition, NeNMF obtains a better approximation.

***Remark 3*:** NeNMF cost much less CPU seconds to reach the same objective value.

***Remark 4*:** NeNMF can be naturally adopted to NMF extensions, e.g., L_1- and L_2-norm regularized NMF, and box-constrained NMF.

NeNMF for NPAF

Euclidean distance based NPAF: Revisit

$$\arg\min_{W\geq 0, H\geq 0} \frac{\gamma}{2}\mathrm{tr}\left(HL^{NPAF}H^T\right)+\frac{1}{2}\left\|X-WH\right\|_F^2$$

Fixed W:

convex

$$\arg\min_{H\geq 0} \varphi(H)=\frac{\gamma}{2}\mathrm{tr}\left(HL^{NPAF}H^T\right)+\frac{1}{2}\left\|X-WH\right\|_F^2$$

***Proposition 2*:** The gradient of $\mathrm{tr}\left(HL^{NPAF}H^T\right)$ is Lipschitz continuous and the corresponding Lipschitz constant is $\left\|L^{NPAF}\right\|_2$.

NeNMF for NPAF

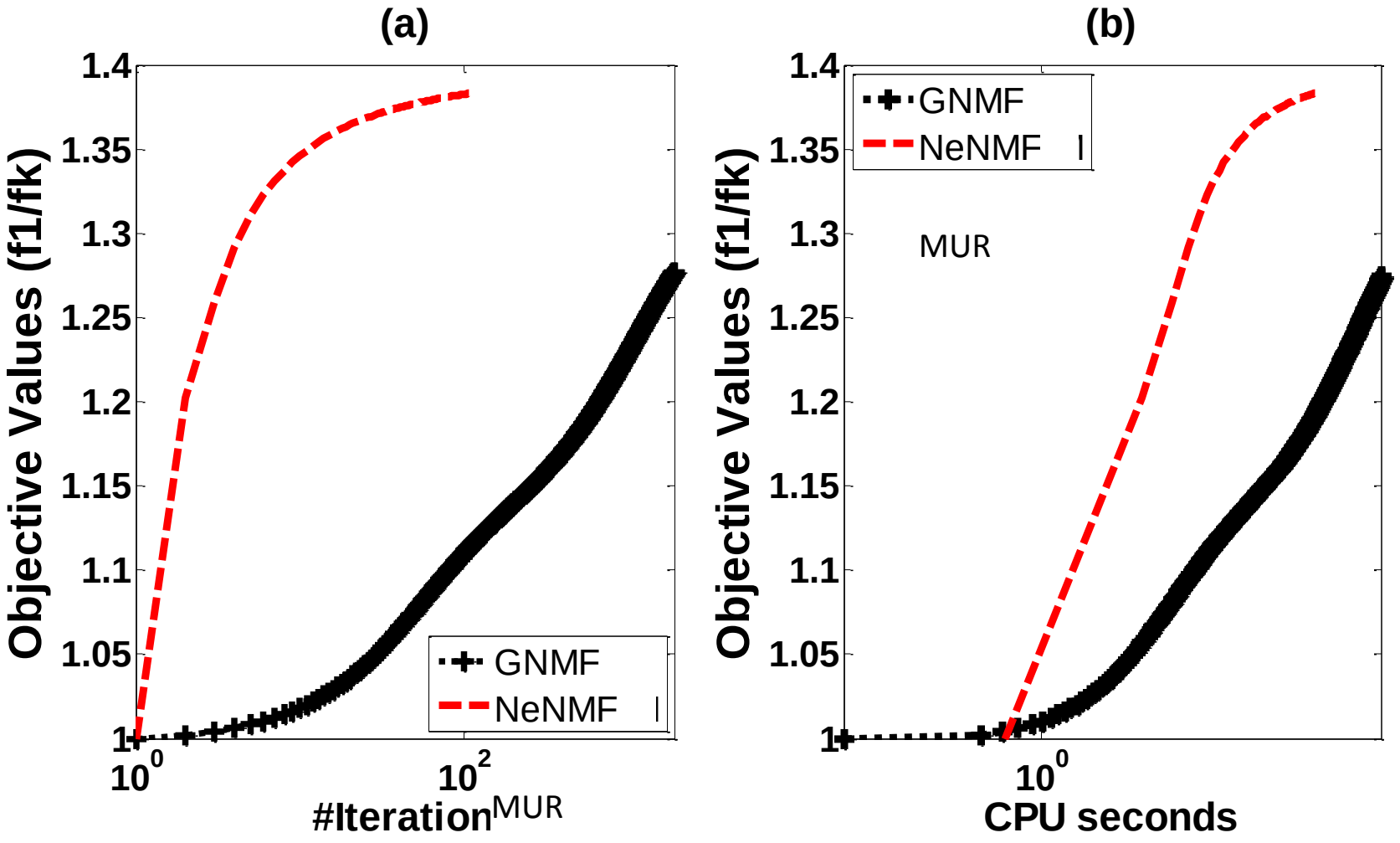


Fig. 11. NeNMF versus MUR for GNMF on 1600x320-D synthetic data matrix.

NeNMF for NPAF

OGM versus MFGD:

1) For optimizing each factor matrix, MFGD converges at the rate of $O(\frac{1}{k})$ while OGM converges at the rate of $O(\frac{1}{k^2})$.

2) MFGD can optimize both *KL*-divergence and Frobenius norm based NPAF while OGM can only optimize Frobenius norm based NPAF.

N. Guan, D. Tao, Z. Luo, and B. Yuan., "NeNMF: An Optimal Gradient Method for Non-negative Matrix Factorization," ***IEEE Transactions on Signal Processing***, vol. 60, no. 6, pp. 2882-2898, 2012.

Online RSA NMF

Drawback of NMF optimization methods for large-scale problems: data resides in the memory.

Table I. Existing Online NMF Optimization Methods.

Online NMF Algorithms	References
Online Non-negative Matrix Factorization (ONMF)	(Cao *et al*. 2007)
Incremental Non-negative Matrix Factorization (INMF)	(Bucak & Gunsel, 2009)
Online Matrix Factorization (OMF)	(Mairal *et al*., 2010)
OMF with Diagonal Approximation (OMF-DA)	(Wang *et al*. 2011)

1) ONMF and INMF: convergence is not guaranteed.
2) OMF and OMF-DA: suffer from numerical instability especially when the dataset is sparse, because they ignore the off-diagonal information in Hessian.

Online RSA NMF

Treat samples as random variables:

$$\arg\min_{W\geq 0} f_n(W) = \arg\min_{W\geq 0} \left\{ \frac{1}{n}\sum_{i=1}^{n} l\left(x^i, W\right) \right\}$$

sample

$$l\left(x^i, W\right) = \min_{h^i\geq 0} \frac{1}{2}\left\| x^i - Wh^i \right\|^2$$

Let *n* go to infinity, we expect:

random variable

$$\arg\min_{W\geq 0} f(W) = \arg\min_{W\geq 0} E_{x\in\wp}\left(l\left(x, W\right)\right)$$

non-convex

Online RSA NMF

Alternating Optimization:

$$h^t = \arg\min_{h^t \geq 0} \frac{1}{2}\left\| x^t - W^{t-1} h^t \right\|^2 \qquad \text{(NNLS)}$$

$$W^t = \arg\min_{W^t \geq 0} E_{x \in \wp_t} \left(\frac{1}{2}\left\| x - W^t h \right\|^2\right), \qquad \text{(S)}$$

$$\wp_t = span\{x^1, ..., x^t\} \subset \wp$$

Denote: $G_t(W, x) = \frac{1}{2}\left\| x - Wh \right\|^2$

convex with respect to W.

Online RSA NMF

Robust Stochastic Approximation (Nemirovski, 09) for (S):

$$W_{k+1} = \Pi_C(W_k - r_k \nabla_W G_t(W_k, x_k)), W_1 = W^{t-1}$$

Projection onto C:

$$C = \{W = [w_1, \ldots, w_r] \mid \|w_j\|_1 = 1, w_j \geq 0, j = 1, \ldots, r\}$$

Smartly Choosing Learning Rate:

$$r_k = \frac{\theta^t D_W}{M_* \sqrt{k}}, \begin{cases} D_W = \max_{W \in C} \{\|W - W_1\|_F\} \\ M_* = \max_{W \in C} \sqrt{E_{x \in \wp_t}(\|\nabla_W G_t(W, x)\|_F^2)} \end{cases}$$

Online RSA NMF

Algorithm 1: Online RSA-NMF **(OR-NMF)**

Input: $x \geq 0, x \in \wp, T, r$

Output: $W \geq 0$

1: Initialize $W^0 \geq 0, B \leftarrow \phi$

For $t = 1$ to T **do**

2: Draw x^t from $\wp$

3: calculate y^t by NNLS

4: add x^t to B

5: $\theta^t = .1\cos(\frac{(t-1)\pi}{2T})$

6: update W^t with Algorithm 2

End For

7: $W = W^T$

Algorithm 2: Update basis matrix

Input: $W^{t-1}, \theta^t, B, \tau$

Output: $W^t \geq 0$

1: $\sum_W \leftarrow 0, \sum \leftarrow 0, W_1 / W_1^t \leftarrow W^{t-1}, k \leftarrow 1$

Repeat

2: $r_k = \theta^t D_W / M_* \sqrt{k}$

3: $\sum_W \leftarrow \sum_W + r_k W_k, \sum \leftarrow \sum + r_k$

4: Draw x_k by cycling on permuted B

5: $W_{k+1} = \Pi_C (W_k - r_k \nabla_W G_t (W_k, x_k))$

6: $W_{k+1}^t = \sum_W / \sum, k \leftarrow k+1$

Until $\left\| W_k^t - W_{k-1}^t \right\|_F \Big/ \left\| W_{k-1}^t \right\|_F \leq \tau$

7: $W^t = W_K^t$

Online RSA NMF

Algorithm settings:

$$D_W = \sqrt{2r}$$ (diameter of simplex is $\sqrt{2}$)

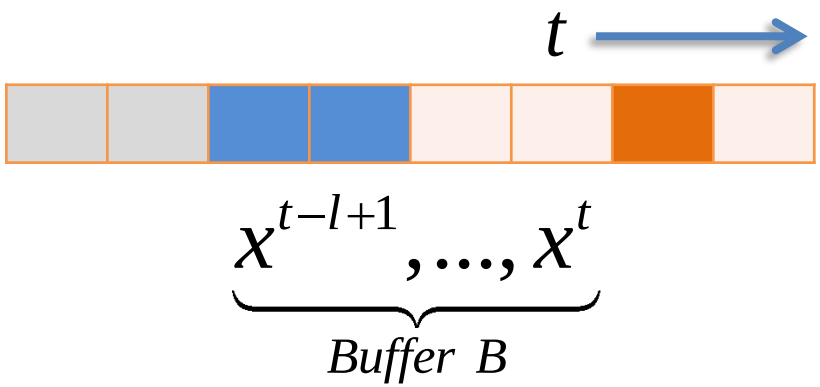
$$M_* = \max_{i=1,\cdots,k} \sqrt{E_{x \in \wp_t} \left\| \nabla_W G_t \left(W_i, x \right) \right\|_F^2}$$ (adaptive updating)

$$\theta^t = .1\cos(\frac{(t-1)\pi}{2T})$$ (avoid oscillation in last iterations)

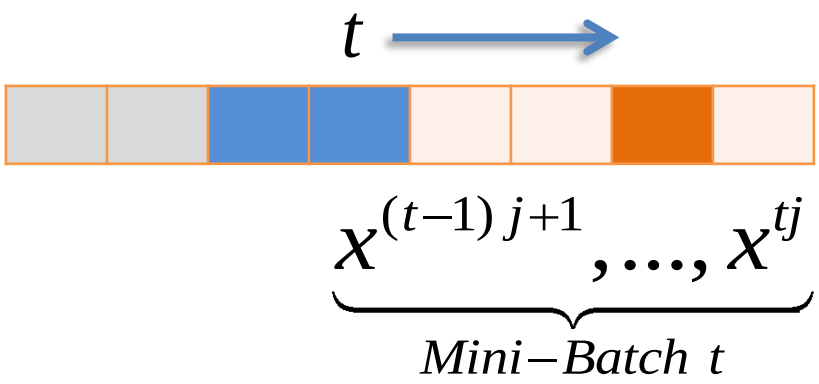
***Theorem 1*:** The objective function $f(W)$ converges almost surely under *Algorithm 1*.

Online RSA NMF

Buffering Strategy:



Mini-Batch Extension:



Algorithm 1: Modified OR-NMF **(MOR-NMF)**

Input: $x \geq 0, x \in \wp, T, r$

Output: $W \geq 0$

1: Initialize $W^0 \geq 0, B \leftarrow \phi$

For $t = 1$ to T **do**

2: Draw x^t from $\wp$

3: calculate y^t by NNLS

4: replace x^{t-l} with x^t in B when $t > l$

5: $\theta^t = .1\cos(\frac{(t-1)\pi}{2T})$

6: update W^t with Algorithm 2

End For

7: $W = W^T$

Online RSA NMF

***Remark 1*:** In each iteration round, OR-NMF almost surely converges to the basis at the rate of $O(\frac{1}{\sqrt{k}})$.

***Remark 2*:** The objective function $f(W)$ almost surely converges under OR-NMF.

***Remark 3*:** OR-NMF can handle L_1- and L_2-norm regularized NMF, box-constrained NMF, and Itakura-Saito (IS) divergence based NMF.

***Remark 4*:** By using the buffering strategy, MOR-NMF saves memory and thus fits large-scale problem well.

Online RSA NMF

Experiment Settings:

	Efficiency Comparison				Face Recognition			Image Annotation			
Dataset	m	n	r	sp	#TR	#TS	r	#TR	#TS	#VC	#KD
CBCL	361	500	10/50	.131	30/50/70	470/450/430	10-80				
ORL	1,024	400	10/50	.042	120/200/280	280/200/120	10-150				
IAPR TC12	100	500	50/80	.745				17, 825	1980	291	4.7

Algorithm Settings (continue):

L_1-ball projection operator: (Duchi *et al.*, 2008).

Online RSA NMF

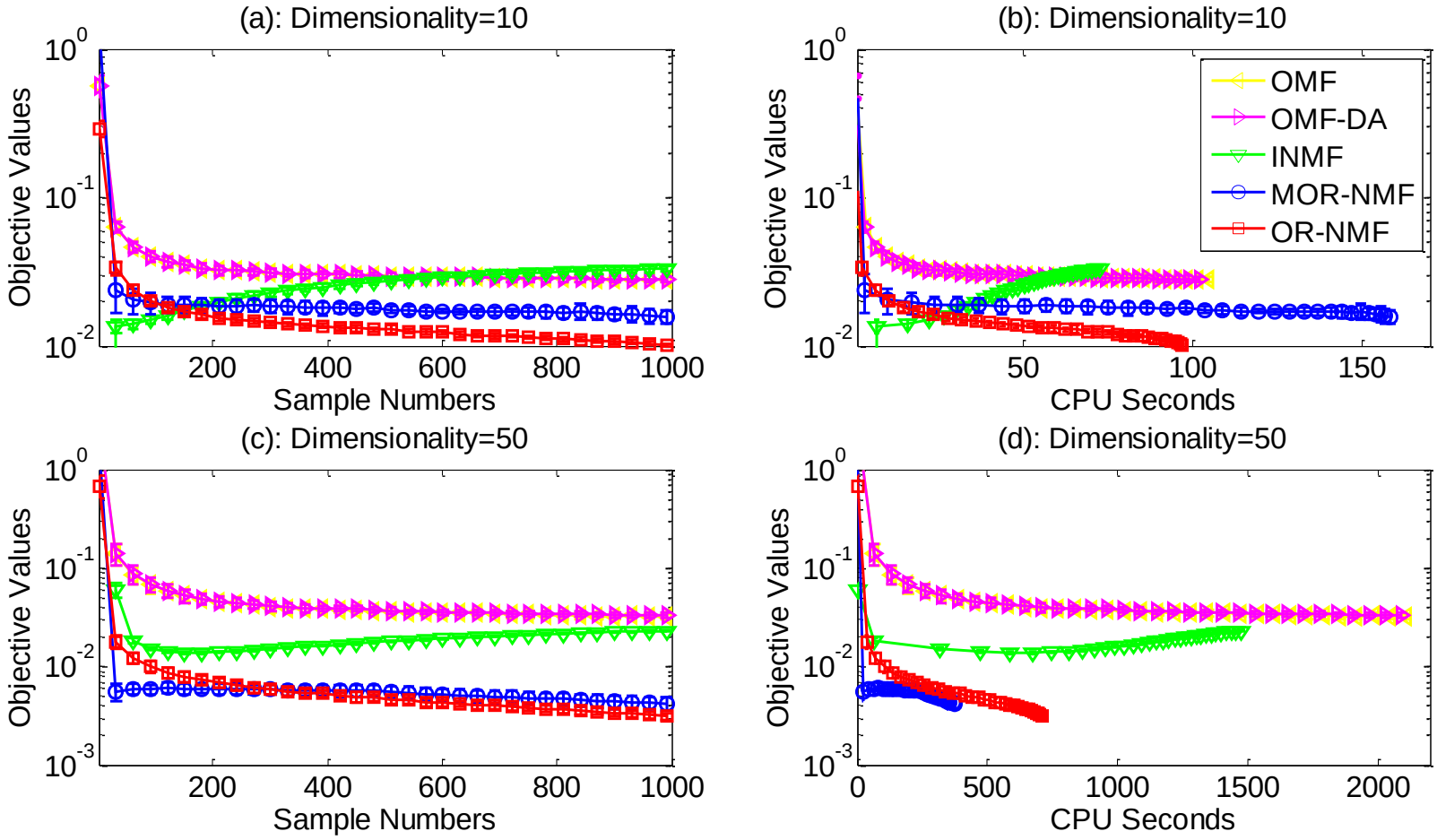


Fig. 2. Objective values versus iteration numbers and CPU seconds of OR-NMF, MOR-NMF, OMF, OMF-DA, and INMF on the CBCL dataset.

Online RSA NMF

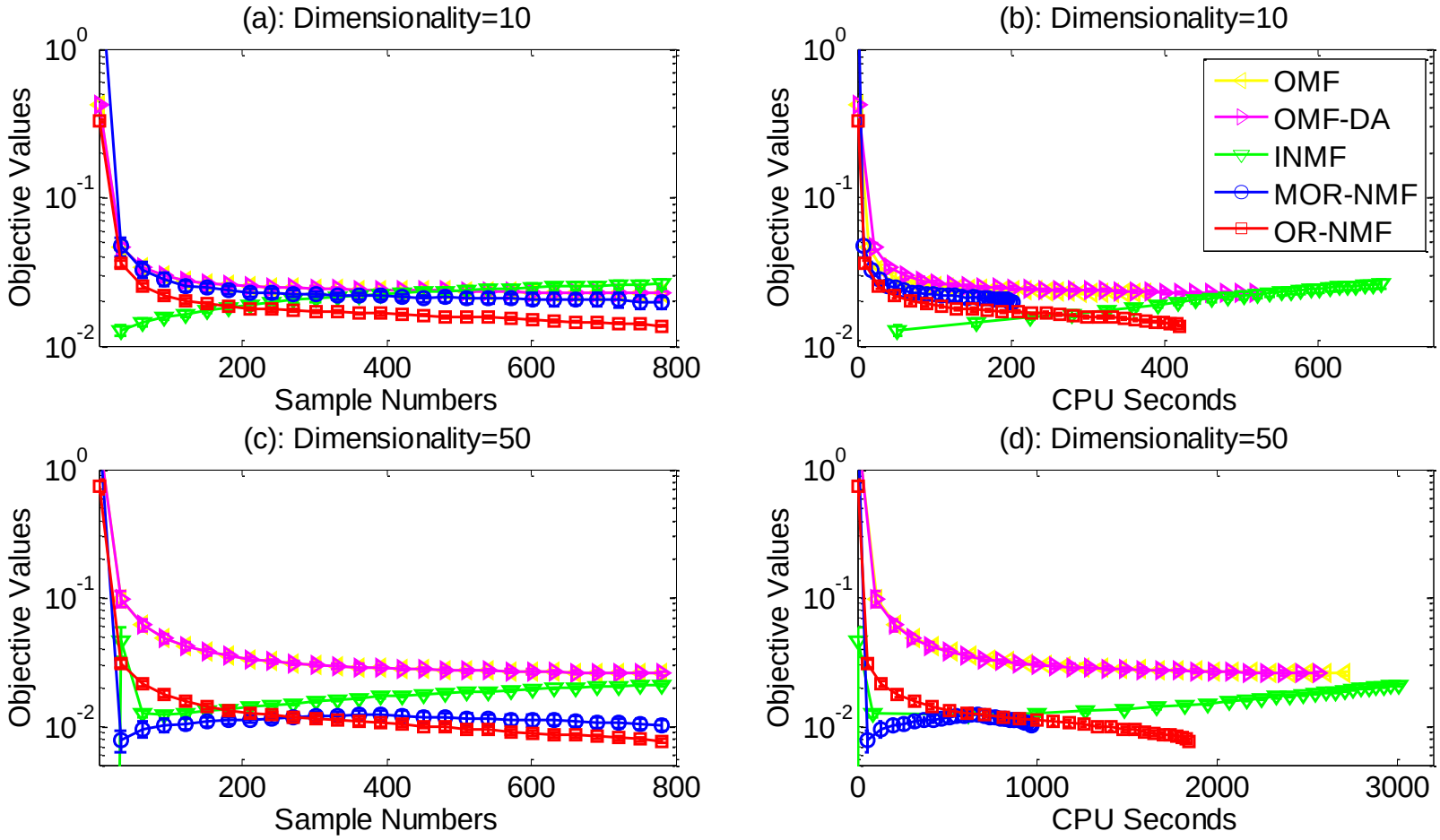


Fig. 3. Objective values versus iteration numbers and CPU seconds of OR-NMF, MOR-NMF, OMF, OMF-DA, and INMF on the ORL dataset.

Online RSA NMF

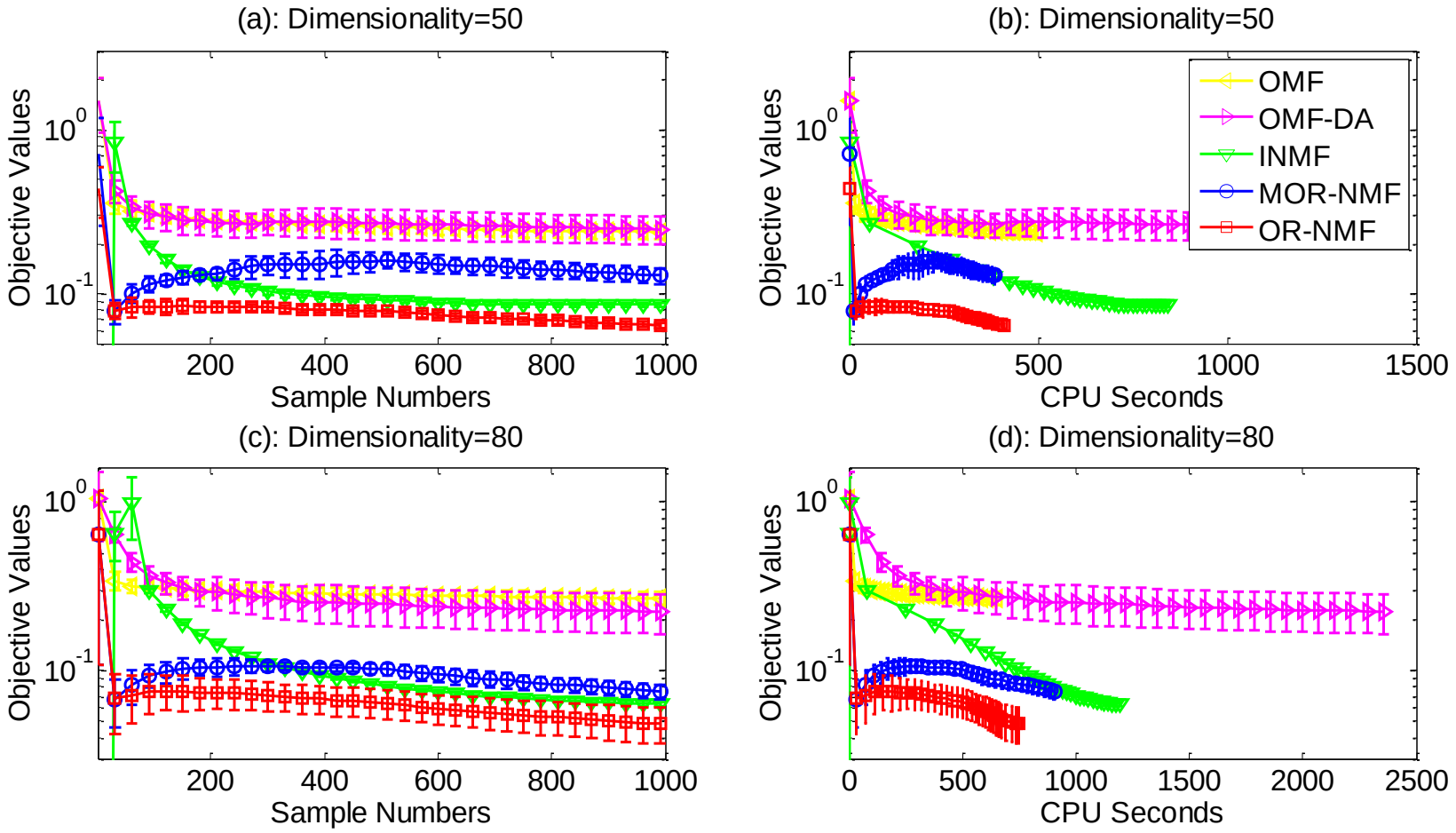


Fig. 6. Objective values versus iteration numbers and CPU seconds of OR-NMF, MOR-NMF, OMF, OMF-DA, and INMF on the ICPR TC12 dataset.

Online RSA NMF

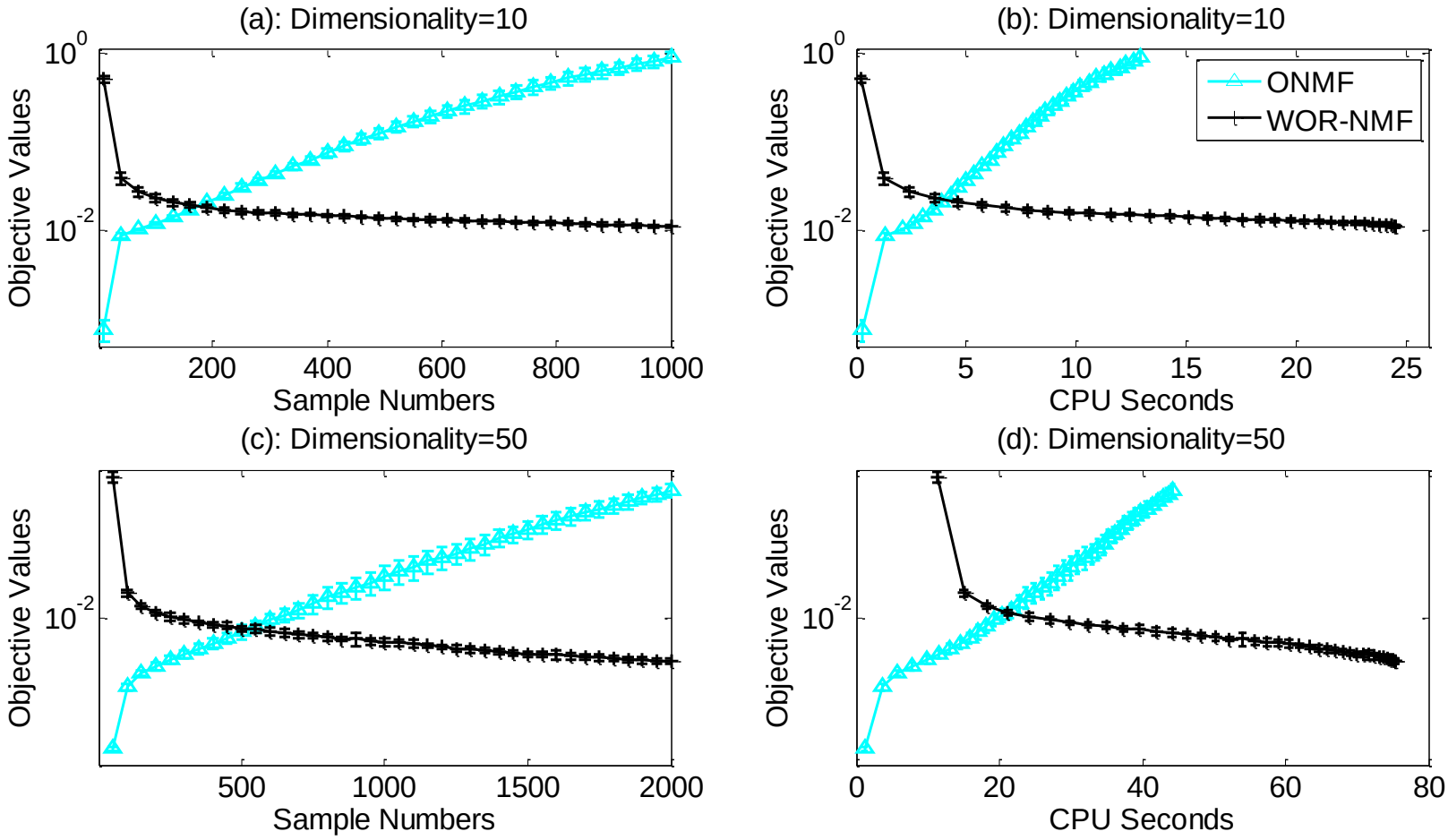


Fig. 4. Objective values versus iteration numbers and CPU seconds of WOR-NMF and ONMF on the CBCL dataset.

Online RSA NMF

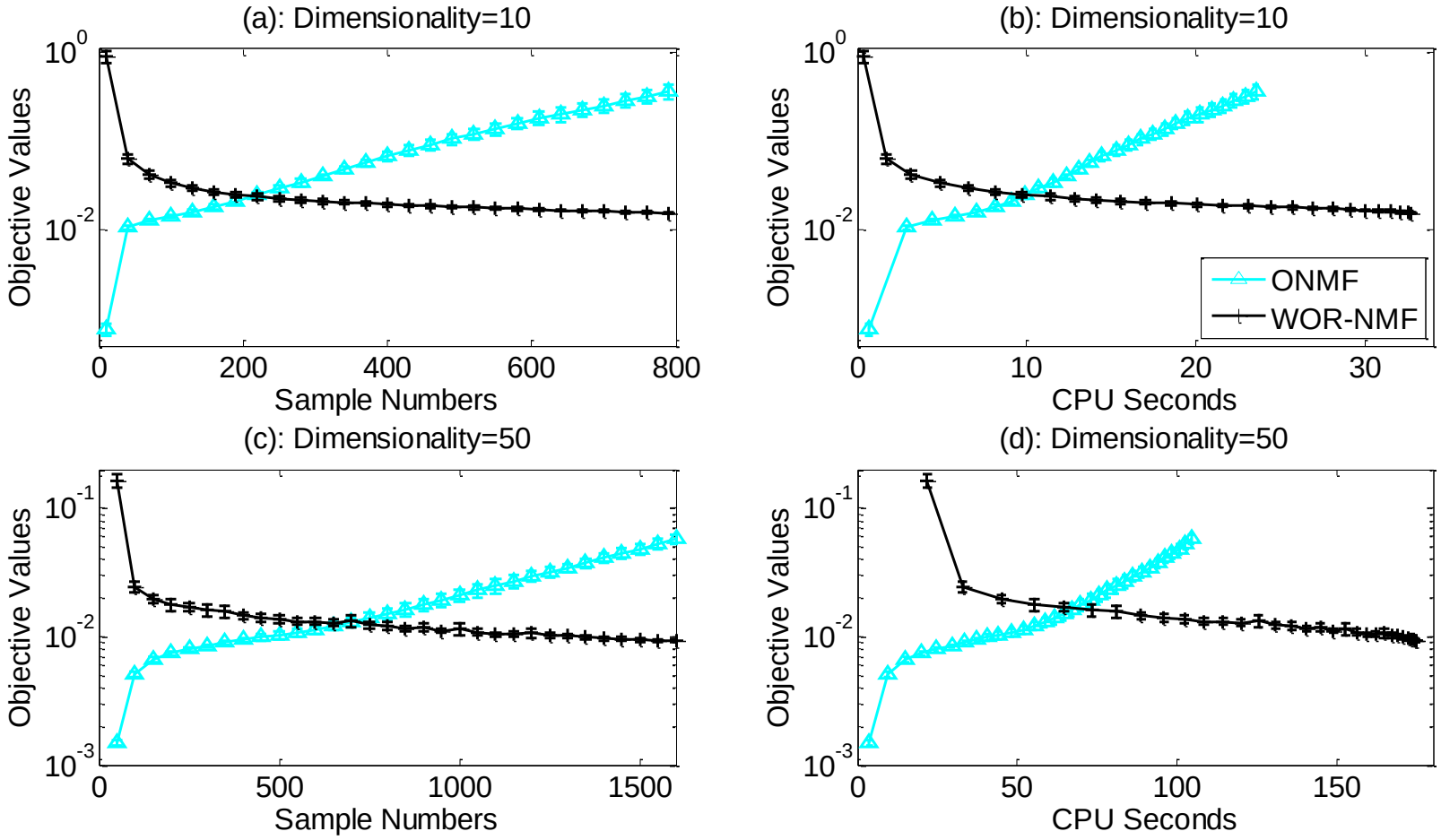


Fig. 5. Objective values versus iteration numbers and CPU seconds of WOR-NMF and ONMF on the ORL dataset.

Online RSA NMF

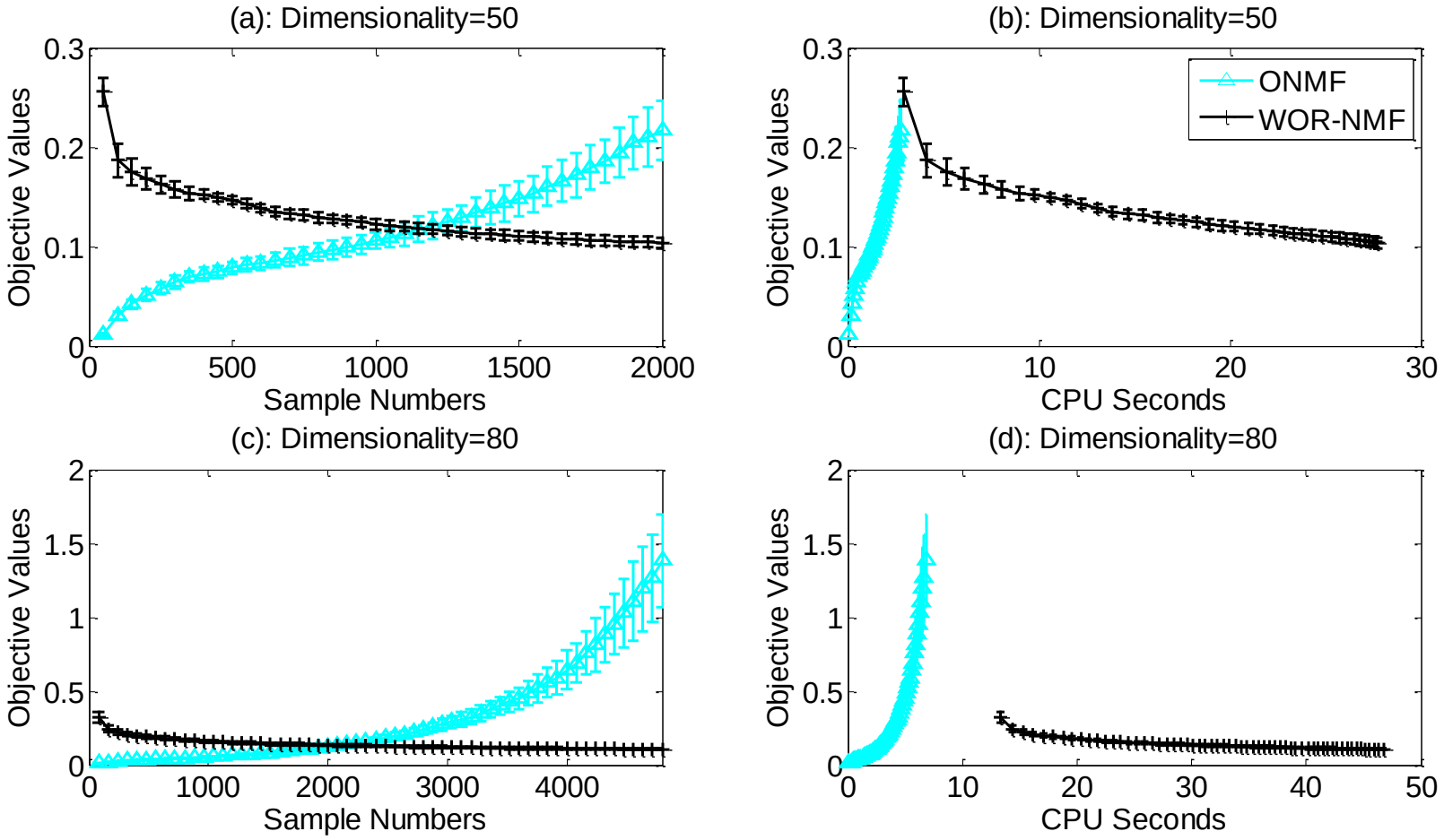


Fig. 8. Objective values versus iteration numbers and CPU seconds of WOR-NMF and ONMF on the IAPR TC12 dataset.

Online RSA NMF

Face Recognition:

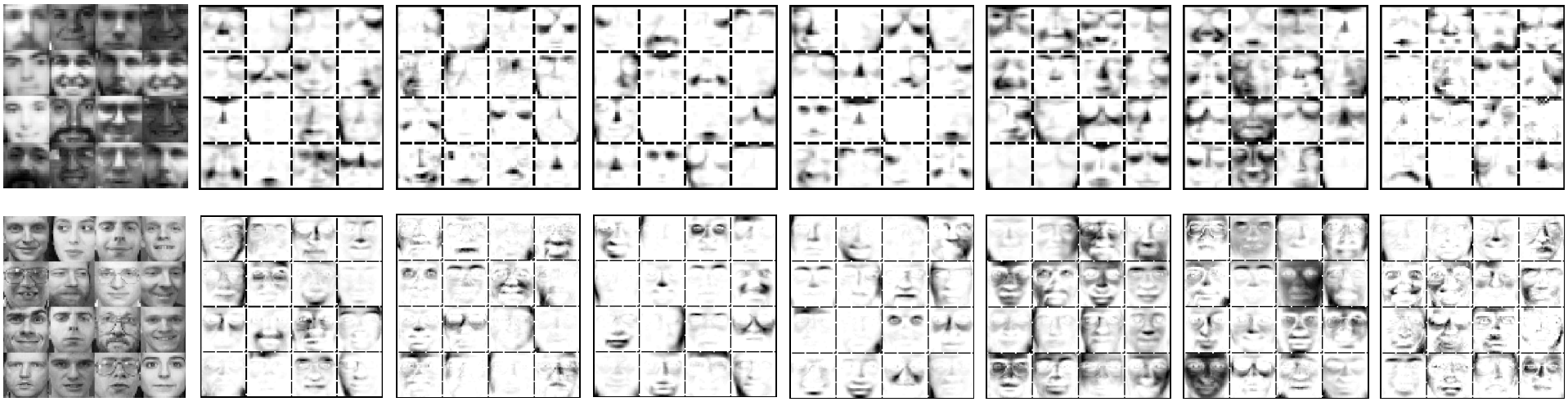


Fig. 1. Image examples (a), learned bases by OR-NMF (b), MOR-NMF (c), OMF (d), OMF-DA (e), INMF (f), WOR-NMF (g), and ONMF (h) on CBCL (1st row) and ORL (2nd row) datasets.

Online RSA NMF

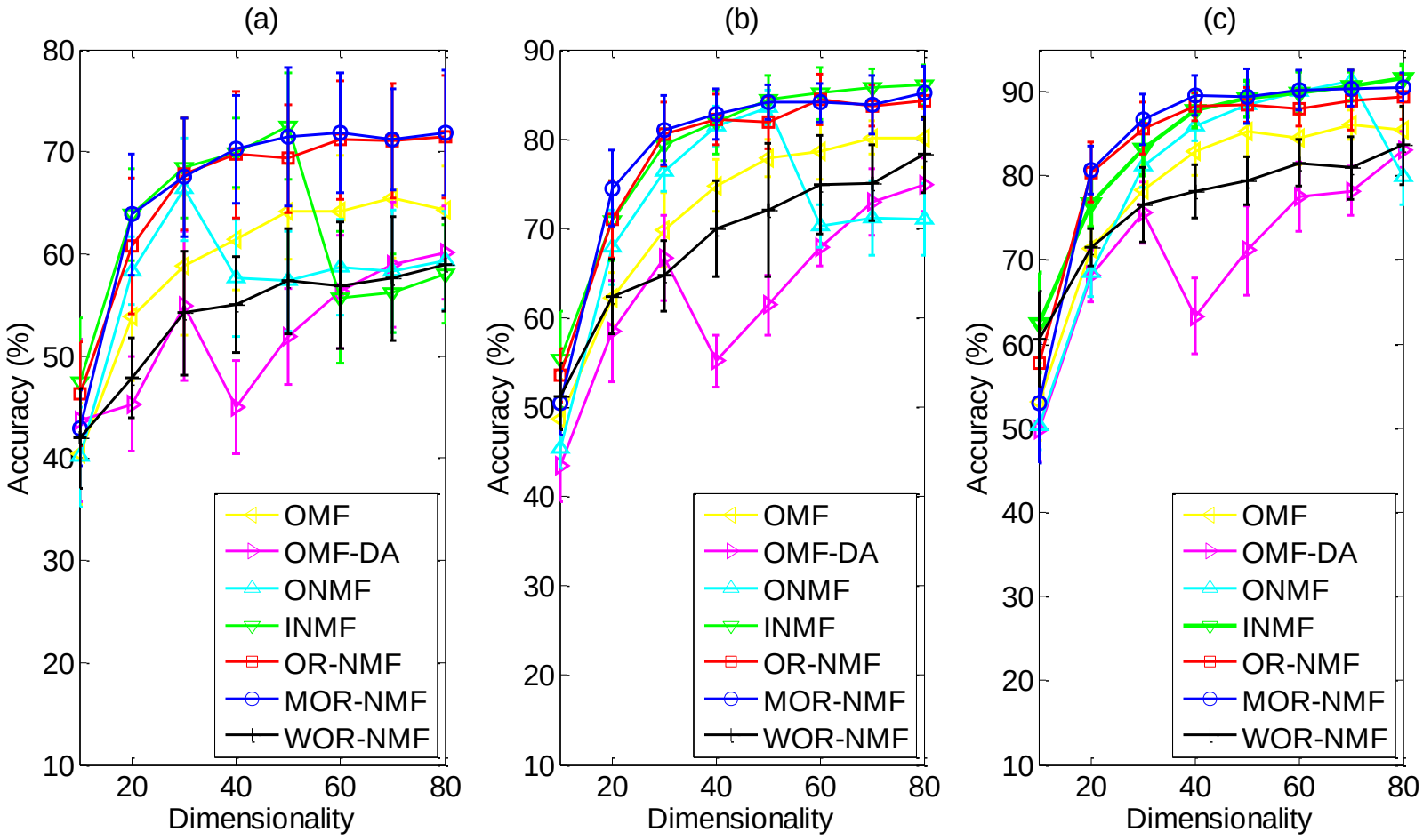


Fig. 10. Mean and deviation of accuracy versus reduced dimensionality of OR-NMF, MOR-NMF, WOR-NMF, OMF, OMF-DA, ONMF, and INMF on the CBCL dataset whereby the training set was composed of (a) three, (b) five, and (c) seven images selected from each individual.

Online RSA NMF

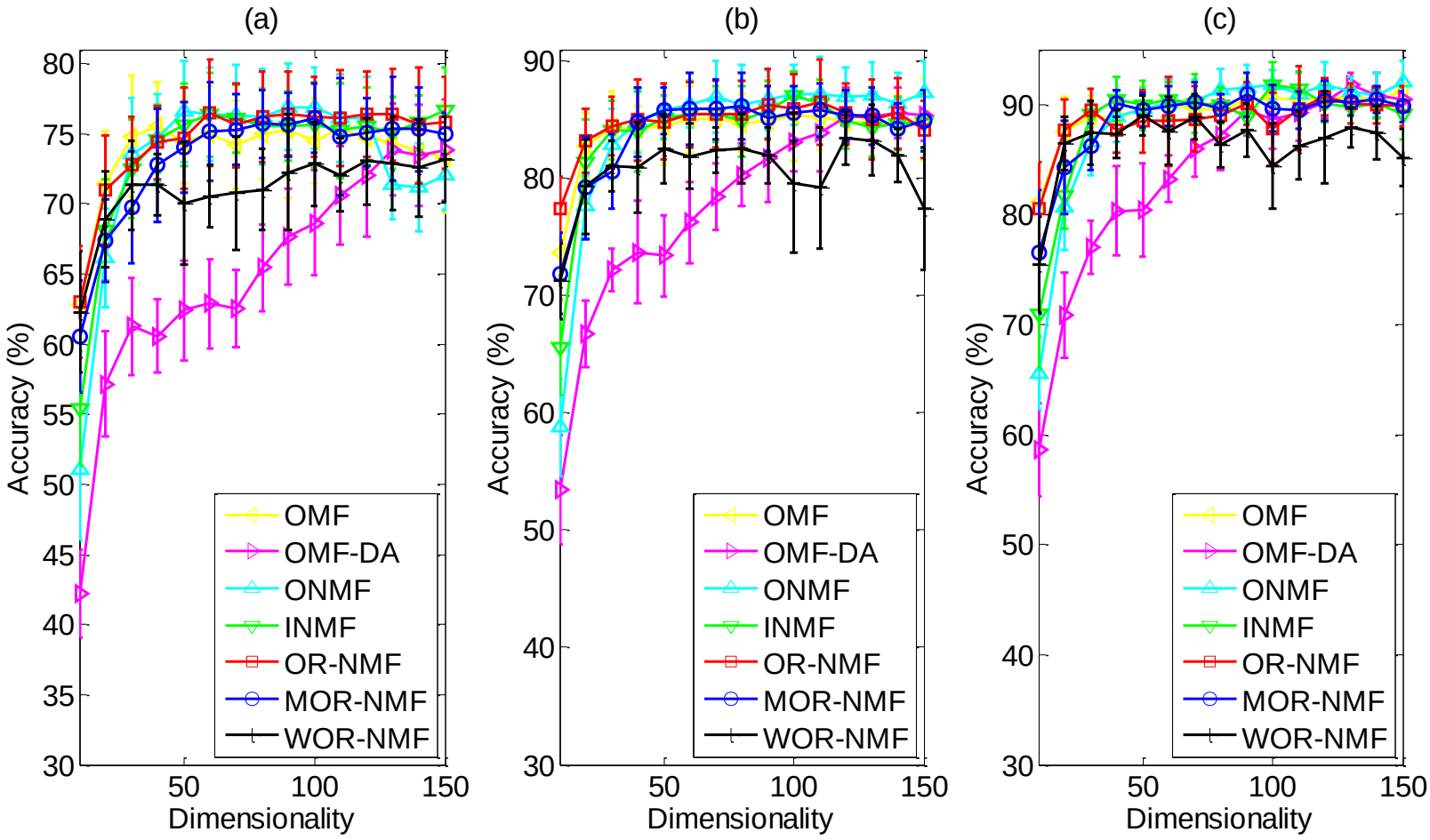
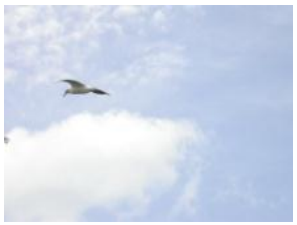


Fig. 11. Mean and deviation of accuracy versus reduced dimensionality of OR-NMF, MOR-NMF, WOR-NMF, OMF, OMF-DA, ONMF, and INMF on the ORL dataset whereby the training set was composed of (a) three, (b) five, and (c) seven images selected from each individual.

Online RSA NMF

Image Annotation:

Table III. Predicted keywords by using OR-NMF versus manually annotated keywords for images in the IAPR TC12 dataset.

					
Predicted Keywords	base, building, horse, man, statue	adult, court, man, player, tennis	forest, railing, sky, snow, tree	horizon, landscape, mountain, range, sun	adult, cloud, grey, sea, sky
Manually Annotated Keywords	base, building, horse, statue	court, man, player, tennis	forest, sky, snow, tree	landscape, mountain, range, sun	cloud, grey, sea, sky

Online RSA NMF

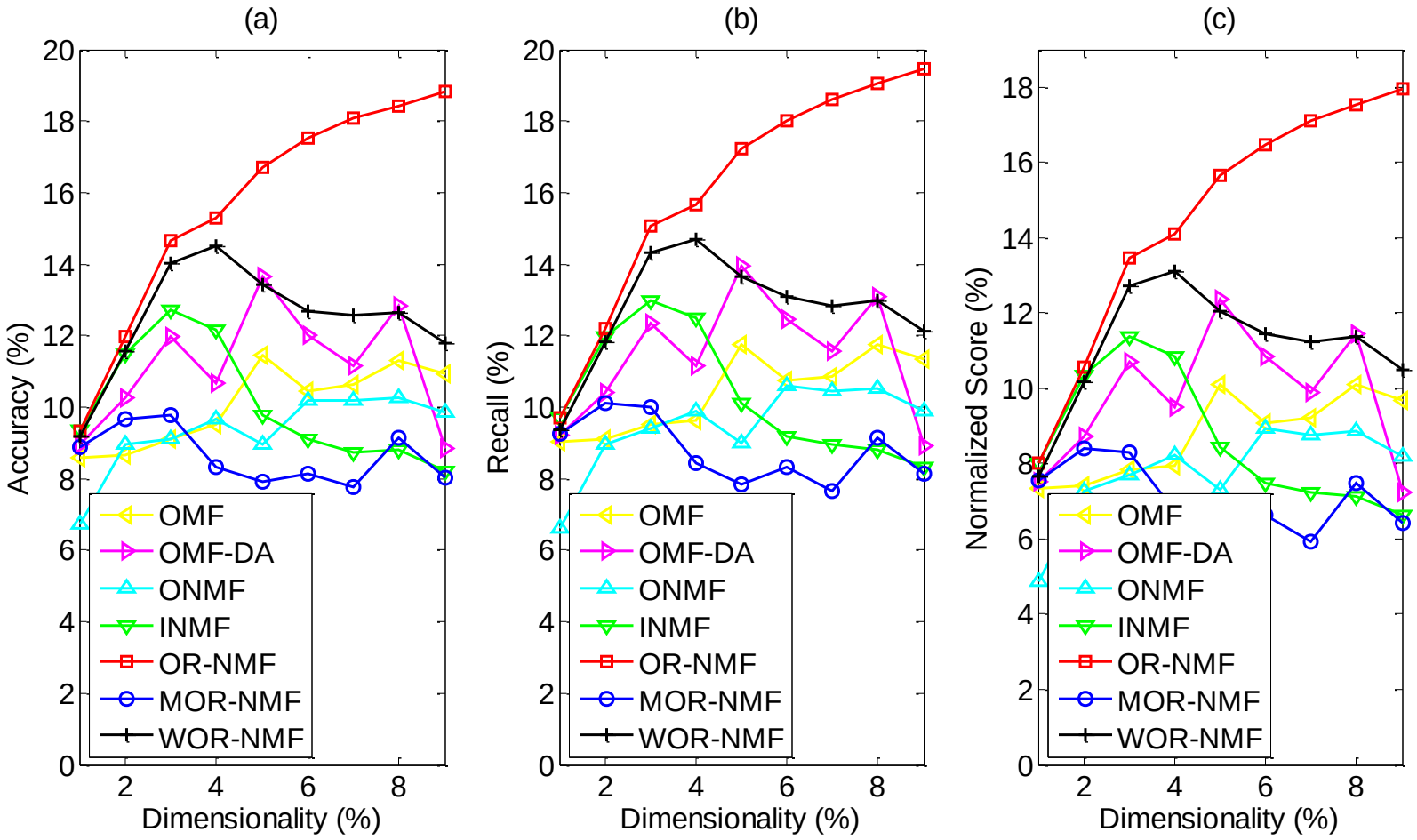


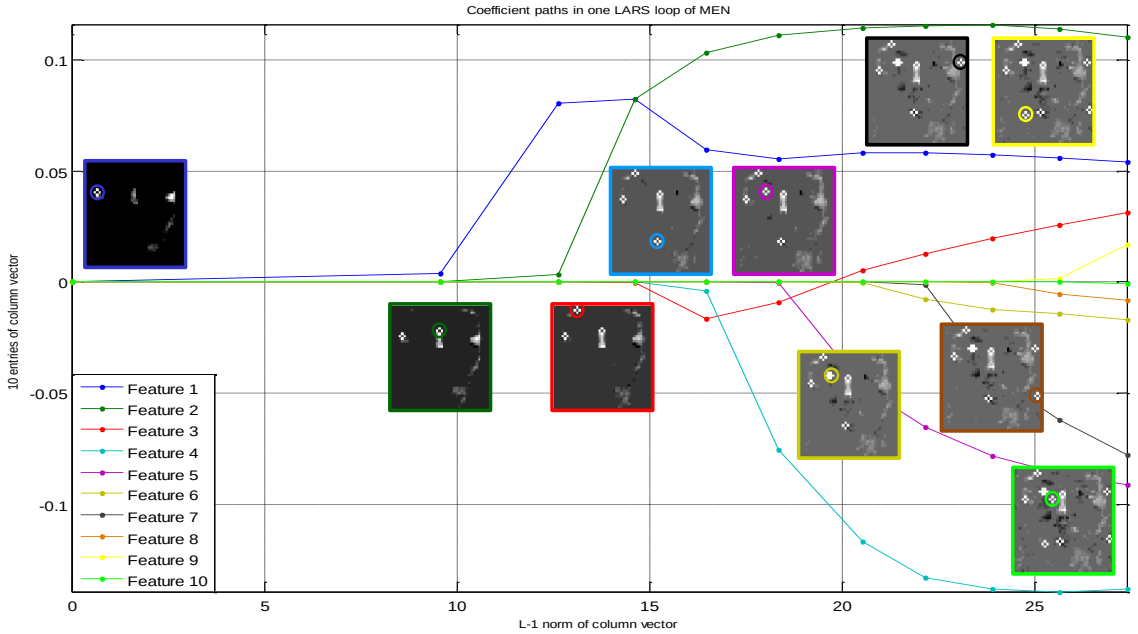
Fig. 11. Accuracy (a), recall (b), and normalized score (c) versus reduced dimensionalities of OR-NMF, MOR-NMF WOR-NMF, ONMF, INMF, OMF, and OMF-DA on the IAPR TC12 dataset.

Online RSA NMF

***Remark 1*:** OR-NMF performs consistently better than existing online NMF optimization method.

***Remark 2*:** By using OR-NMF, we can easily solve large-scale NPAF problem in an online manner.

N. Guan, D. Tao, Z. Luo, and B. Yuan., "Online Non-negative Matrix Factorization with Robust Stochastic Approximation," ***IEEE Transactions on Neural Networks and Learning Systems***, vol. 23, no. 7, pp. 1087-1099, 2012.



Sparse Patch Alignment Framework – Manifold Elastic Net

PART FOUR – 2

UTS:QCIS
QUANTUM COMPUTATION & INTELLIGENT SYSTEMS

Least squares

The term **least squares** describes a frequently used approach for solving over-determined or inexactly specified systems of equations in an approximate sense. Instead of solving the equations exactly, it seeks only to **minimize the sum of squares of residuals**.

- More than 200 years, introduced by Gauss
- Model parameterization

- What about under-determined systems, such as inverse problems?

Ill-posed problem

The mathematical term well-posed problem is given by Hadamard and well-posed problems should have the following three properties:

- A solution exists
- The solution is unique
- The solution depends continuously on the data

- Ill-posed problems: additional assumptions, such as smoothness of solution. [regularization]

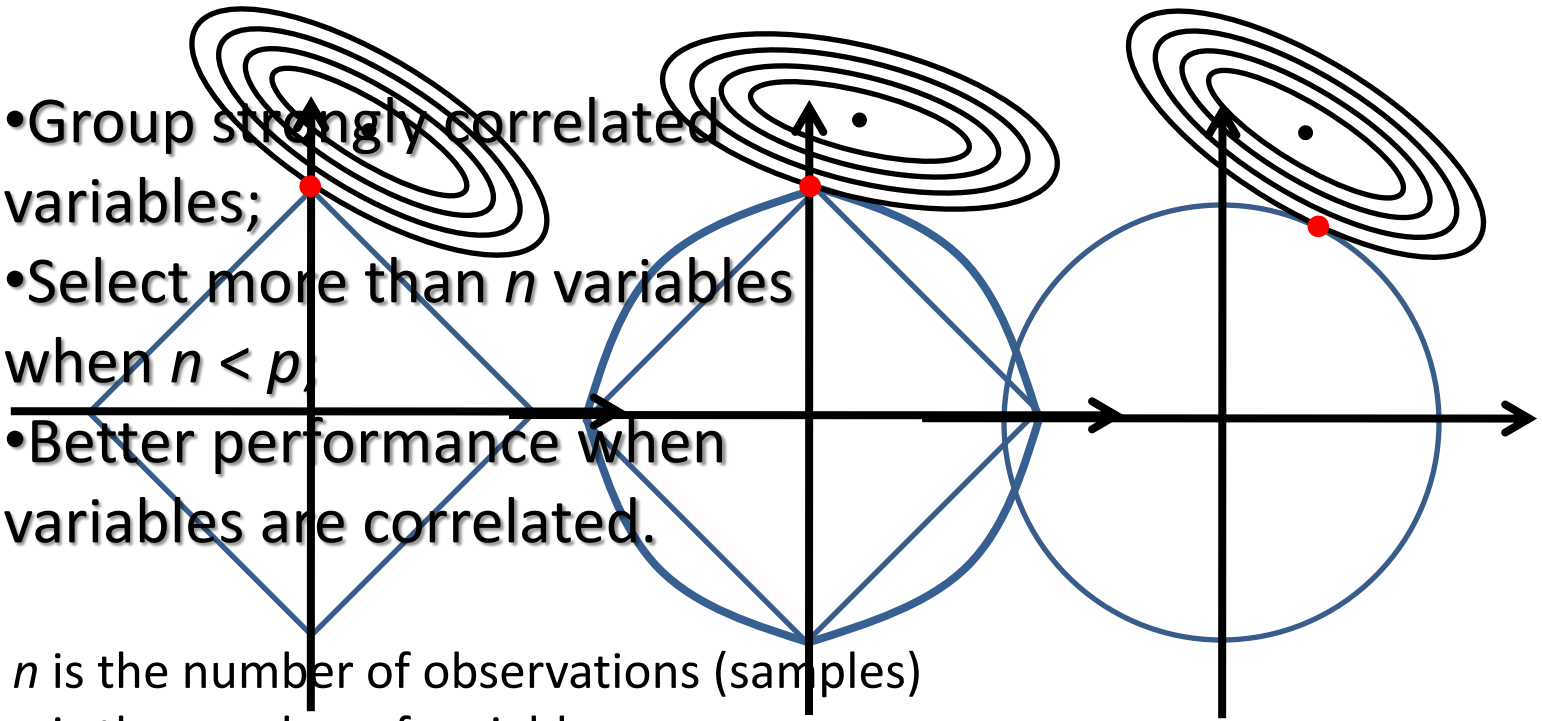
Ill-posed problem (examples)

- Classification and regression
  - L_p norm and elastic net
  - Manifold regularization
  - Bregman divergence based regularization
- Image and video processing (de-noising, de-blur, and super-resolution):
  - Tikhonov regularization
  - L_1 norm
  - Besov norm
  - Gradient consistency

Elastic net penalty

Elastic net penalty can:

- Group strongly correlated variables;
- Select more than *n* variables when *n* < *p*;
- Better performance when variables are correlated.



n is the number of observations (samples)
p is the number of variables

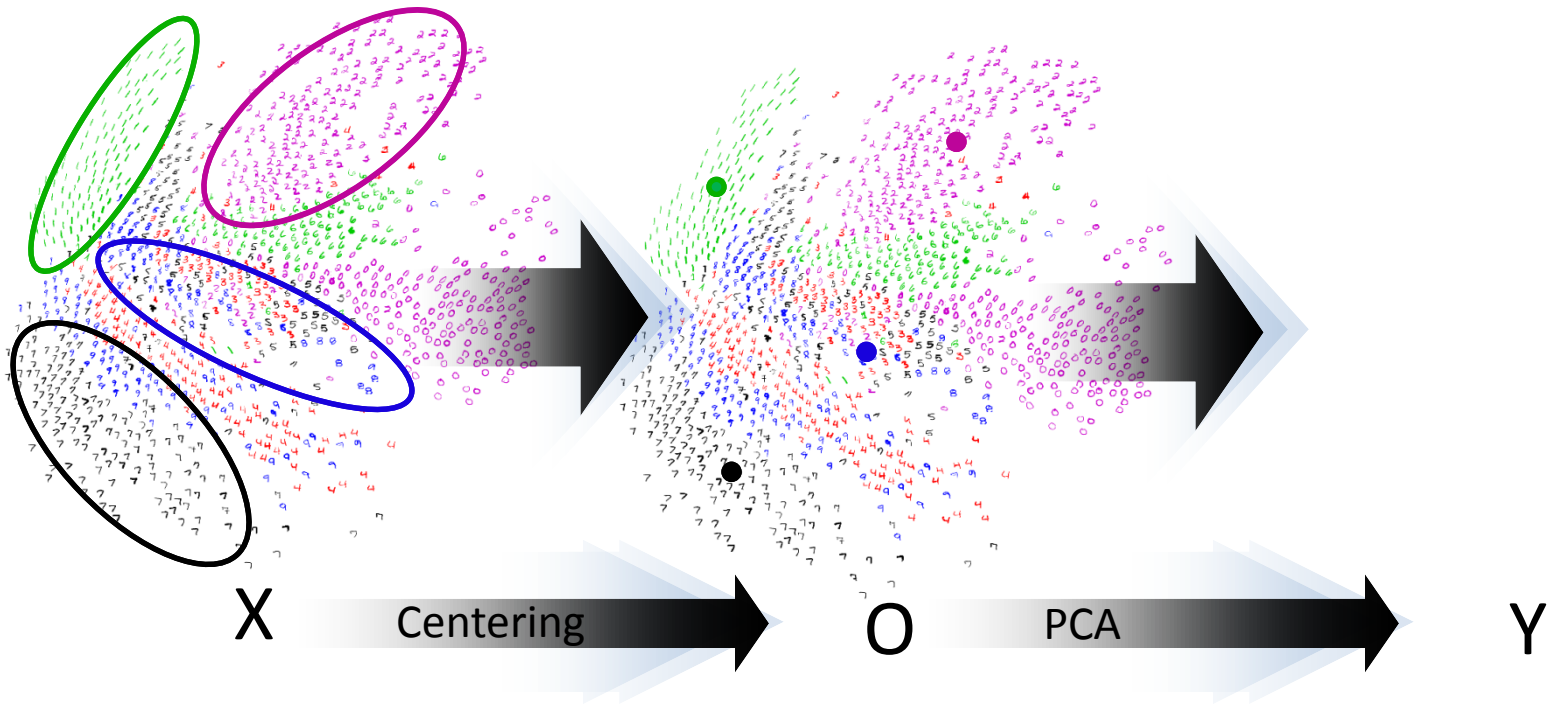
Regression for dimension reduction

- Direct least square regression formulation based on {0,1} label matrix is unstable.
  - $\min_W \|Y-XW\|_2$
  - **Y** is of size *n* (sample) by *c* (class).

- A better solution: generate an indicator matrix from both labels and data.
  - What's a suitable distribution of data representation in classification given the label matrix?

Regression for dimension reduction

Center Anchored Encoding of Indicator matrix Y:



Four different classes indicated by the indicator matrix Y

Manifold elastic net

$$\min_{Z,W} \|Y - XW\|_2^2 + \alpha \mathrm{tr}\left(Z^T L Z\right) + \beta \|Z - XW\|_2^2 + \lambda_1 \|W\|_1 + \lambda_2 \|W\|_2^2$$

Classification error

Patch alignment framework

Elastic net penalty

1. Eliminate Z

2. Least square

$$\frac{\partial F}{\partial Z} = \left(I + \alpha L^T\right) Z + 2\beta \left(Z - XW\right) = 0$$

$$A = \alpha \left(\beta \left(\alpha L + \beta I\right)^{-1}\right)^T L \left(\beta \left(\alpha L + I\right)\right) +$$

3. Eliminate L2

4. Eliminate double shrinkage

$$Y^* = \begin{bmatrix} \left(D^{1/2} U^T\right)^T X \\ \sqrt{\lambda_2} I_{p \times p} \end{bmatrix}^{-1/2} \begin{bmatrix} \left(\left(D^{1/2} U^T\right)^T\right)^{-1} Y \\ 0 \end{bmatrix}$$

$$W = \sqrt{1 + \lambda_2} W^*$$

$$\lambda = \lambda_1 / \left(1 + \lambda_2\right)$$

$$W^* = \sqrt{1 + \lambda_2} W$$

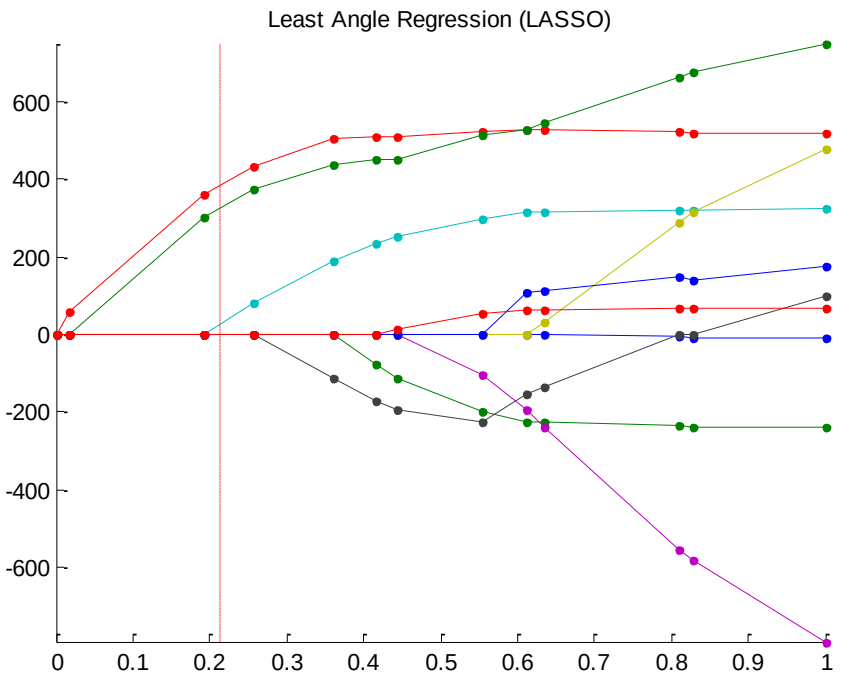
$$\min_{W^*} \|Y^* - X^* W^*\|_2^2 + \lambda \|W^*\|_1$$

Least angle regression

$$\min_{W^*} \|Y^* - X^*W^*\|_2^2 + \lambda\|W^*\|_1$$ can be solved by LARS:

In each iteration:

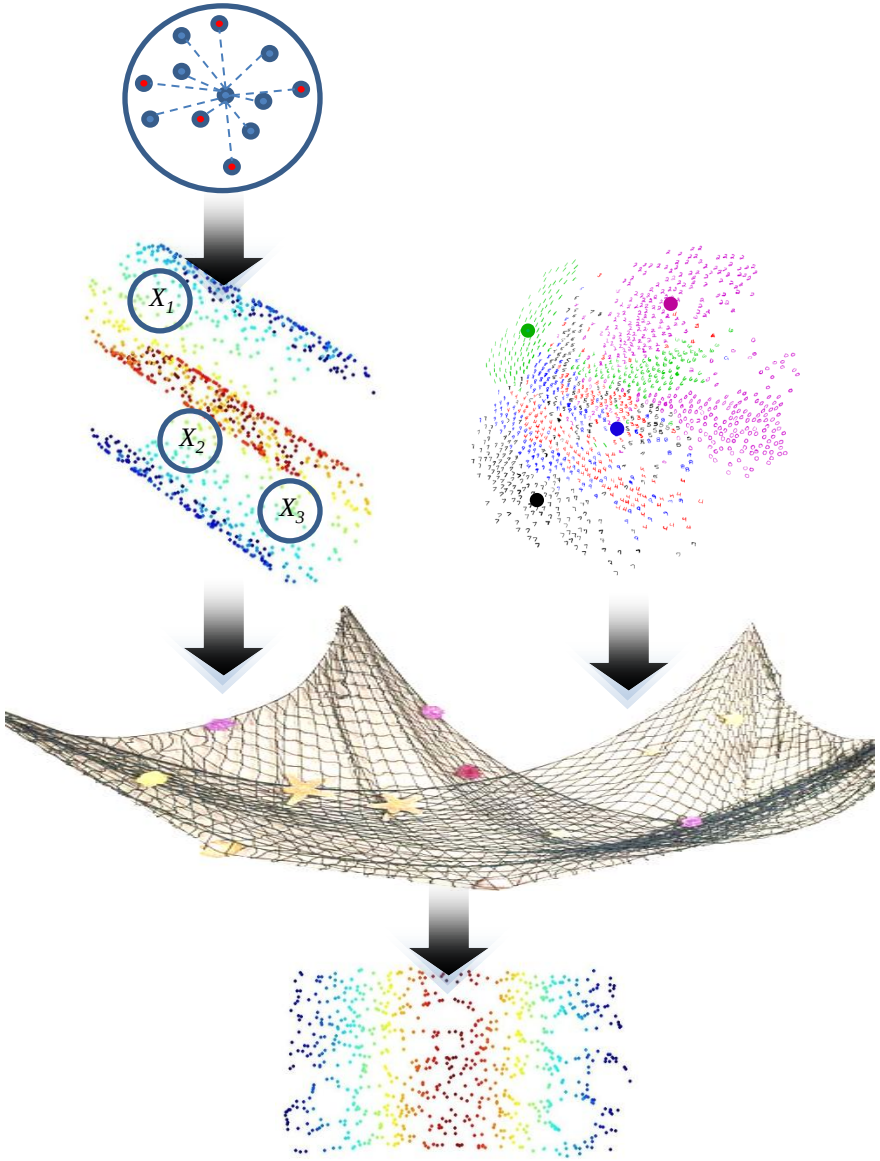
- Pick a variable most correlated with response and add it into model;
- Proceed in a direction along which the correlations of variables in model are equally increased;
- Stop when a variable out of model has the same correlation as the ones in the model.



Fast LARS can update the matrix inverse computation from the last iteration and thus decrease the time complexity from $O(p^3)$ to $O(p^2+5p)$.

Manifold elastic net

1. Calculate L_i in Part optimization;
2. Calculate alignment matrix L in Whole alignment;
3. Calculate indicator matrix Y in Classification error minimization;
4. Calculate X^*, Y^* in Manifold Elastic Net;
5. Run Fast LARS to obtain the sparse solution W;
6. Project data to subspace Z.



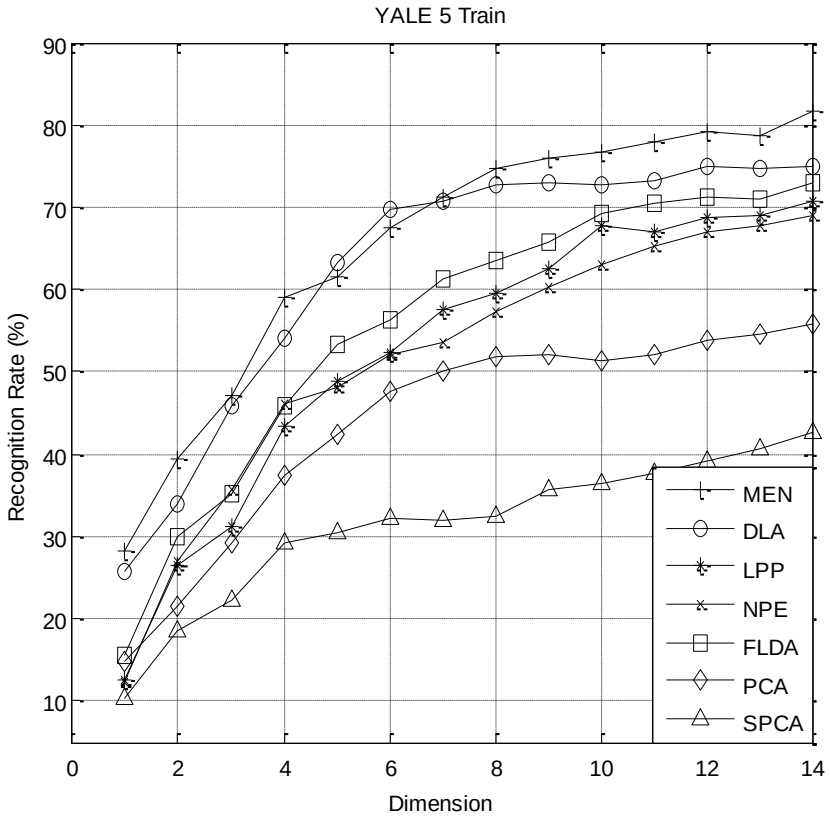
Experimental settings

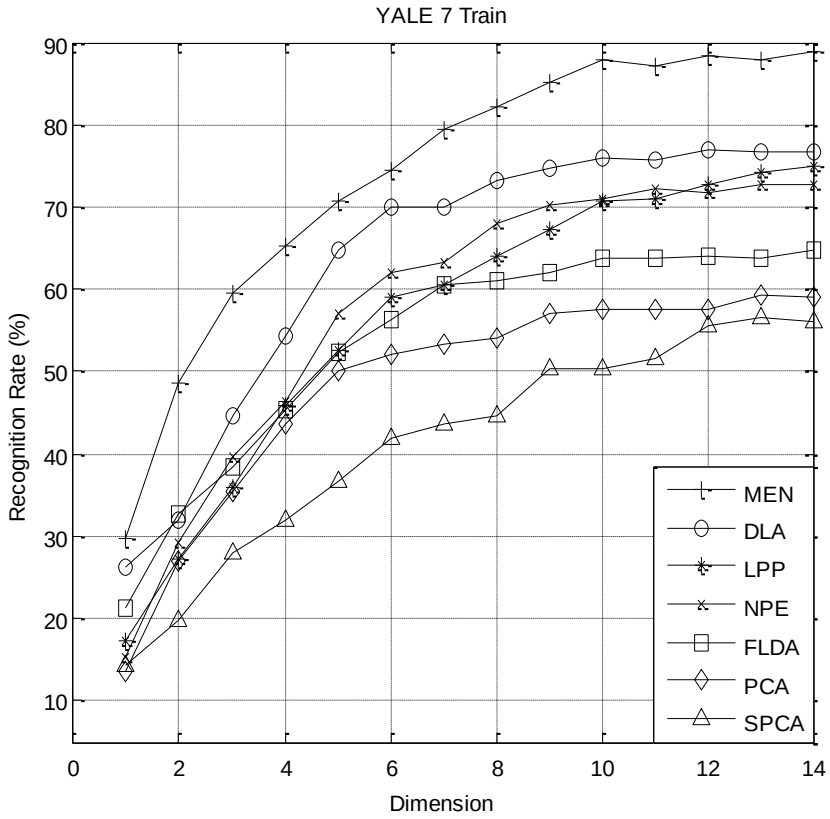
Datasets: FERET, 700 40X40 images, 100 individuals (7)
UMIST, 565 40X40 images, 20 individuals (28)
YALE, 165 40X40 images, 15 individuals (11)
Training: FERET(4,5), UMIST(5,7), YALE(5,7)

Me d

Set

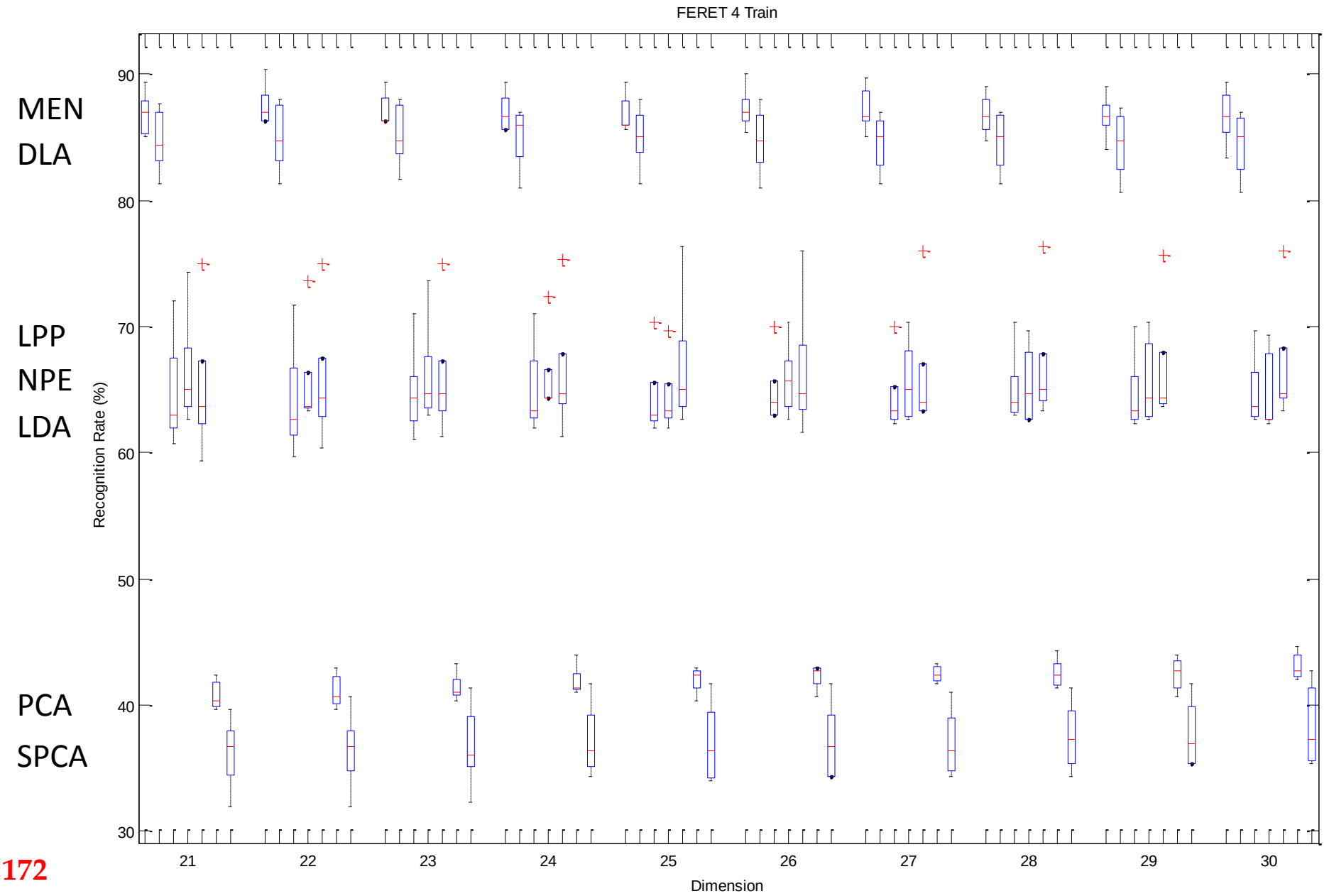
Manifold elastic net by fast LARS





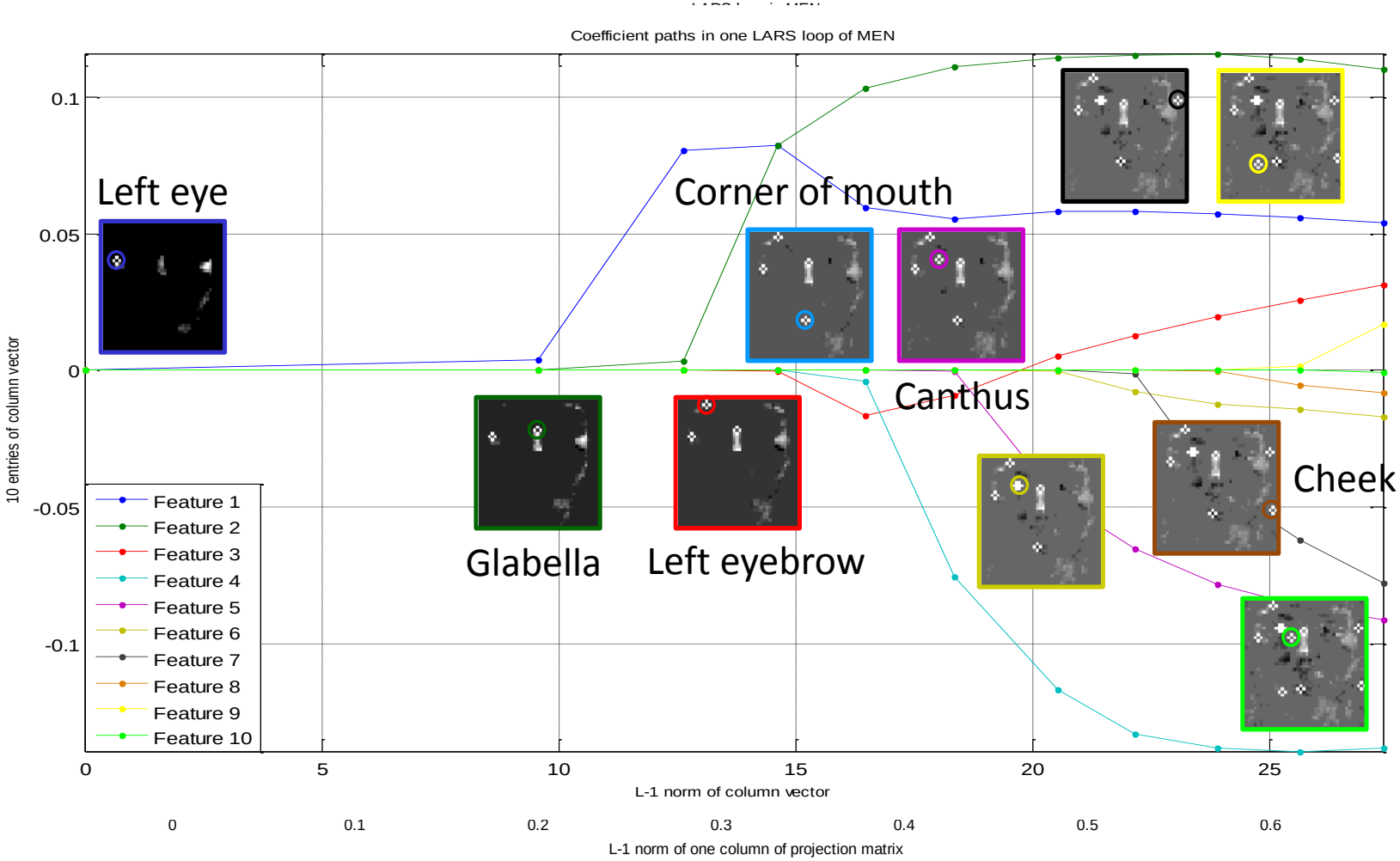
Sparse eigenfaces

Robustness on FERET (4)



Feature selection in MEN

Sorted feature set of Manifold Elastic Net



Transfer Patch Alignment Framework – Bregman Divergence-based Regularization

PART FOUR – 3

UTS:QCIS
QUANTUM COMPUTATION & INTELLIGENT SYSTEMS

Cross-domain image annotation

Cougar

Zebra

Tiger

Elk

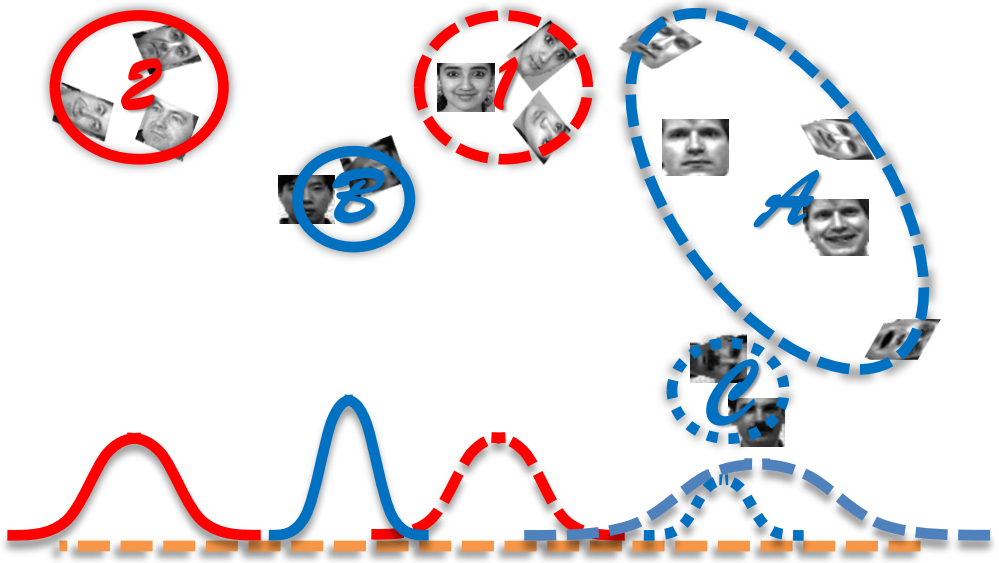
Transfer Dimension Reduction

Training and test samples may be not i.i.d

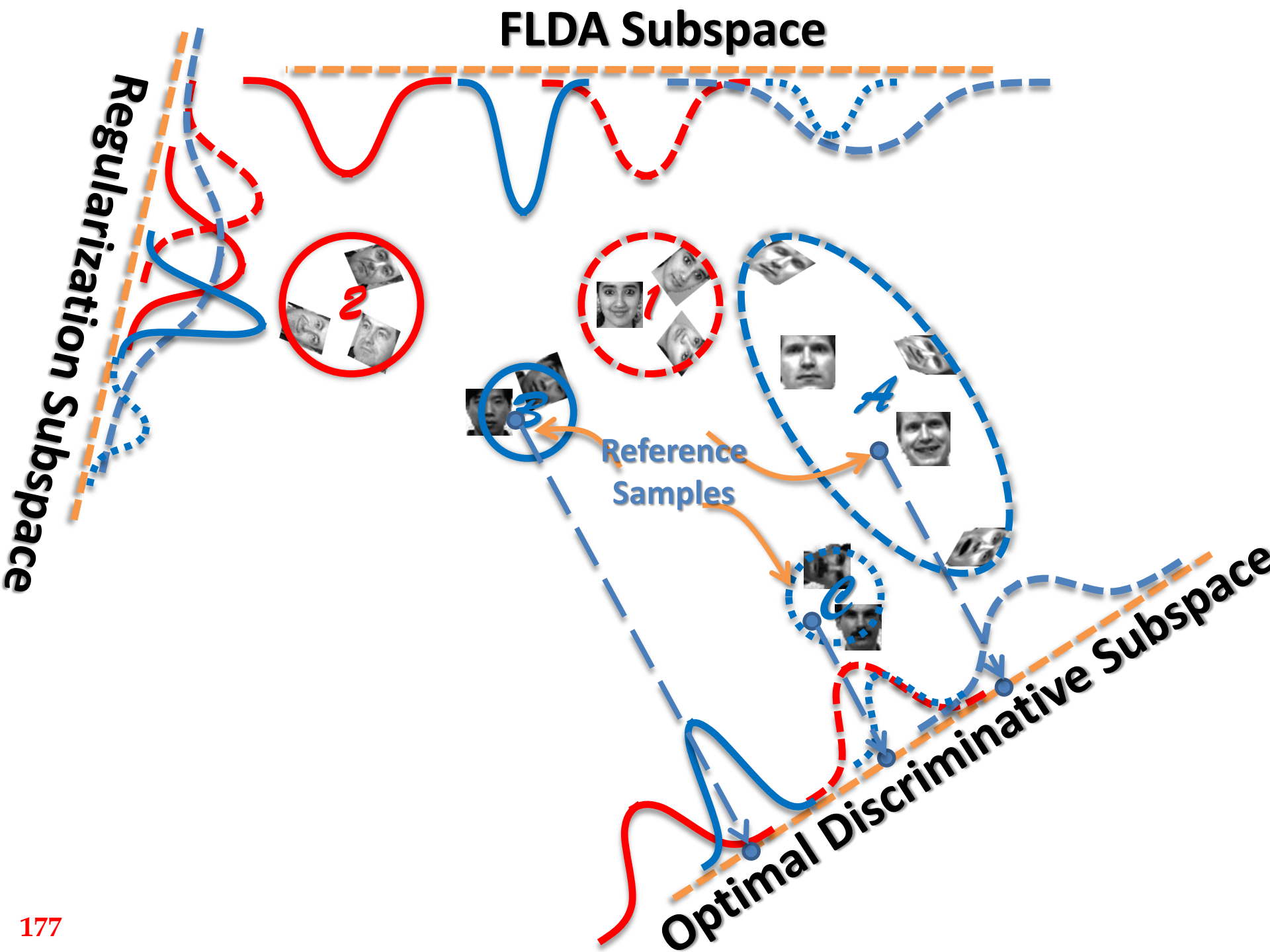
Transfer learning - transfer information from auxiliary tasks to target tasks

Cross domain problems

- text categorization
- image annotation
- sign language processing
- ⋮



Can we learn a subspace, in which samples are approximately i.i.d ?


FLDA Subspace
Regularization Subspace
1
2
3
4
5
Reference
Samples
Optimal Discriminative Subspace

TSL - regularized patch alignment

Objective function in the patch alignment framework

$$U = \arg\min_{U} F(U) + \lambda D_U(P_l \| P_u), \quad U^T U = I$$

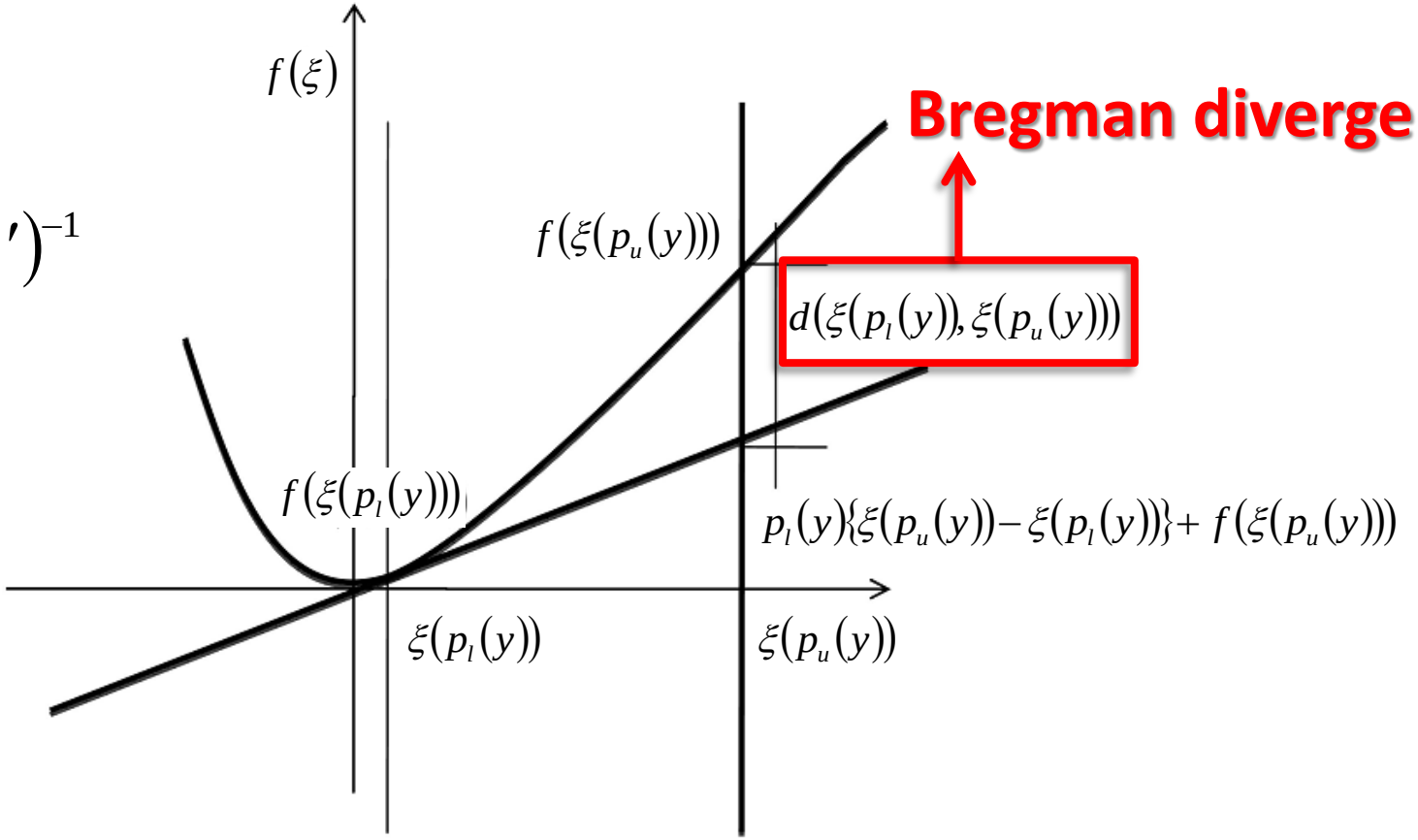
Bregman divergence based regularization

P_l : probability density of the training samples

P_u : Probability density of the test samples

Bregman divergence

$$\xi = (f')^{-1}$$



$$D_U(P_l \parallel P_u) = \int d(\xi(p_l(y)), \xi(p_u(y)))d\mu$$

Implementation

A special form $f(y) = y^2$ quadratic divergence

$$D_U(P_l \| P_u) = \int (p_l(y) - p_u(y))^2 dy$$

estimate using KDE

Kernel density estimation

$$p(y_i) = \frac{1}{N} \sum_{j=1}^{N} G_\Sigma(y_i - y_j)$$

$G_\Sigma(y)$: Gaussian kernel with covariance Σ

Optimization

Optimize using gradient descent

$$U_{k+1} = U_k - \eta_k \left(\frac{\partial F(U)}{\partial U} + \lambda \frac{\partial D_U(P_l \| P_u)}{\partial U} \right)$$

Algorithm specific

$$\sum_{i=1}^{N_l+N_u} \frac{\partial D_U(P_l \| P_u)}{\partial y_i} \frac{\partial y_i}{\partial U}$$

Examples of TSL

1. Transferred PCA

$$F(U) = -\mathrm{tr}(U^T R U)$$

covariance matrix

2. Transferred FLDA

$$F(U) = \mathrm{tr}^{-1}(U^T S_B U)\,\mathrm{tr}(U^T S_W U)$$

between-class scatter matrix

within-class scatter matrix

Examples of TSL

3. Transferred LPP

$$F(U) = 2\mathrm{tr}\left(U^T X (D - W) X^T U\right)$$

diagonal matrix with $D_{ii} = \sum_{j=1}^{N_l} W_{ij}$

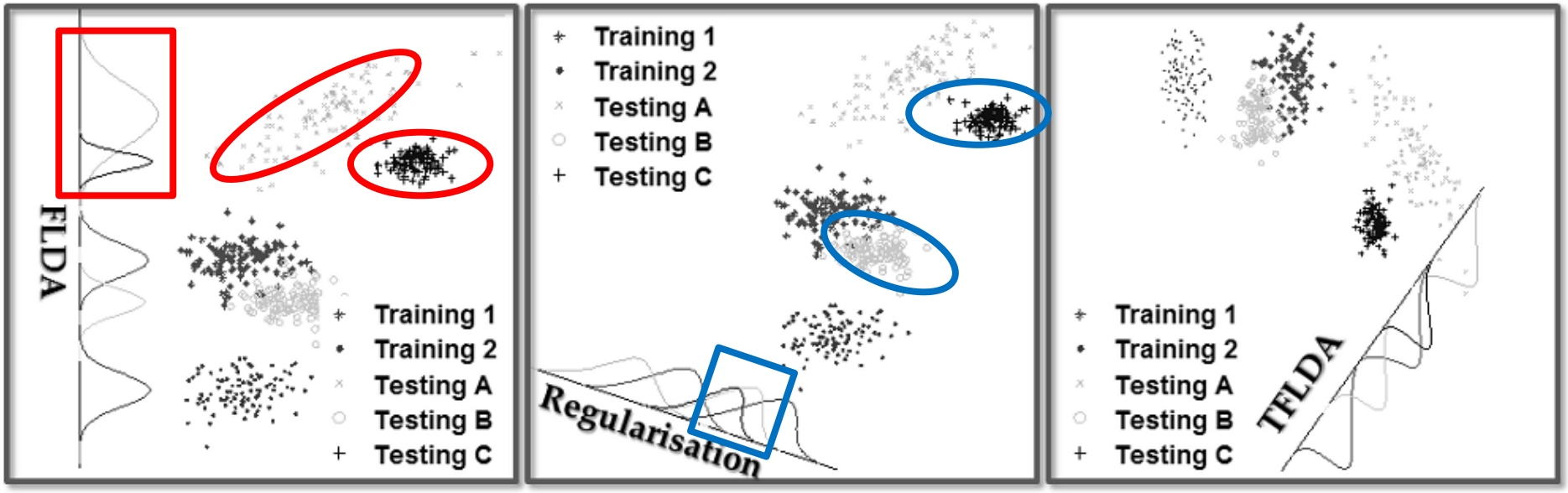
weighted graph

4. Transferred DLA

$$F(U) = \mathrm{tr}\left(U^T X L X^T U\right)$$

global optimization representation

Synthetic data experiment



test A, test C merged test B, test C merged

Cross-domain face datasets

FERET: 700 images, 100 individuals, a subset of FERET

UMIST: 564 images, 20 people

YALE: 165 images, 15 individuals

Cross-domain face reconstruction

Reconstruct a projected $y_i = U^T x_i$ sample as

$$\tilde{x}_i = U y_i = U U^T x_i$$

Reconstruction error

$$\| x_i - \tilde{x}_i \|_F$$

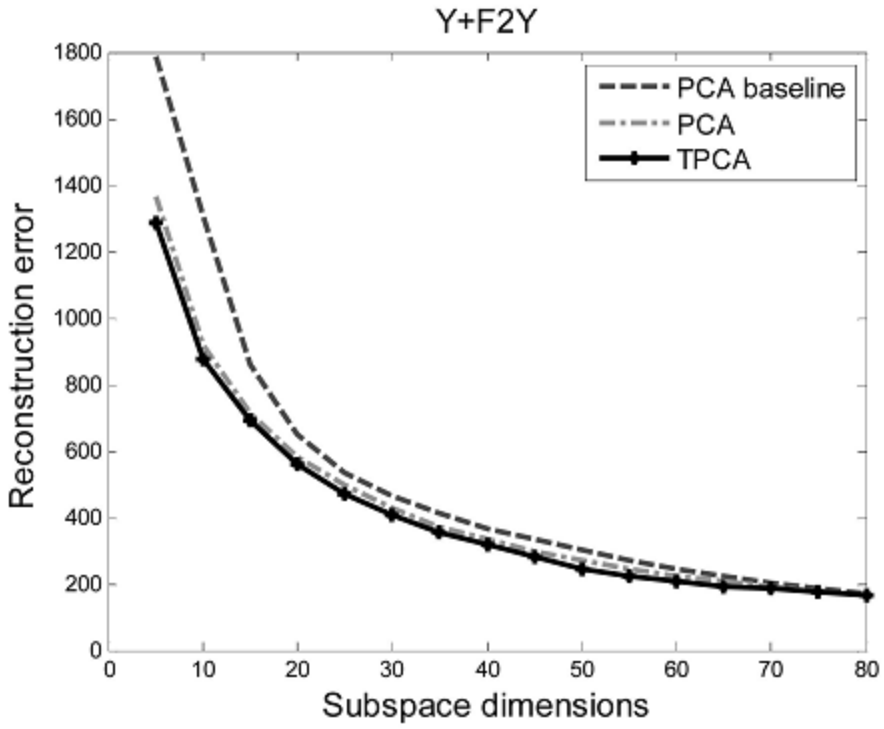
Experimental setting

Y+F2F: YALE+FERET for training, FERET for test

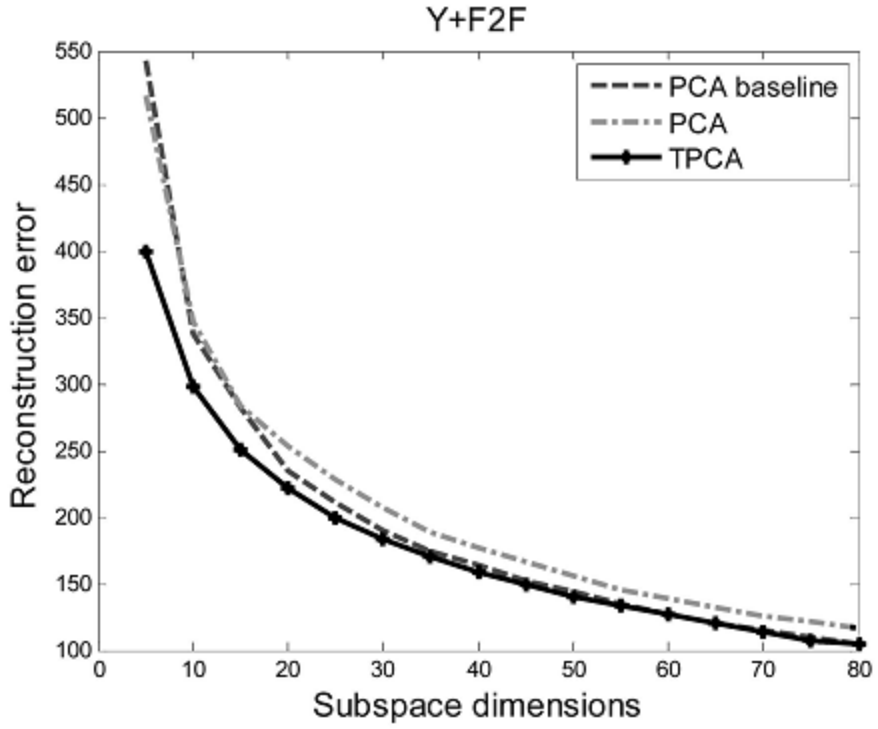
Y+F2Y: YALE+FERET for training, YALE for test

Reconstruction error

PCA baseline: both training and test are conducted on test set



(a)



(b)

TPCA outperforms others, esp. when using insufficient eigenvectors

(a): PCA baseline much worse than PCA and TPCA

(b): PCA baseline similar to TPCA using sufficient eigenvectors

Y+F2Y, reconstruction using 5-80 (step 5) eigenvectors

Y+F2F, reconstruction using 5-80 (step 5) eigenvectors

Cross-domain face recognition

Experimental settings

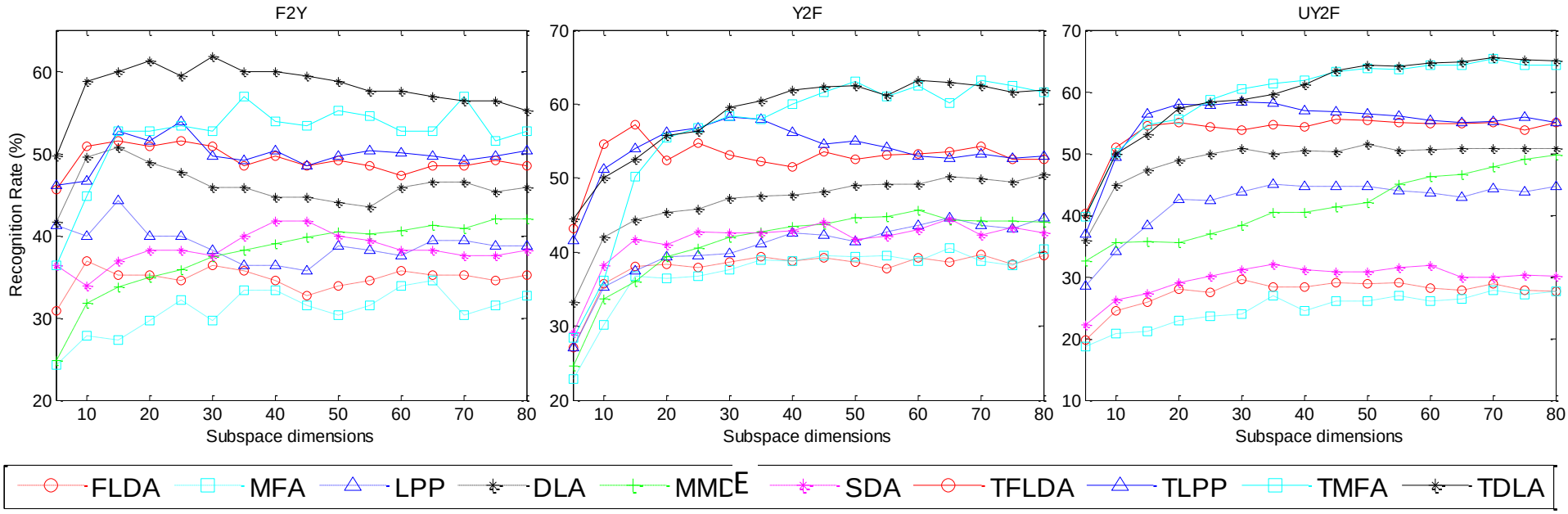
Y2F: YALE for training, FERET for test

F2Y: FERET for training, YALE for test

YU2F: YALE+UMIST for training, FERET for test

Nearest neighbor rule for classification

Recognition rates vs. dimensions



FLDA, LPP, MFA: sample i.i.d. assumption

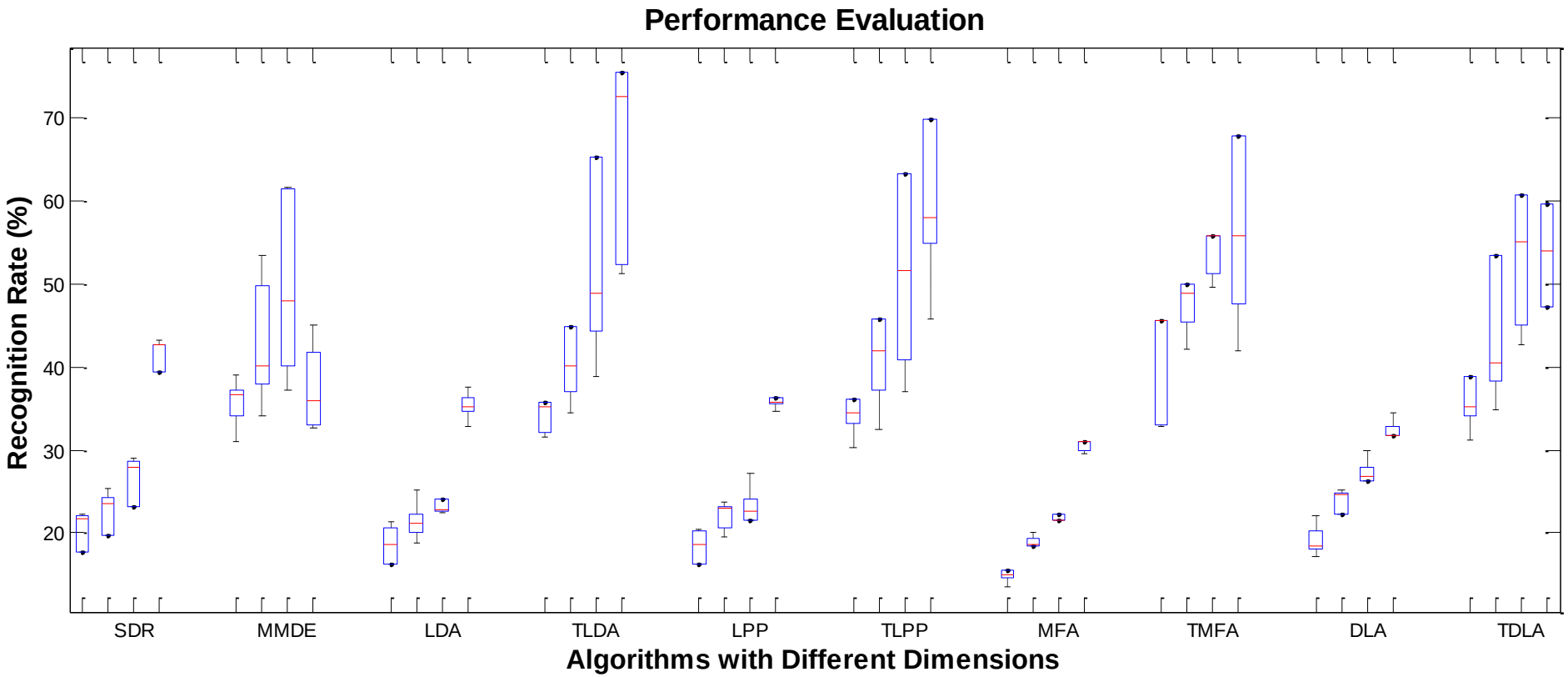
SDA: assume samples in a same class drawn from identical manifold

MMDE: ignore discriminative information

Best recognition rates

	Y2F	F2Y	YU2F
LDA	39.71 (70)	36.36 (30)	29.57 (30)
LPP	44.57 (65)	44.24 (15)	45.00 (35)
MFA	40.57 (65)	34.54 (60)	27.85 (70)
DLA	50.43 (80)	50.73 (15)	50.86 (65)
SDA	44.42 (65)	41.81 (40)	32.00 (35)
MMDE	45.60 (60)	42.00 (75)	49.75 (80)
TLDA	57.28 (15)	50.51 (20)	55.57 (45)
TLPP	58.28 (30)	53.93 (25)	58.42 (30)
TMFA	63.14 (70)	56.96 (35)	65.42 (70)
TDLA	63.12 (60)	61.82 (30)	65.57 (70)

Cross-domain text categorization

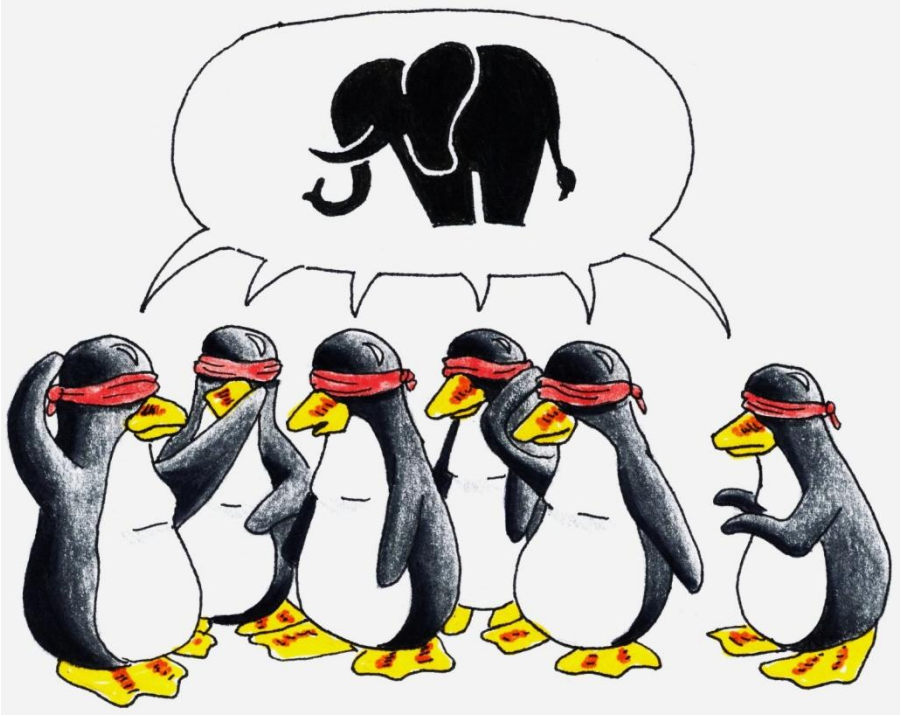


20Newsgroups

10 for training and the rest 10 for testing

Each class contains 50 documents

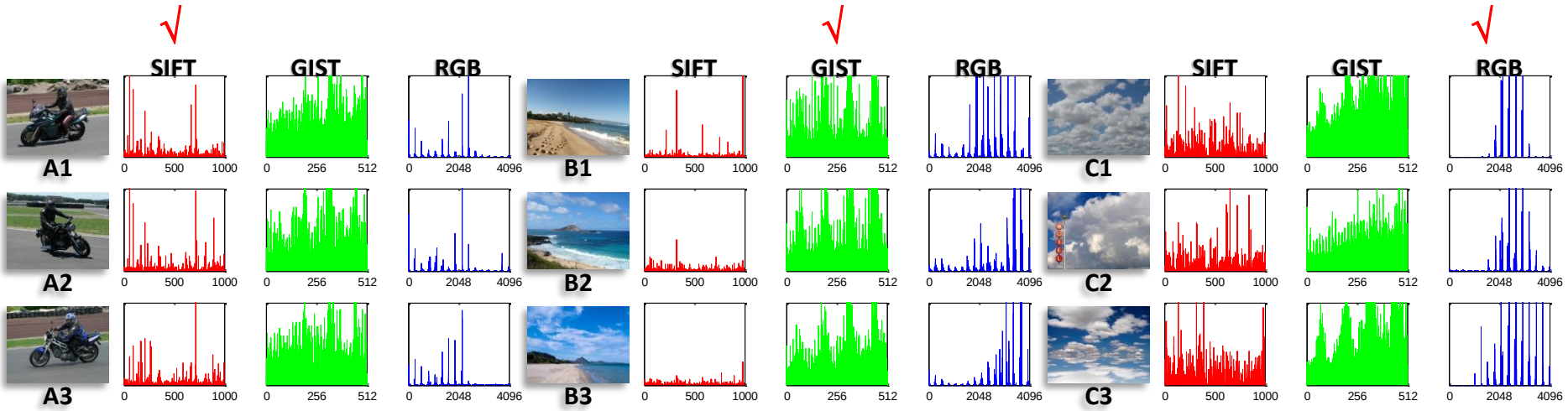
Evaluation: 5, 10, 15, 20 dimensions



Multiview Patch Align Framework – Integrating Multiple Features for Performance Improvement

PART FOUR – 4

UTS:QCIS
QUANTUM COMPUTATION & INTELLIGENT SYSTEMS



	A1	A2	A3
A1	0	0.22	0.21
A2	0.22	0	0.25
A3	0.21	0.25	0

(a)

	A1	A2	A3
A1	0	0.28	0.26
A2	0.26	0	0.25
A3	0.24	0.22	0

(b)

	A1	A2	A3
A1	0	0.41	0.35
A2	0.41	0	0.38
A3	0.35	0.38	0

(c)

	B1	B2	B3
B1	0	0.32	0.29
B2	0.32	0	0.19
B3	0.29	0.19	0

(d)

	B1	B2	B3
B1	0	0.17	0.20
B2	0.17	0	0.15
B3	0.20	0.15	0

(e)

	B1	B2	B3
B1	0	0.43	0.46
B2	0.43	0	0.30
B3	0.46	0.30	0

(f)

	C1	C2	C3
C1	0	0.34	0.37
C2	0.34	0	0.45
C3	0.37	0.45	0

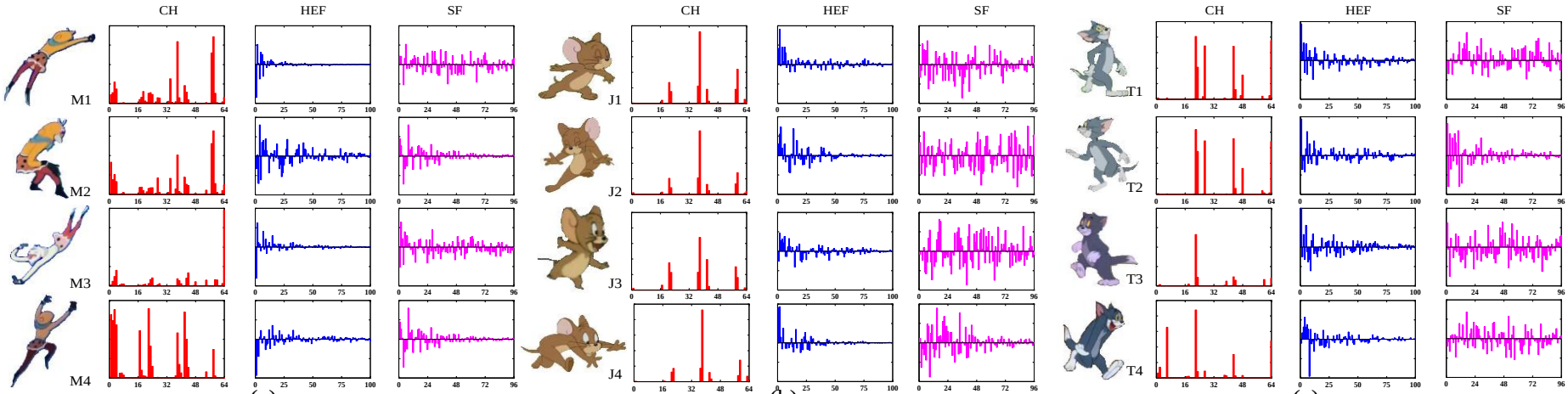
(g)

	C1	C2	C3
C1	0	0.33	0.22
C2	0.33	0	0.32
C3	0.22	0.32	0

(h)

	C1	C2	C3
C1	0	0.24	0.32
C2	0.24	0	0.29
C3	0.32	0.29	0

(i)



(a) (b) (c)

	M1	M2	M3	M4
M1	0	0.05	0.56	0.50
M2	0.05	0	0.61	0.48
M3	0.56	0.61	0	0.58
M4	0.50	0.48	0.58	0

Distance on CH (d)

	M1	M2	M3	M4
M1	0	0.47	0.09	0.29
M2	0.47	0	0.41	0.37
M3	0.09	0.41	0	0.34
M4	0.29	0.37	0.34	0

Distance on HEF (e)

	M1	M2	M3	M4
M1	0	0.41	0.18	0.13
M2	0.41	0	0.62	0.53
M3	0.18	0.62	0	0.21
M4	0.13	0.53	0.21	0

Distance on SF (f)

	J1	J2	J3	J4
J1	0	0.04	0.50	0.25
J2	0.04	0	0.48	0.28
J3	0.50	0.48	0	0.63
J4	0.25	0.28	0.63	0

Distance on CH (g)

	J1	J2	J3	J4
J1	0	0.27	0.13	0.25
J2	0.27	0	0.21	0.34
J3	0.13	0.21	0	0.39
J4	0.25	0.34	0.39	0

Distance on HEF (h)

	J1	J2	J3	J4
J1	0	0.18	0.23	0.28
J2	0.18	0	0.17	0.15
J3	0.23	0.17	0	0.30
J4	0.28	0.15	0.30	0

Distance on SF (i)

	T1	T2	T3	T4
T1	0	0.11	0.86	0.96
T2	0.11	0	0.80	0.91
T3	0.86	0.80	0	0.98
T4	0.96	0.91	0.98	0

Distance on CH (j)

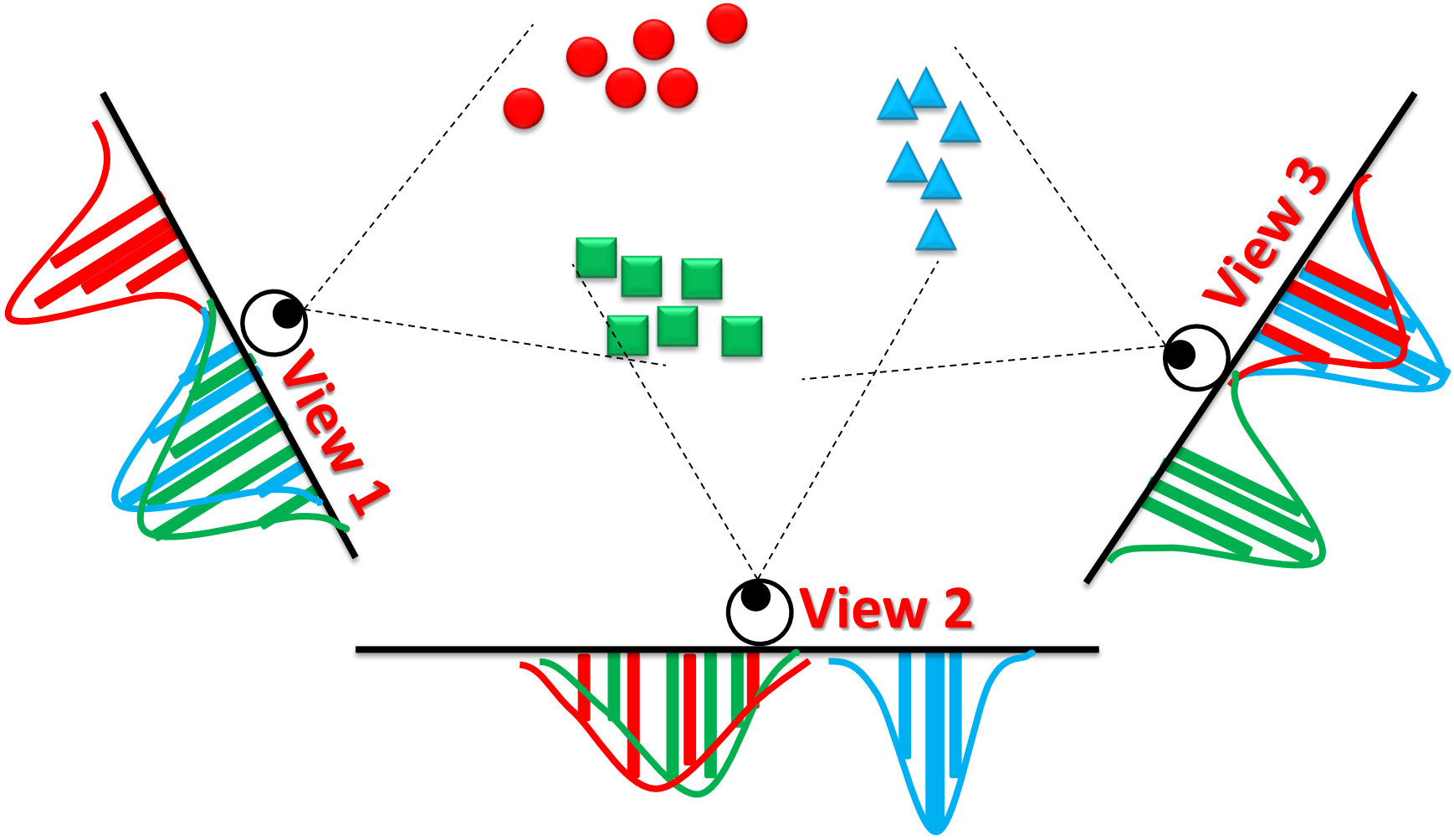
	T1	T2	T3	T4
T1	0	0.20	0.24	0.36
T2	0.20	0	0.18	0.37
T3	0.24	0.18	0	0.35
T4	0.36	0.37	0.35	0

Distance on HEF (k)

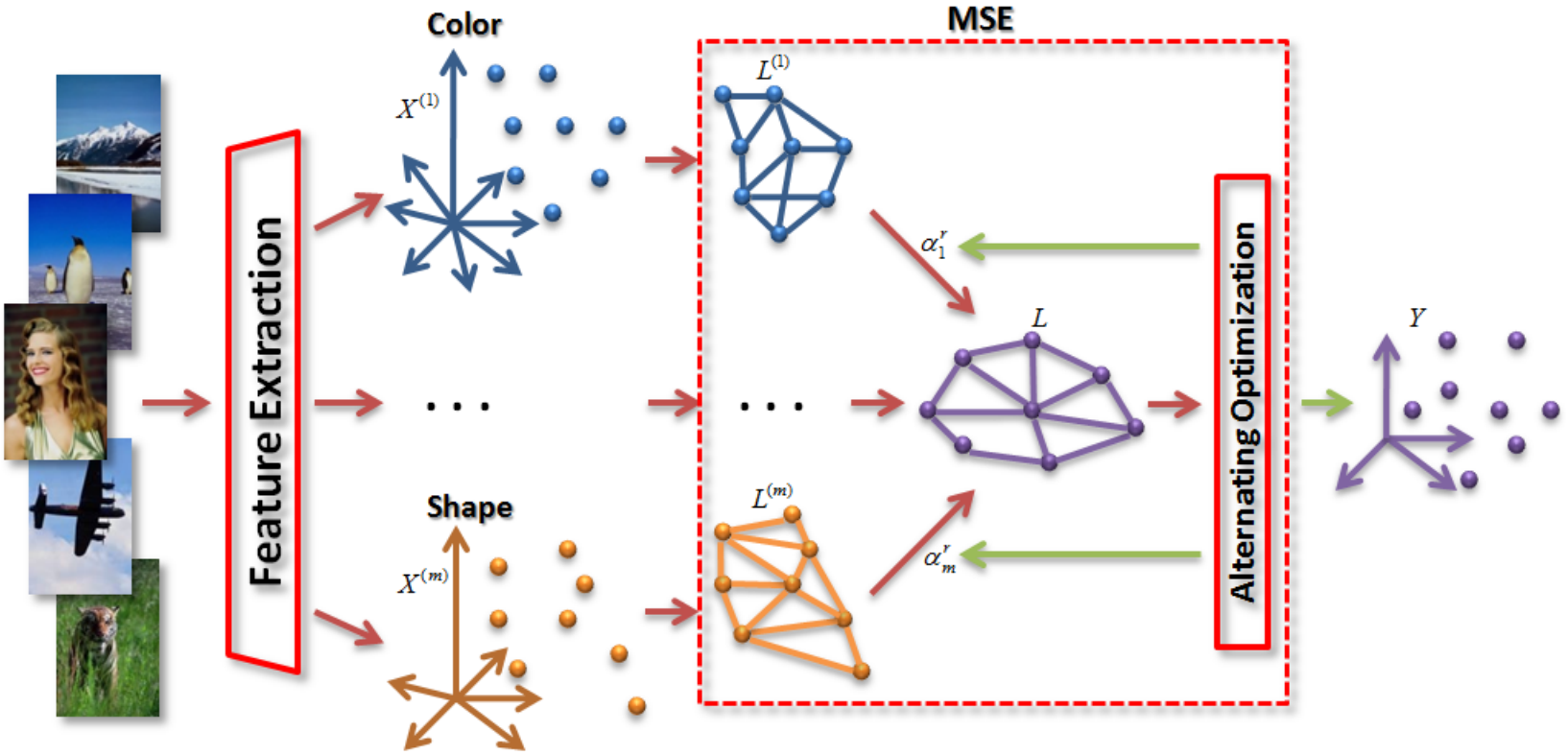
	T1	T2	T3	T4
T1	0	0.67	0.22	0.10
T2	0.67	0	0.65	0.67
T3	0.22	0.65	0	0.21
T4	0.10	0.67	0.21	0

Distance on SF (l)

Patch (1/5)



Framework


Color
MSE
Feature Extraction
Shape
Alternating Optimization
$X^{(1)}$
$X^{(m)}$
$L^{(1)}$
$L^{(m)}$
L
Y
α_1^r
α_m^r

Relation

$$\left.\begin{aligned}
&\arg\min_{Y_i^1} tr(Y_i^1 L_i^1 (Y_i^1)^T) \\
&\arg\min_{Y_i^2} tr(Y_i^2 L_i^2 (Y_i^2)^T) \\
&\quad\vdots \\
&\arg\min_{Y_i^V} tr(Y_i^V L_i^V (Y_i^V)^T)
\end{aligned}\right\}$$ **Part optimization**

Multiview Spectral Embedding (MSE)

$$\arg\min_{Y,\alpha} \sum_{i=1}^{N}\sum_{v=1}^{V} \alpha_v tr(Y S_i^v L_i^v (S_i^v)^T Y^T)$$

$$= \arg\min_{Y,\alpha} \sum_{v=1}^{V} \alpha_v tr(Y \sum_{i=1}^{N} (S_i^v L_i^v (S_i^v)^T) Y^T)$$

$$= \arg\min_{Y,\alpha} \sum_{v=1}^{V} \alpha_v tr(Y L^v Y^T)$$

Formulation

Unnormalized graph Laplacian matrix

Normalized graph Laplacian matrix

$$L^v = D^v - W^v \longrightarrow \tilde{L}^v = I - (D^v)^{-1/2} W^v (D^v)^{-1/2}$$

$$\arg\min_{Y,\alpha} \sum_{v=1}^{V} \alpha_v tr(YL^vY^T) \longrightarrow \arg\min_{Y,\alpha} \sum_{v=1}^{V} \alpha_v tr(Y\tilde{L}^vY^T)$$

$$s.t. YY^T = I, \sum_{v=1}^{V} \alpha_v = 1, \alpha_v \geq 0$$

Combination Coefficients

$$\arg\min_{Y,\alpha} \sum_{v=1}^{V} \alpha_v tr(Y\tilde{L}^v Y^T)$$

$$\downarrow$$

$$\begin{cases} \alpha_k = 1, \ for \quad k = \arg\min_i tr(Y\tilde{L}^i Y^T) \\ \alpha_k = 0, \ for \quad others \end{cases}$$

?

Trick: $\alpha_v \leftarrow \alpha_v^s, s > 1 \longrightarrow \arg\min_{Y,\alpha} \sum_{v=1}^{V} \alpha_v^s tr(Y\tilde{L}^v Y^T)$!

Alternating optimization

$$\arg\min_{Y,\alpha} \sum_{v=1}^{V} \alpha_v^s tr(Y\tilde{L}^v Y^T) \qquad s.t. YY^T = I, \sum_{v=1}^{V} \alpha_v = 1, \alpha_v \geq 0$$

Fix *Y* update α

$$L(\alpha, \lambda) = \sum_{v=1}^{V} \alpha_v^s tr(Y\tilde{L}^v Y^T) - \lambda(\sum_{v=1}^{V} \alpha_v - 1)$$

Lagrange multiplier

$$\alpha_v = \frac{(1/tr(Y\tilde{L}^v Y^T))^{1/(s-1)}}{\sum_{v=1}^{V} (1/tr(Y\tilde{L}^v Y^T))^{1/(s-1)}}$$

Fix α update *Y*

$$\arg\min_{Y} Y\tilde{L}Y^T, s.t. YY^T = I$$

$$Y = U^T$$

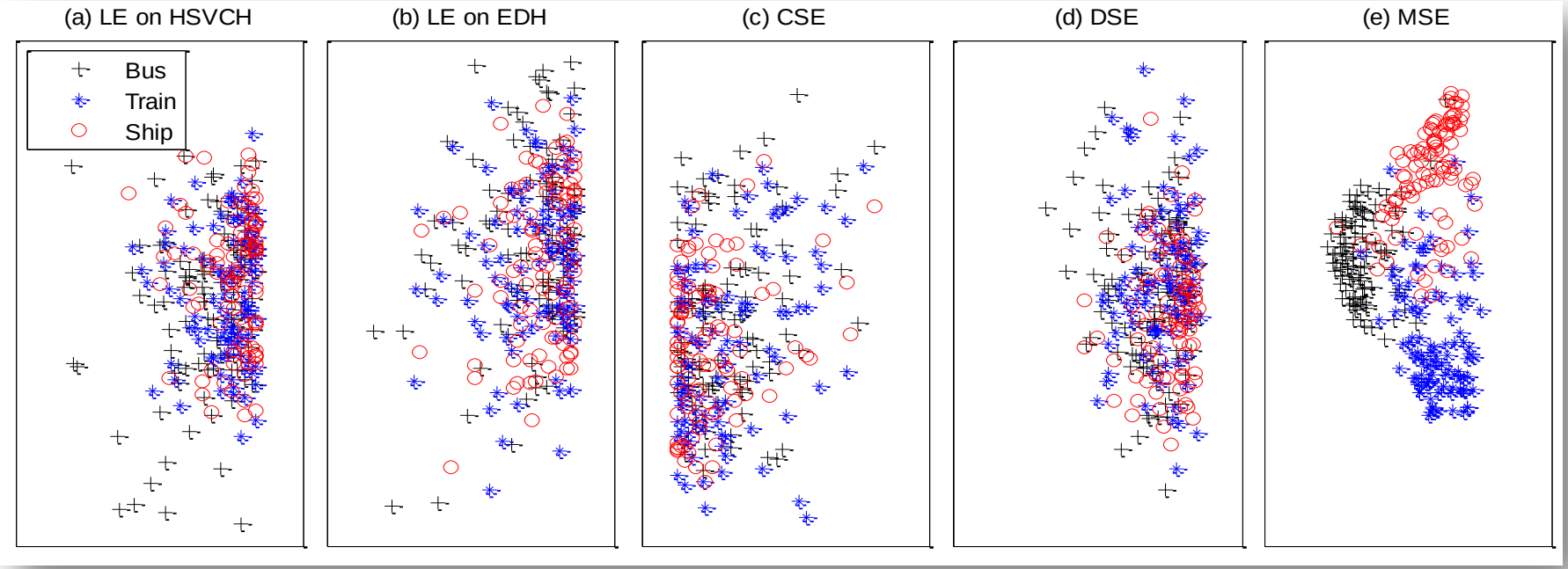
Ky-Fan theorem

$U = [u_1, \ldots, u_d]$ **the *d* eigenvectors corresponding to *d* smallest eigenvalues of $\tilde{L}$**

Exp. (1) toy dataset

A subset of Corel image gallery (each 100 images)

Exp. (1) MSE vs. CSE, DSE



2 Features:

HSVCH: HSV Colour Histogram

EDH: Edge Directional Histogram

4 Algorithms:

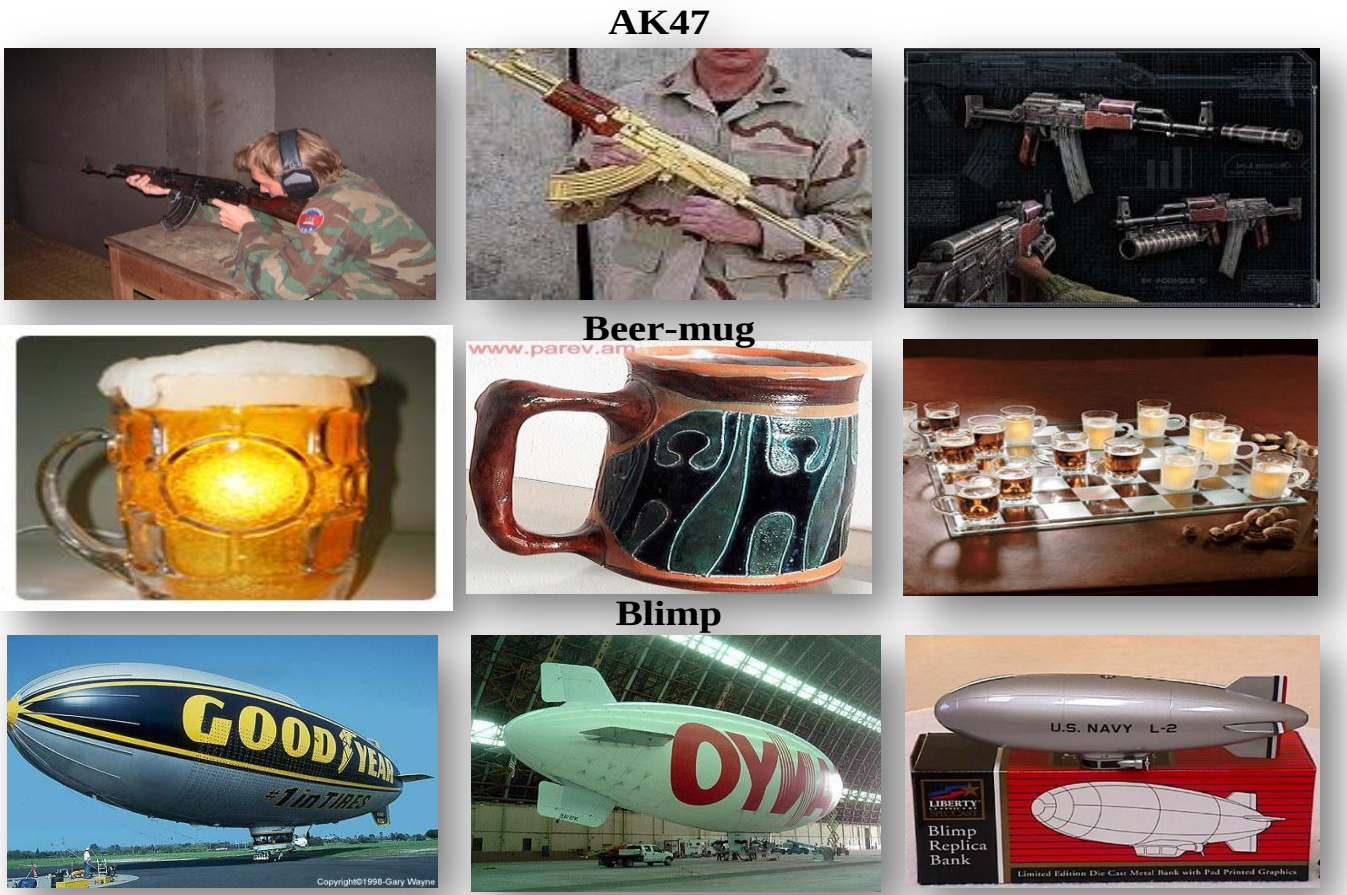
LE: Laplacian Eigenmaps

CSE: Concatenation based SE

DSE: Distributed SE

MSE: Multiview based SE

Exp. (2) Caltech256-2045

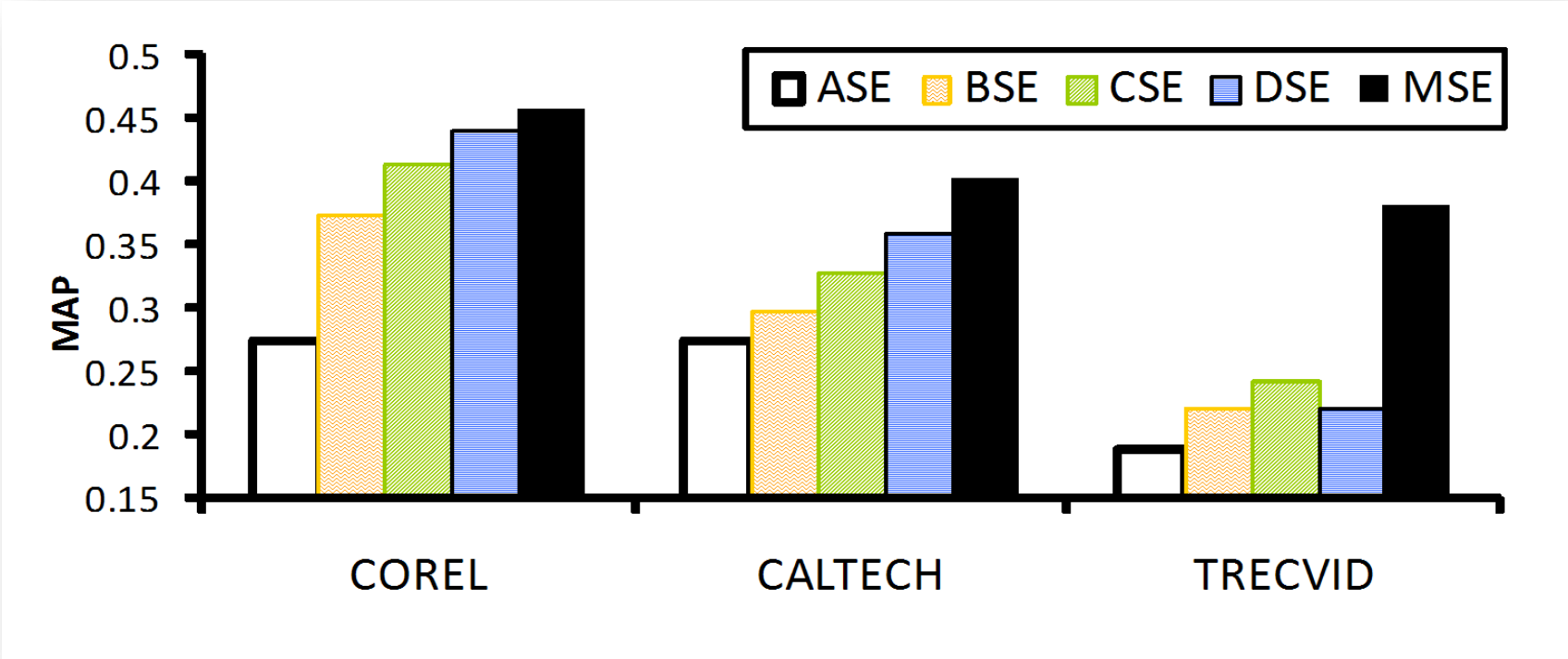


Caltech256-2045, 15 categories, 2045 imags

Exp. (2) TRECVID-2008

TRECVID-2008, 20 concepts, 39674 shots

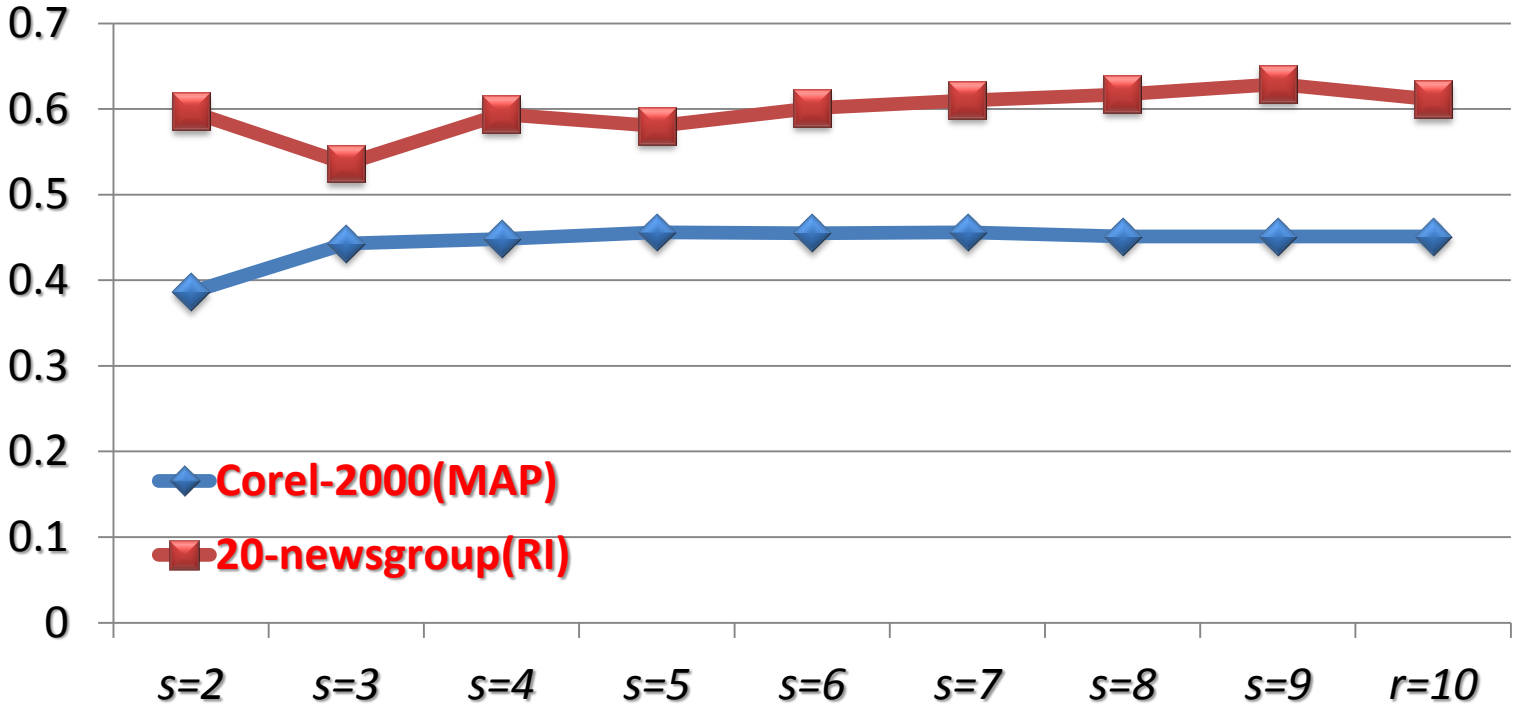
Exp. (2) Image Retrial and Video Annotation



5 Features :
colour moment, color correlogram, HSVCH, edge directional histogram, and wavelet texture

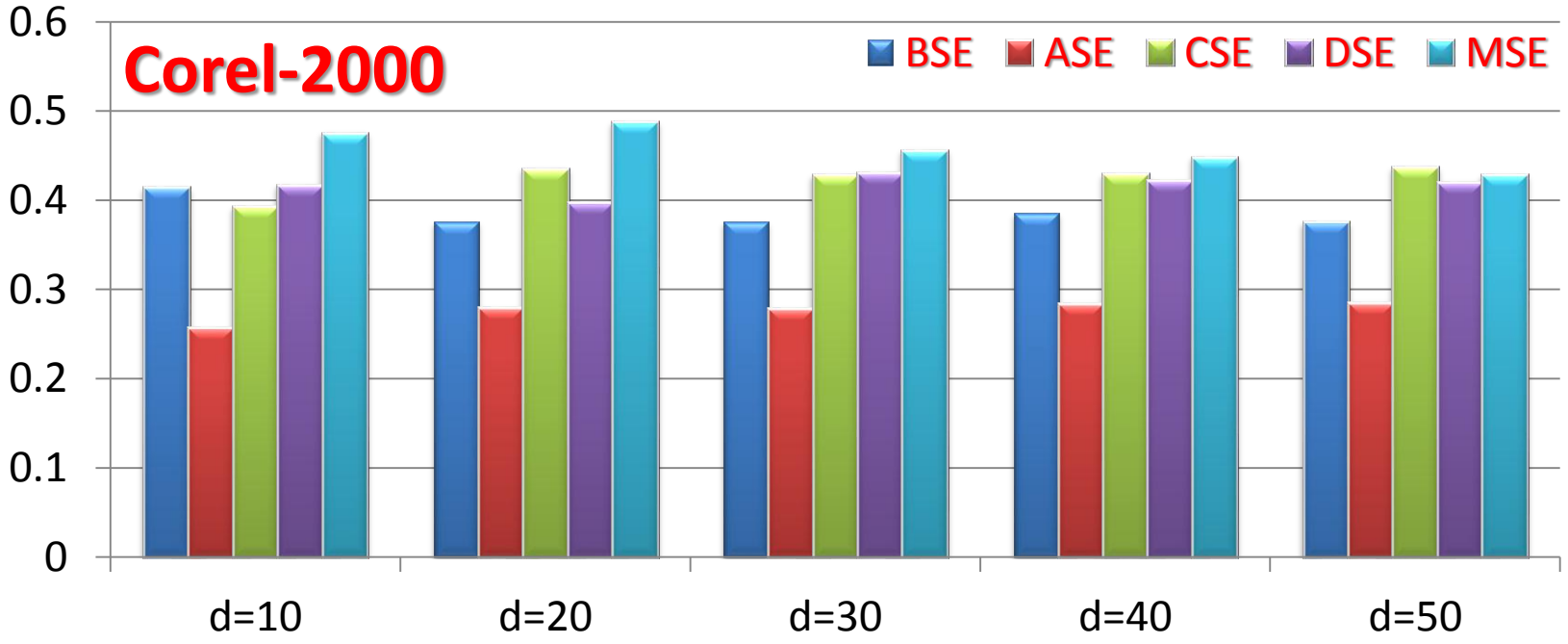
5 Methods:
ASE: Average performance of single-view-based SE
BSE: Best performance of single-view-based SE
CSE: Concatenation based SE
DSE: Distributed SE
MSE: Multiview based SE

Exp. (3) The Effect of Parameter *s*



RI	s=2	s=3	s=4	s=5	s=6	s=7	s=8	s=9	s=10
Corel-2000 (MAP)	0.386	0.443	0.448	0.456	0.455	0.456	0.451	0.451	0.451
20-newsgroup (RI)	0.597	0.536	0.594	0.58	0.601	0.61	0.617	0.629	0.611

Exp. (4) The Performance with *d*



RI	BSE	ASE	CSE	DSE	MSE
d=10	0.414	0.257	0.393	0.417	0.475
d=20	0.376	0.279	0.435	0.397	0.488
d=30	0.376	0.278	0.428	0.431	0.456
d=40	0.386	0.284	0.429	0.422	0.448
d=50	0.375	0.285	0.437	0.42	0.428

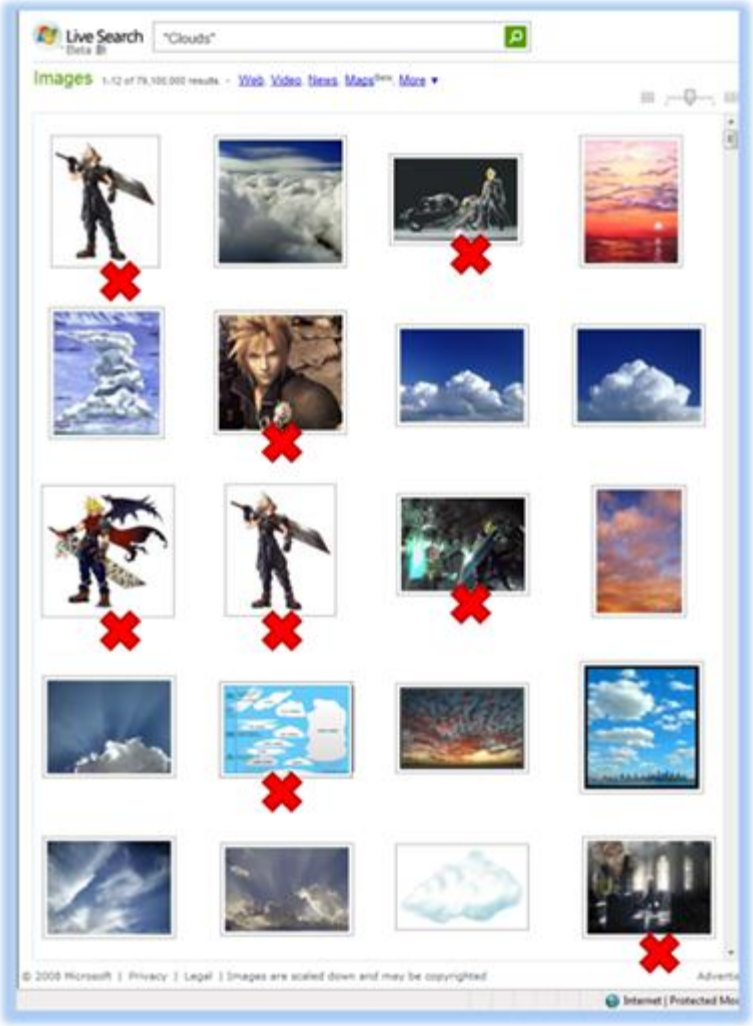
Active Patch Alignment Framework – Choosing Optimal Samples for Performance Improvement

PART FOUR – 5

UTS:QCIS
QUANTUM COMPUTATION & INTELLIGENT SYSTEMS

Background – Reranking

- Ranking
  - Given a query term, the deployed ranking model measures the relevance of each document to the query
  - sorts all images based on their relevance scores
  - presents a list of top-ranked ones to the user.

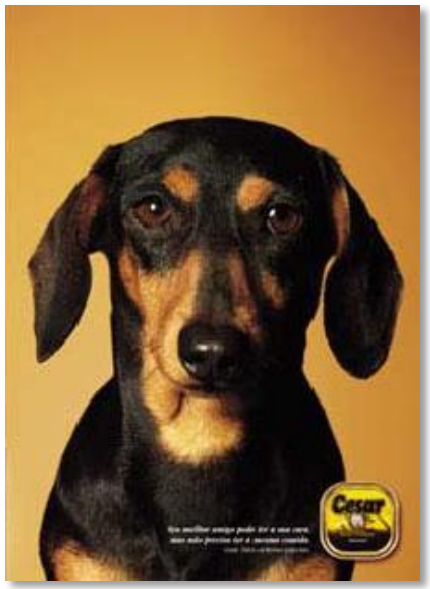


Cloud

Background – Reranking

- Reranking (unsupervised)
  - exploit the visual information for refining the text-based search result.
  - E.g. Bayesian Reranking, PageRank, Random Walk, Clustering
  - Pseudo Relevance Feedback in CBIR

Do visual features work?

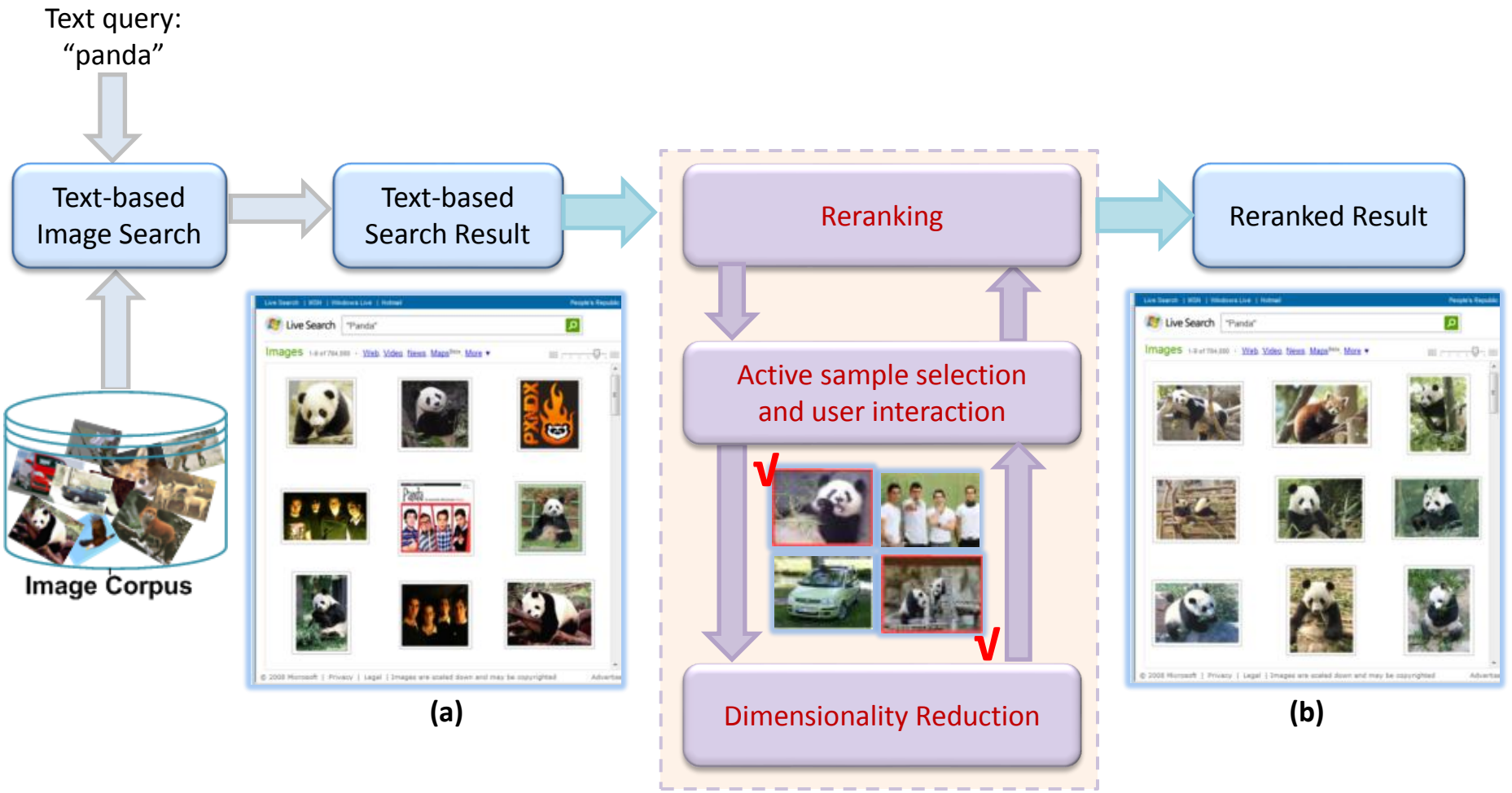


Background – Reranking

- Reranking problems
  - Cannot infer different user's different search intentions;
  - Especially for ambiguous queries, e.g., "Apple".
  - Generate same reranking results for all users;

Active reranking!

Framework


Text query:
“panda”
Text-based Image Search
Text-based Search Result
Reranking
Reranked Result
Active sample selection and user interaction
Dimensionality Reduction
Image Corpus
(a)
(b)

Two key problems

Active Sample Selection

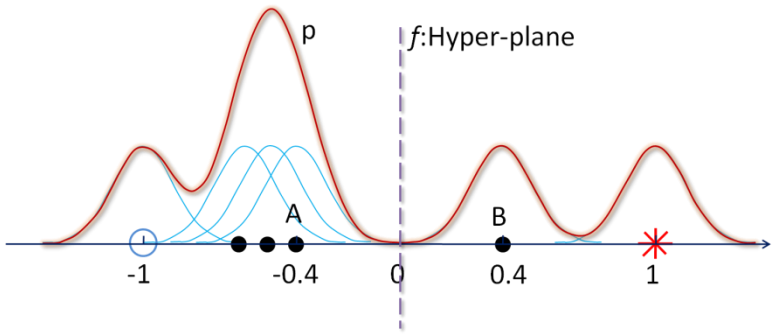
Aim:

Select the most informative samples

Algorithm:

SInfo: Structural information based sample selection

- Ambiguity
- Representativeness



Dimension Reduction

Aim:

Learning effective features to encode the user's search intentions

Algorithm:

LGD: Local and global discriminative dimension reduction

- Local patch
  learning local geometry and discriminative information
- Global patch
  transferring knowledge

SInfo: active sample selection

- Ambiguity denotes the uncertainty whether an image is relevant or not

$$H(I_i) = \alpha H_{\mathrm{r}}(I_i) + (1-\alpha) H_{\bar{\mathrm{r}}}(I_i)$$

where

$$H_{\mathrm{r}}(I_i) = -r_i \log r_i - (1-r_i)\log(1-r_i)$$

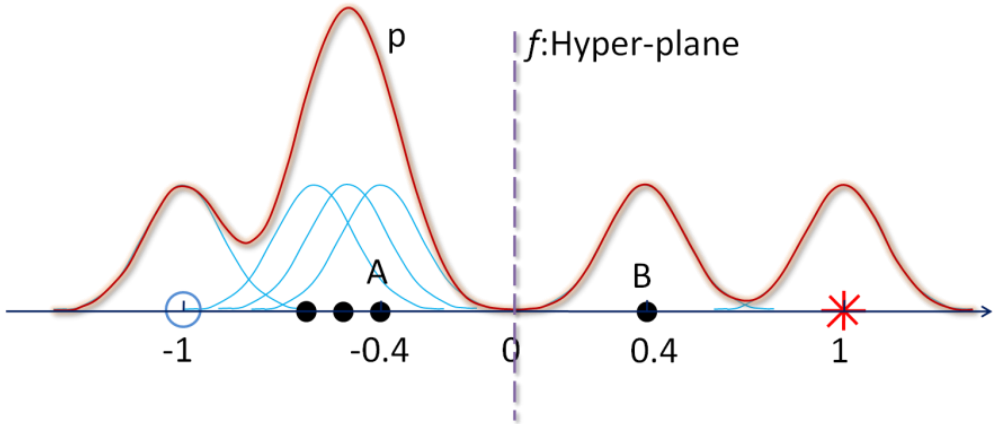
is ambiguity by the ranking scores

$$H_{\bar{\mathrm{r}}}(I_i) = -\bar{r}_i \log \bar{r}_i - (1-\bar{r}_i)\log(1-\bar{r}_i)$$

is ambiguity in the text-based search result

SInfo : active sample selection

- Representativeness denotes information shared by its neighbors.



- estimated by using KDE

$$p(I_i) = \frac{1}{|N_i|} \sum_{I_j \in N_i} k(x_i - x_j)$$

SInfo : active sample selection

- structural information of image is measured

$$\mathrm{SI}(I_i) = p(I_i)H(I_i)$$

- maintain the diversity by angle-diversity

$$I^* = \arg\max_{I_i \in U}(\eta \mathrm{SI}(I_i) + (1-\eta)\min_{I_j \in S} \frac{-x_i \cdot x_j}{\|x_i\|\|x_j\|})$$

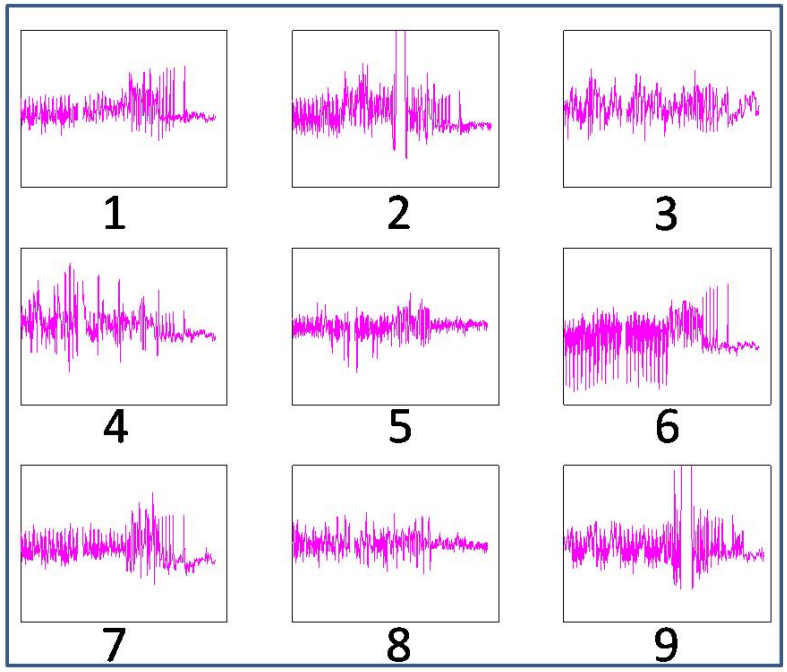
LGD: dimension reduction

- LGD considers both information:
  - the local information
    - local geometry of the labeled relevant images
    - the discriminative information in the labeled images
  - the global information of the whole database

LGD: dimension reduction

- Local Patches for Labeled relevant
  - query relevant samples may vary in appearance

(a) Query relevant images



(b) The visual features

Local Patches for Labeled Relevant

separate relevant image from all irrelevant ones

$$\min \operatorname{tr}\left(Y_i^+ L_i^+ \left(Y_i^+\right)^T\right) = \min \left\| y_i^+ - \sum_{j=1}^{k_1} (c_i)_j \, y_{i_j} \right\|^2 - \beta \sum_{j=k_1+1}^{k_1+k_2} \left\| y_i^+ - y_{i_j} \right\|^2$$

preserve the local geometry of relevant images

k_1 : nearest neighbors in the relevant set

k_2 : nearest neighbors in the irrelevant set

Local Patches for Labeled Irrelevant

labeled irrelevant image

$$\min tr\left(Y_i^- L_i^- \left(Y_i^-\right)^T\right) = \min -\sum_{j=1}^{k} \left\| y_i^- - y_{i_j} \right\|^2$$

The *k* nearest neighbors in the relevant image set

Global Patches for All Images

- Local patches use labeled data.
- Global patches exploit the information contained in both labeled and unlabeled data.
- Conventional manifold regularizations prone to over-fit to unlabelled samples.

Principle component analysis

$$\max tr\left(\left(y_i - y^m\right)\left(y_i - y^m\right)^T\right)$$

$$\max tr\left(Y_i L_i^{PCA} \left(Y_i\right)^T\right)$$

Patch Coordinate Alignment

Local Patches for Labeled Irrelevant Images

$$\max tr\left(U^T XLX^T U\right) = \max tr\left(YLY^T\right)$$

$$= \max tr\left(-Y_i^+ L_i^+ \left(Y_i^+\right)^T\right) + \max tr\left(-Y_i^- L_i^- \left(Y_i^-\right)^T\right) + \gamma \max tr\left(Y_i L_i^{PCA} \left(Y_i\right)^T\right)$$

Global Patches for All Images

Local Patches for Labeled Relevant Images

Overall Procedure

1: Initialization: the image set Γ, the number of interaction rounds T, labeled image set $S = \phi$ and $Y = X$.

2: /*Perform Bayesian reranking to get r */

$r \leftarrow$ Bayesian reranking $\{S, Y\}$

3: For *t*=1 to T do

1) /*Perform Sinfo to select a set of image S_t* /

$S_t^* \leftarrow SInfo\{r, Y\}, S_t \subseteq (I - S)$

/*Update S */

$S \leftarrow S \cup S_t$

2) /*Perform LGD to learn a new Y */

$Y \leftarrow LGD\{S\}$

3) /*Perform Bayesian reranking to derive a new r*/

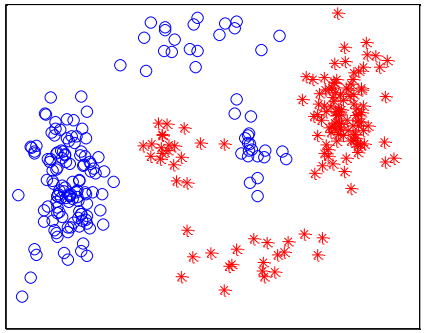
$r \leftarrow$ Bayesian reranking $\{S, Y\}$

4: End for

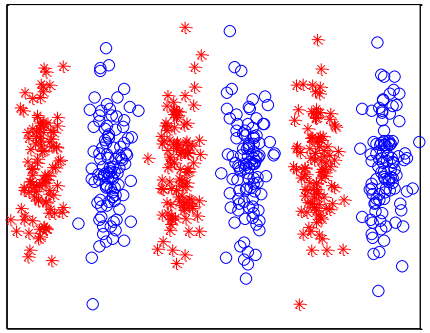
5: Return r

Experiments on Synthetic Datasets

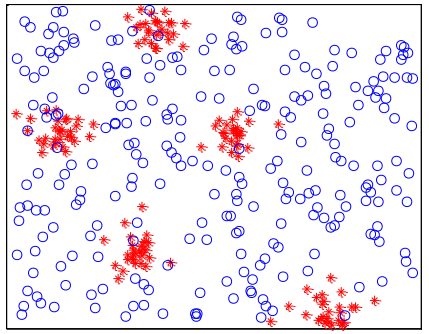
- Active Sample Selection
  - SInfo vs. Error Reduction, Most Uncertain and Random



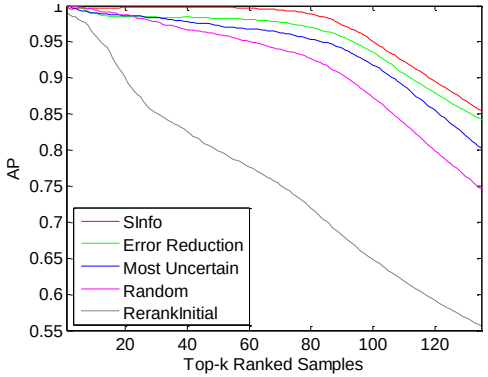
Synthetic data 1



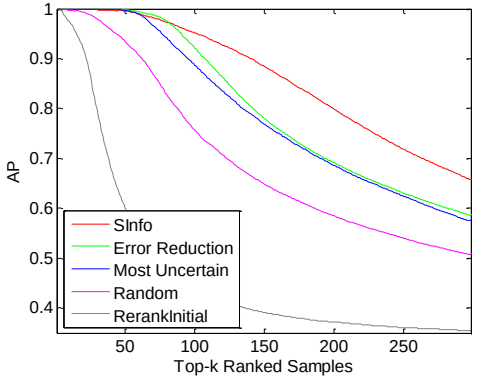
Synthetic data 2



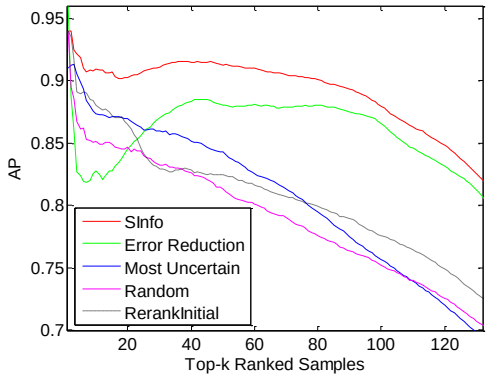
Synthetic data 3



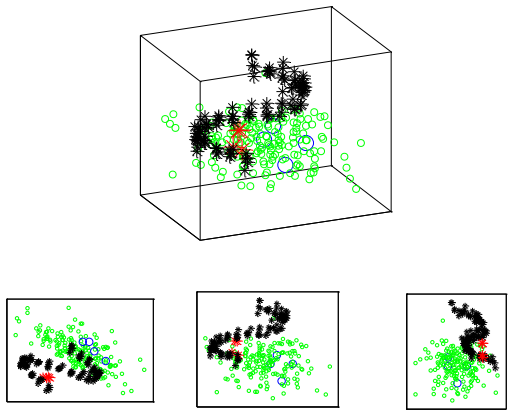
AP on synthetic data 1



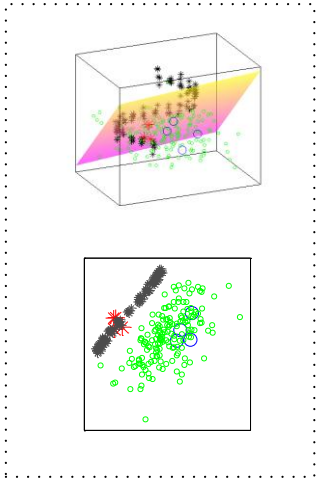
AP on synthetic data 2



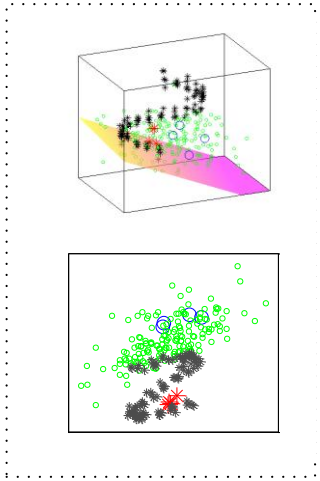
AP on synthetic data 3



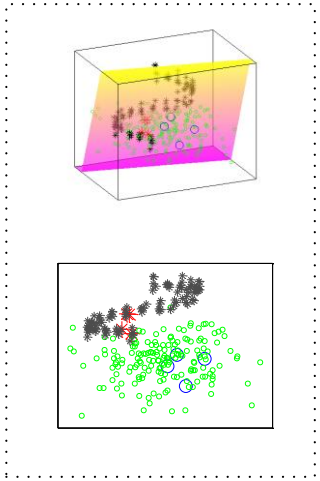
(a)



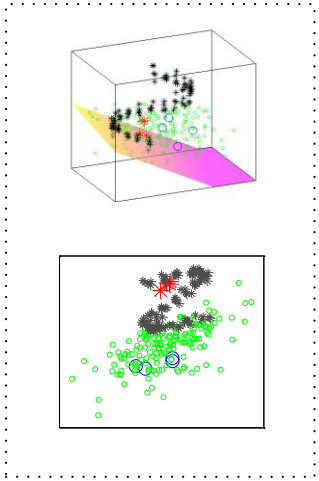
(b) LGD



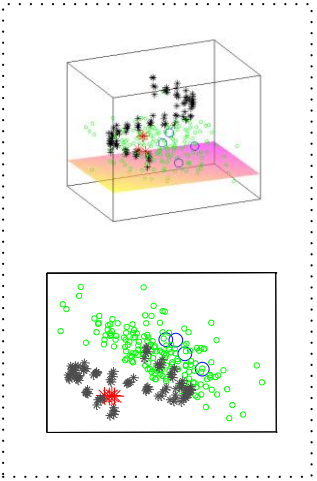
(c) L-patch



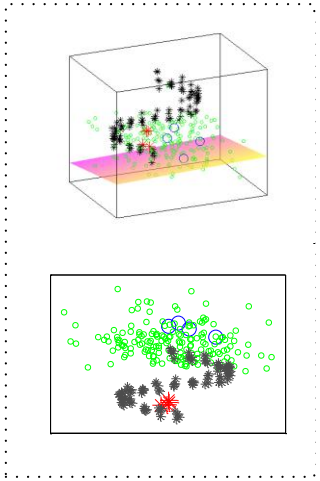
(d) G-patch



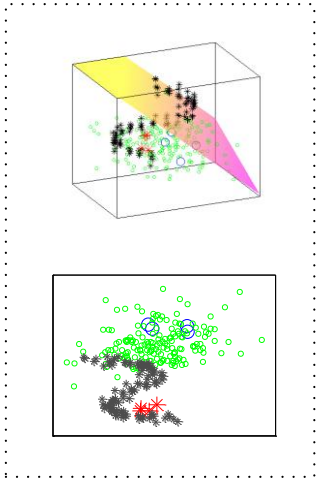
(e)LGD-LE



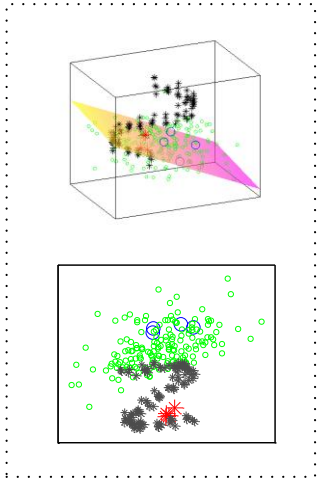
(f) BDA



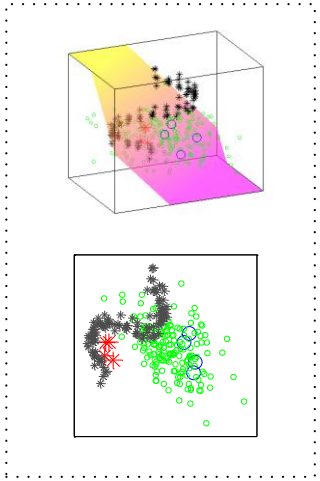
(g) BMFA



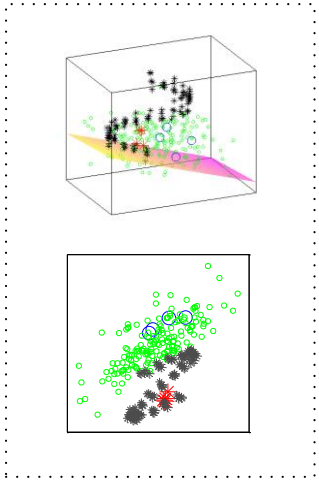
(h) LDE



(i) SLPP



(j) SDA

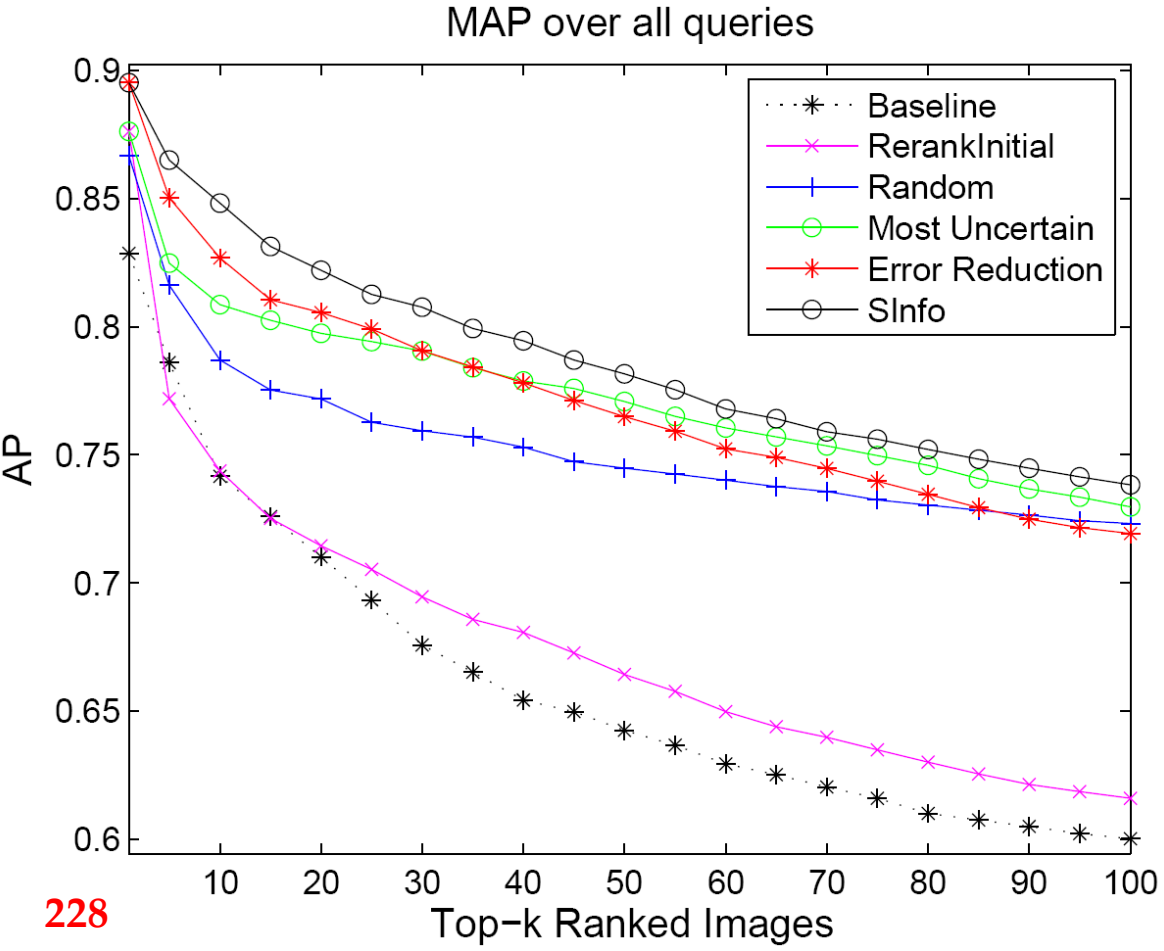


(k) SML

(a) The 3-D synthetic data and its 2-D projections (b) LGD, (c) Local patches, (d) Global patches, (e) LGD-LE, (f) BDA, (g) BMFA, (h) LDE, (i) SLPP, (j) SDA, (k) SML

Experiments on Real Search Dataset

- Active Sample Selection
  - SInfo vs. Error Reduction, Most Uncertain and Random



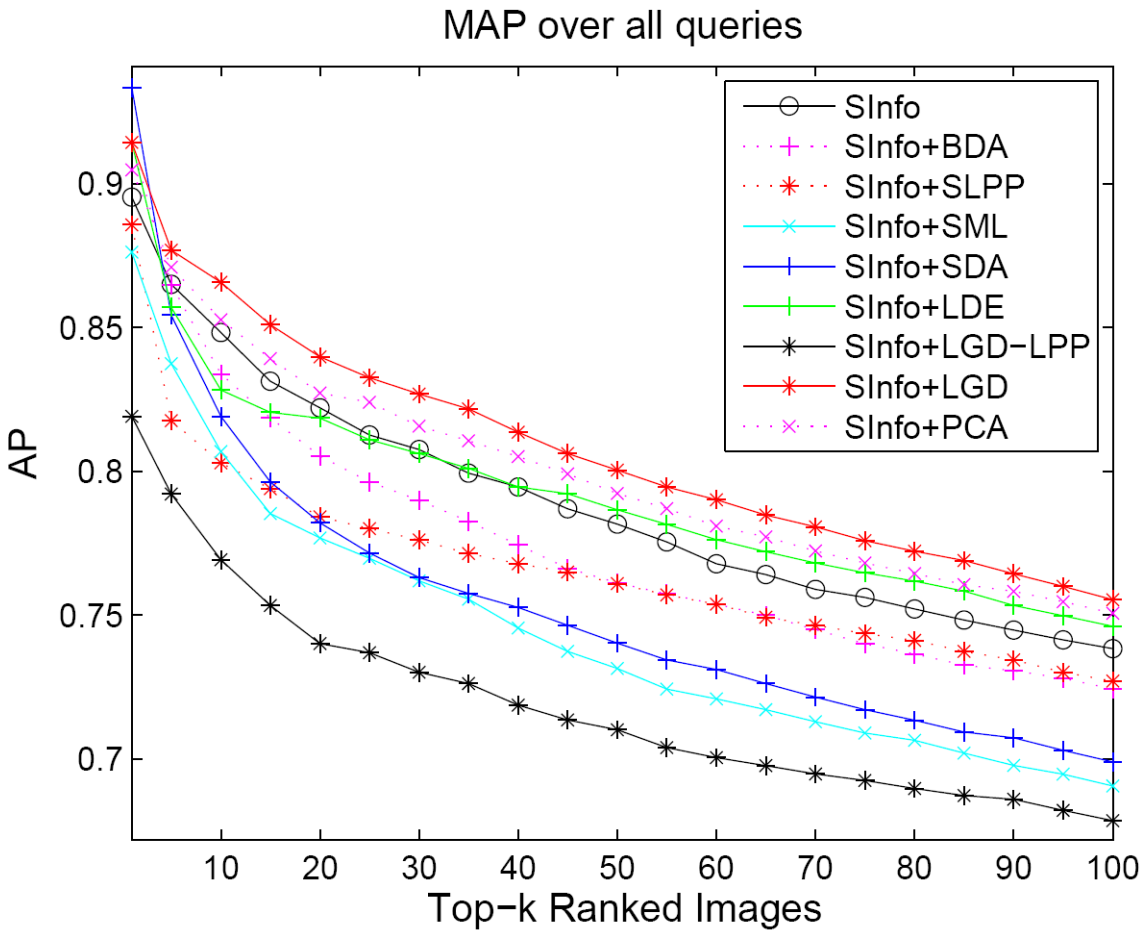
105 queries, 94341 images

225-D color moment, 128-D wavelet texture, 75-D edge distribution histogram

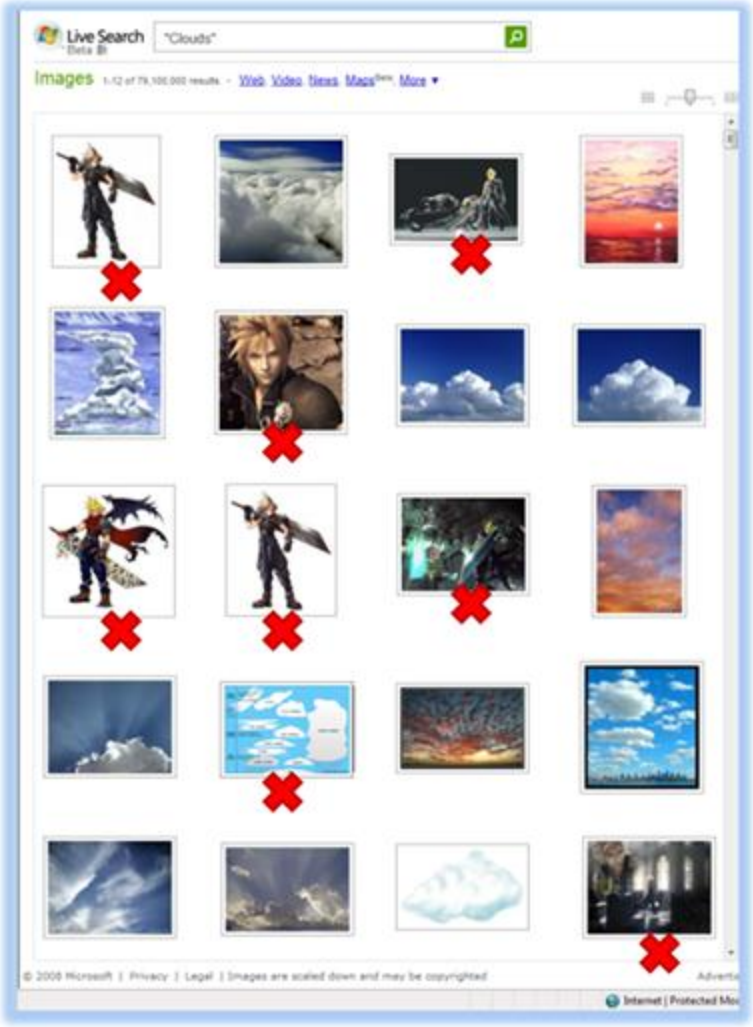
5 images in each interaction round, 4 rounds.

Experiments on Real Search Dataset

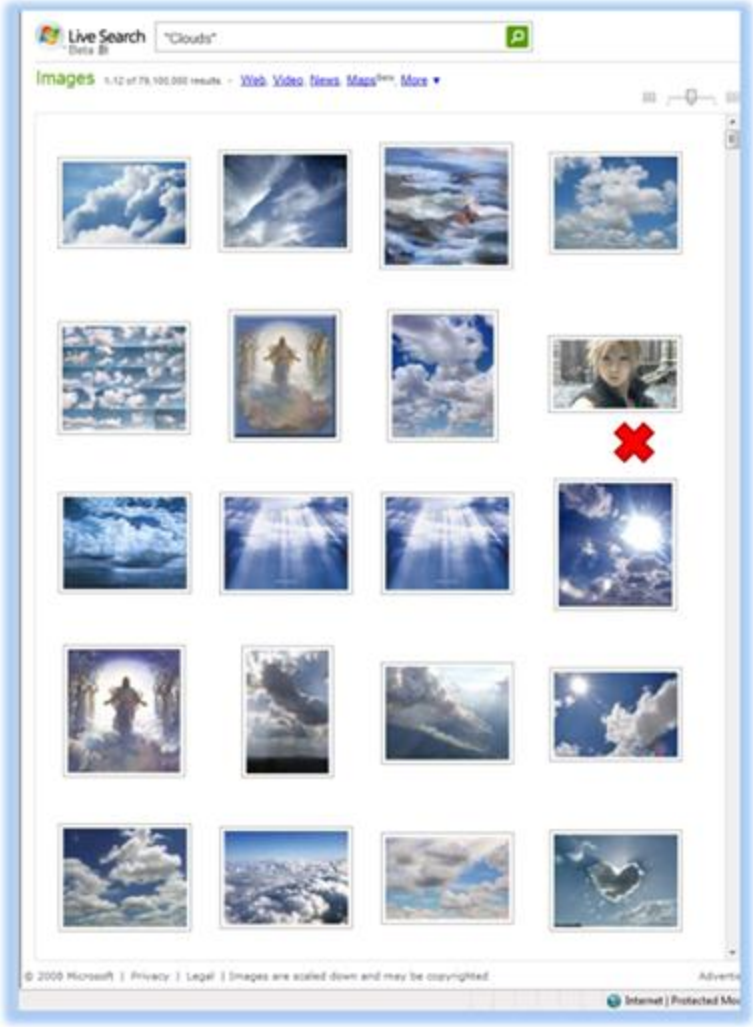
- Dimensionality Reduction
  - LGD vs. PCA, LGD-LPP, BDA, LDE, SLPP, SDA and SML



Results of Active Reranking: Clouds



Text-based search result

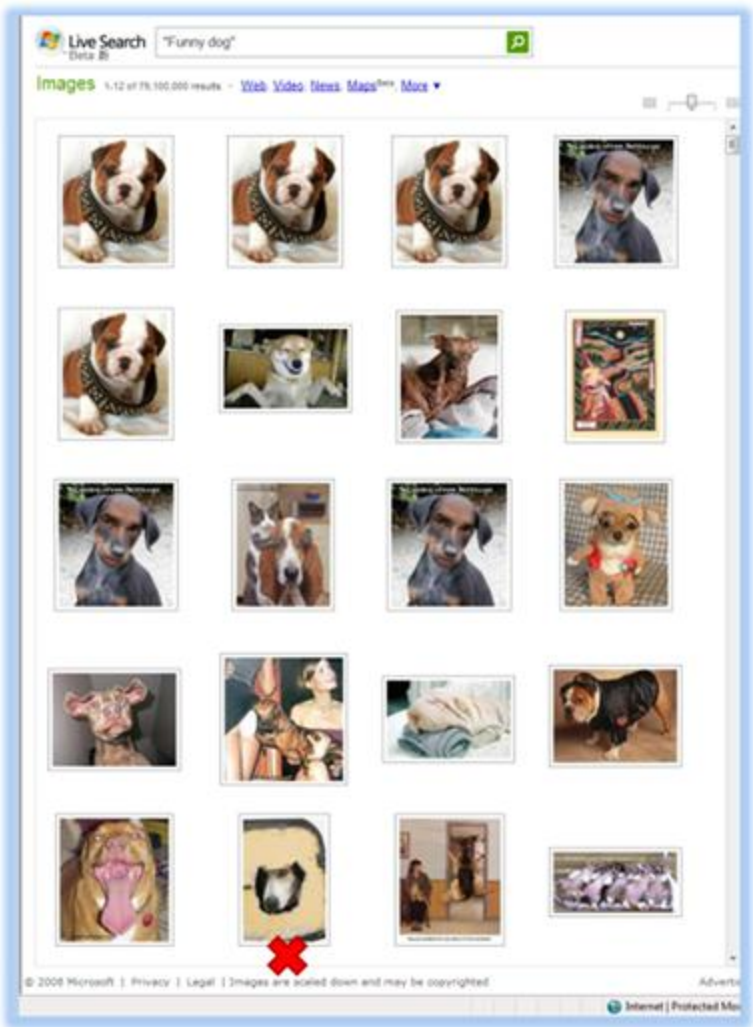


Reranked result

Results of Active Reranking:

Funny dog

Text-based search result



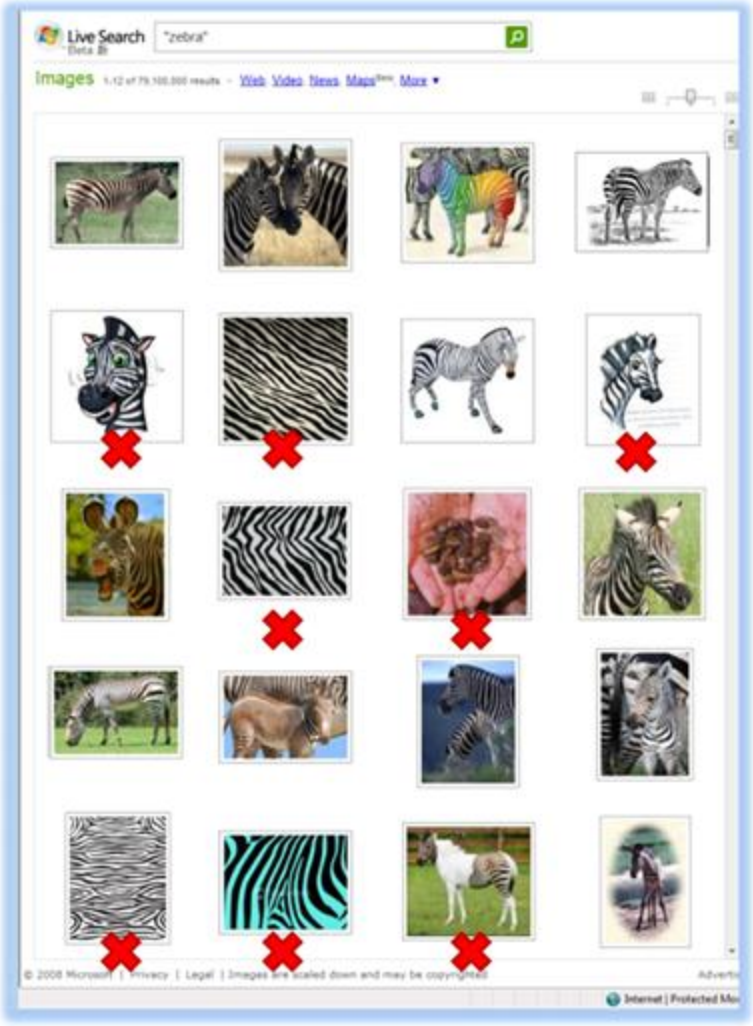
Reranked result

Results of Active Reranking: Panda

Text-based search result

Reranked result

Results of Active Reranking: zebra



Text-based search result

Reranked result

Applications, Challenges and Future Directions

PART FIVE

UTS:QCIS
QUANTUM COMPUTATION & INTELLIGENT SYSTEMS

QUERY DIFFICULTY ESTIMATION (QDE)

Applications, Challenges and Future Directions

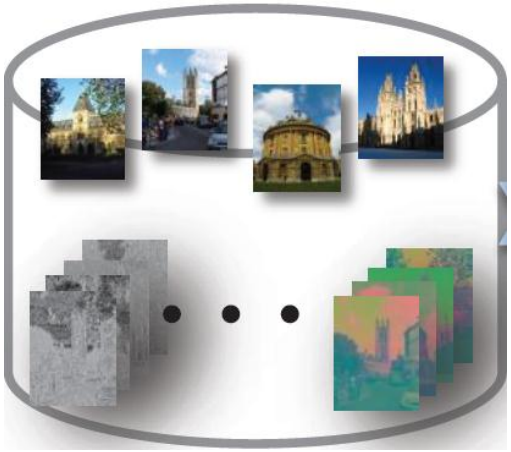
PART FIVE

Query difficulty estimation (QDE)

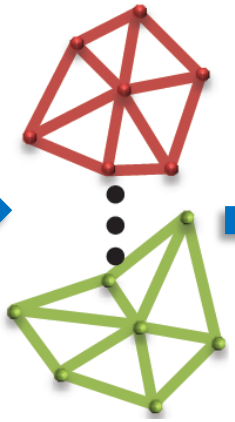
- "Difficult" queries degrade effectiveness of subsequent query-dependent reranking

- Query difficulty estimation
  - quantitatively estimate the retrieval performance of a given query on a given dataset.

The MSE based QDE System



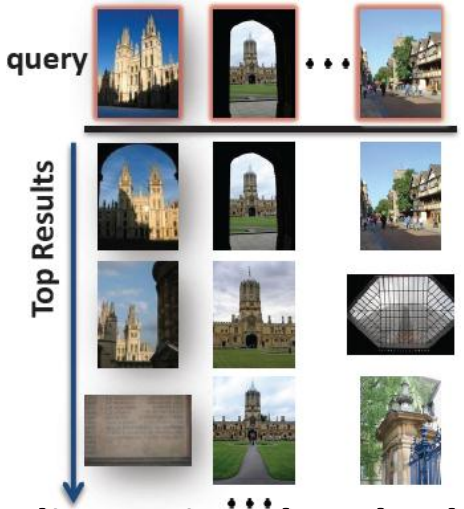
Multiple feature data



Multigraphs

Linear MSE

$$Y = U^T X$$



Low-dimensional embedding

Difficulty features

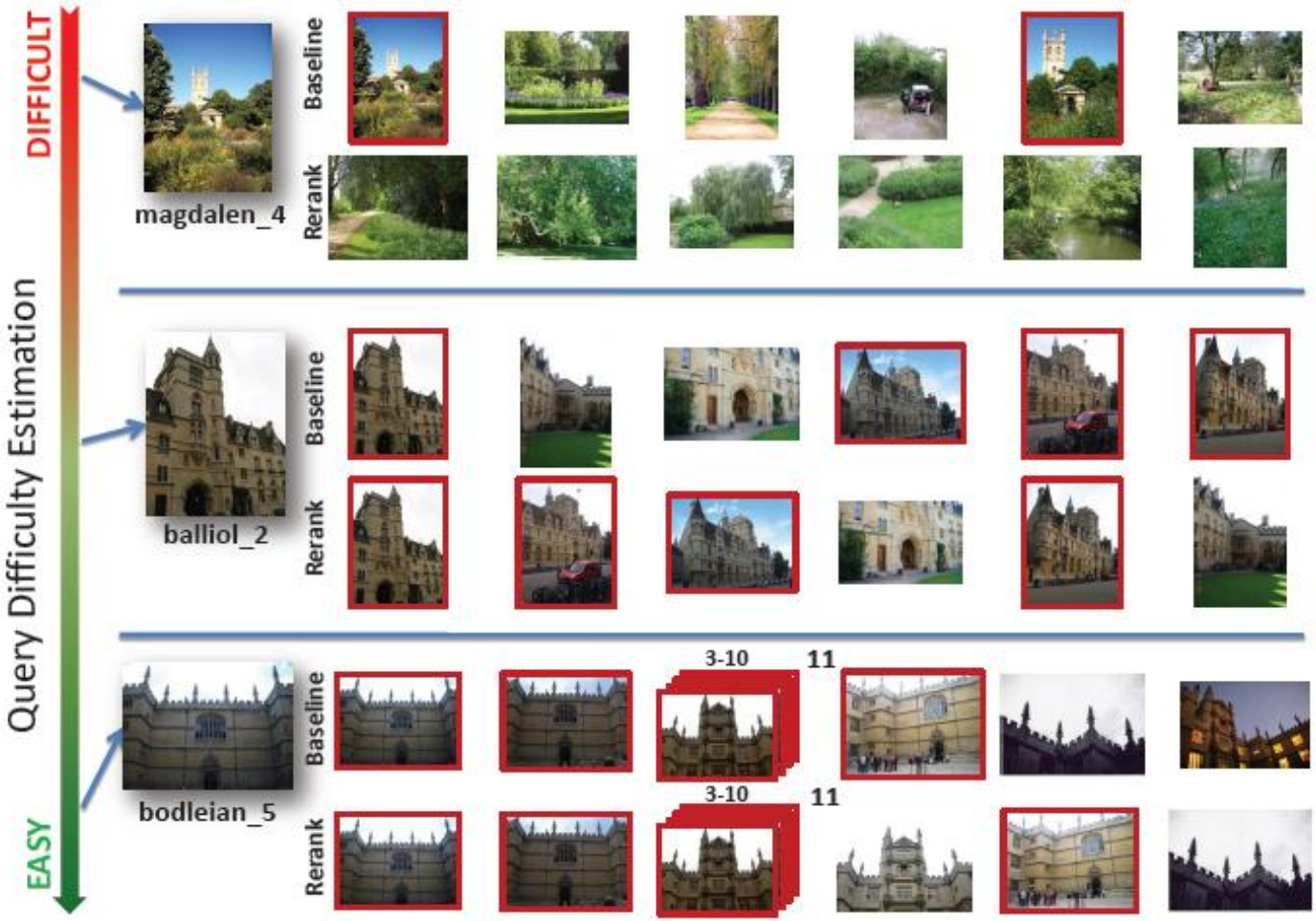
(Coherence, Prominence)

Query difficulty estimation

(classification)

Difficulty guided applications

Results (1)



Reranking deteriorates initial results of difficult query, while has little effect on easy queries

Results (2)

In QDE guided reranking, rearrangement is only conducted on moderately difficult queries

	Dataset	JSD	RQDE	LQDE	Proposed
PRF	*Oxford*	0.624	0.633	0.625	**0.643**
	Oxford+1M	0.534	0.532	0.528	**0.542**
	NUS-WIDE	0.155	0.149	0.163	**0.172**
Bayesian	*Oxford*	0.646	0.655	0.648	**0.661**
	Oxford+1M	0.543	0.539	0.540	**0.548**
	NUS-WIDE	0.162	0.159	0.172	**0.176**

mAP

Oxford: 55 queries, 5062 images

Oxford+1M: Oxford + 1M ImageNet irrelevant images

NUS-WIDE: 50 object queries, 30,000 images

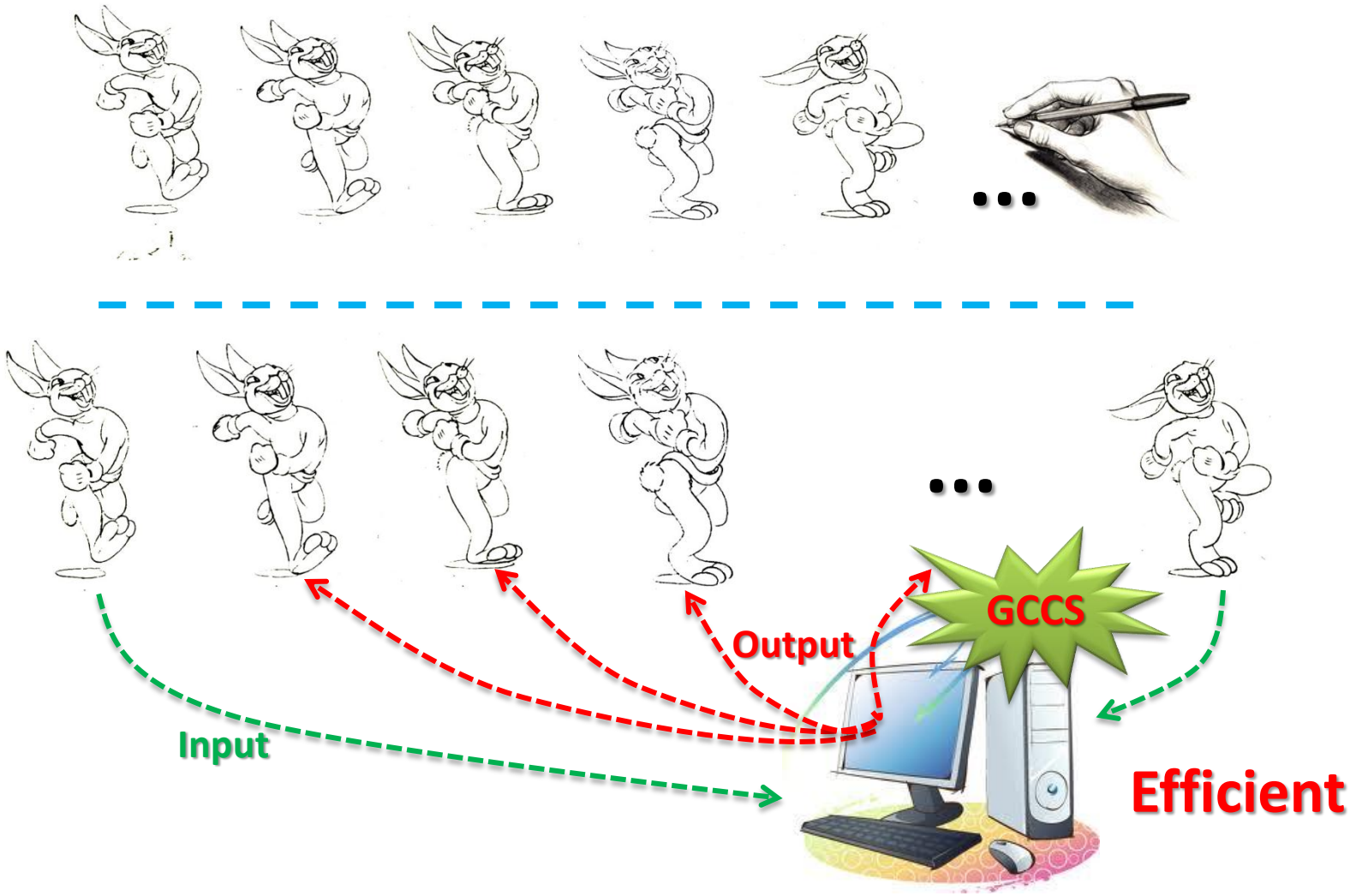
CARTOON SYNTHESIS

Applications, Challenges and Future Directions

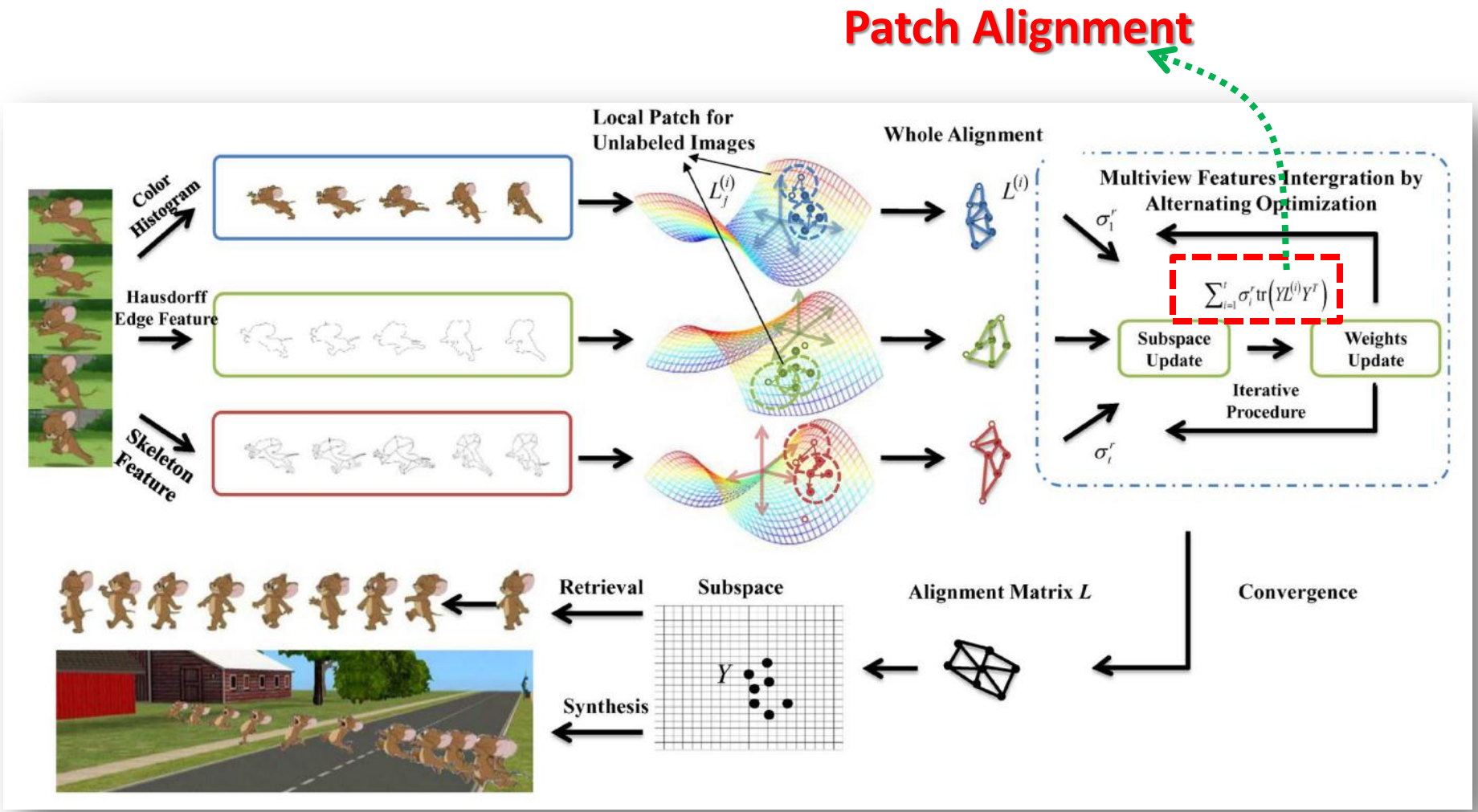
PART FIVE

UTS:QCIS
QUANTUM COMPUTATION & INTELLIGENT SYSTEMS

Cartoon Synthesis



Framework


Patch Alignment
Local Patch for Unlabeled Images
Whole Alignment
Color Histogram
Hausdorff Edge Feature
Skeleton Feature
$L_j^{(i)}$
$L^{(i)}$
Multiview Features Intergration by Alternating Optimization
σ_1^r
$\sum_{i=1}^{t}\sigma_i^r \mathrm{tr}\left(YL^{(i)}Y^T\right)$
Subspace Update
Weights Update
Iterative Procedure
σ_t^r
Retrieval
Subspace
Alignment Matrix L
Convergence
Y
Synthesis

Cartoon Characters Retrieval

Semi-MSL

MSL

UMSL

UBDML

DMSL

FCSL

MSL: multiview subspace learning

Semi-MSL: semi-superivised MSL

UMSL: unsuperivised MSL

UBDML: unsupervised bidistance metric learning

DMSL: distributed MSL

FCSL: concatenation based subspace learning

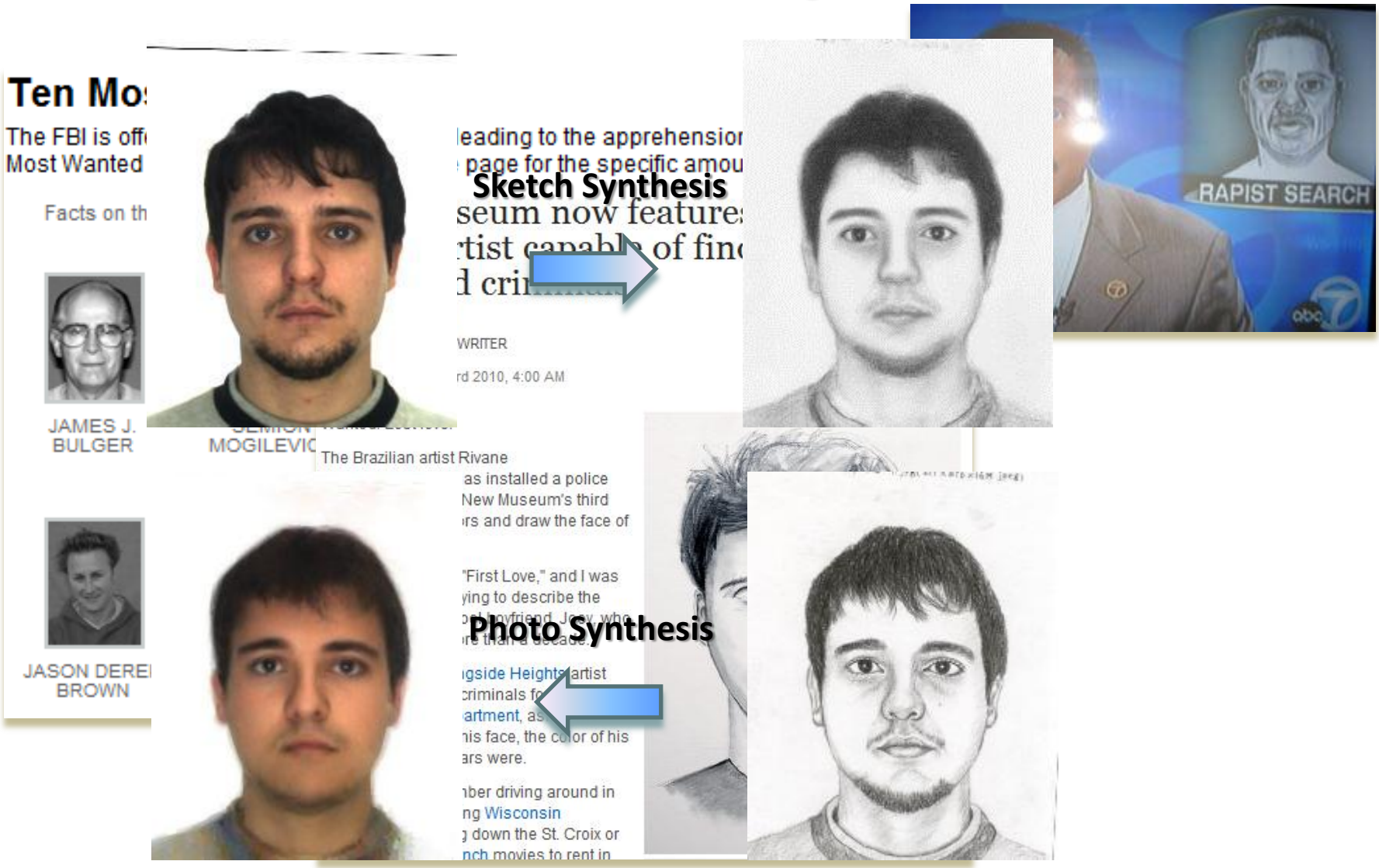
FACE SKETCH-PHOTO SYNTHESIS

Applications, Challenges and Future Directions

PART FIVE

UTS:QCIS
QUANTUM COMPUTATION & INTELLIGENT SYSTEMS

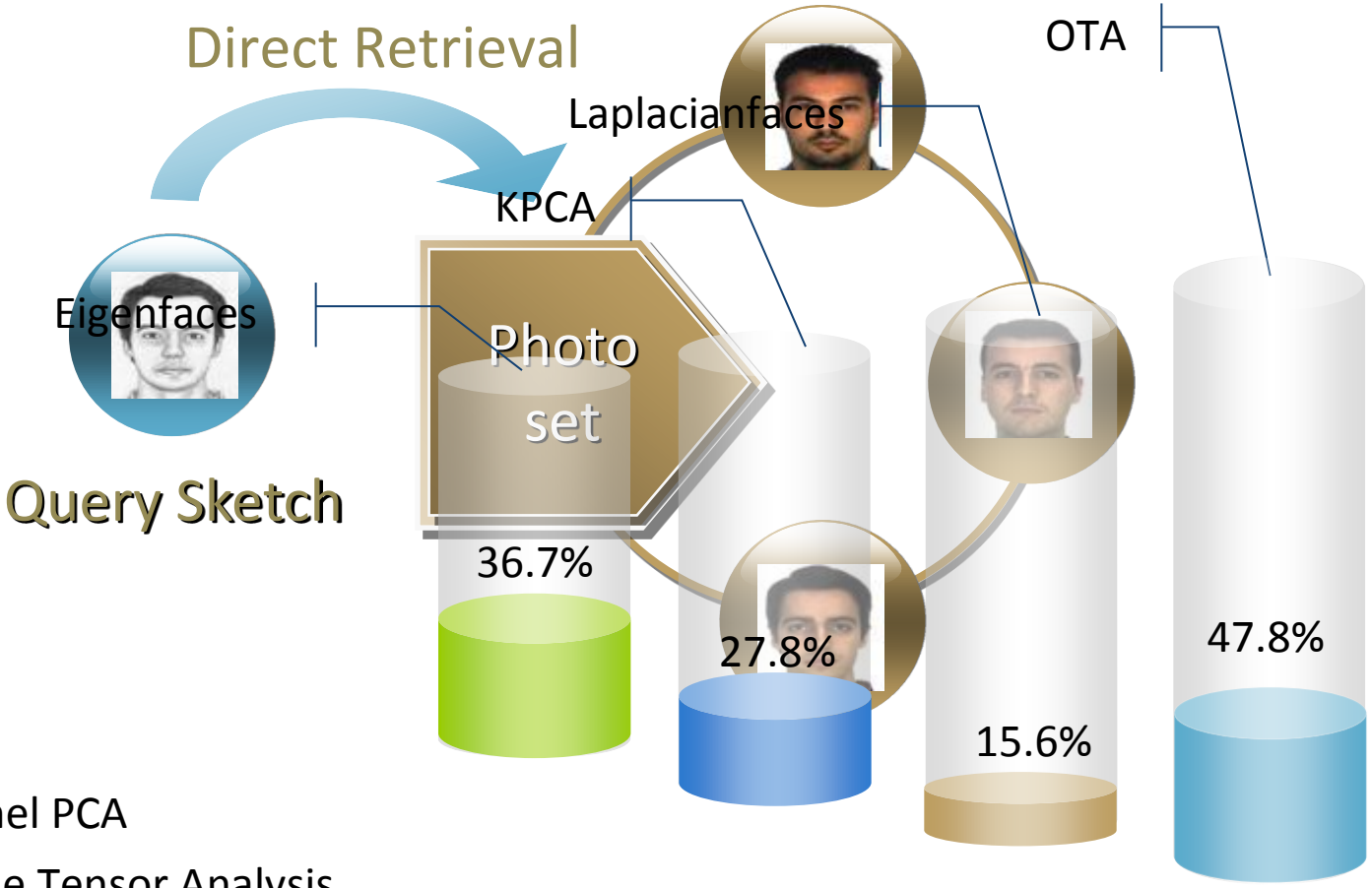
Face Sketch-Photo Synthesis



Direct Face Retrieval

All ≤ 50 !

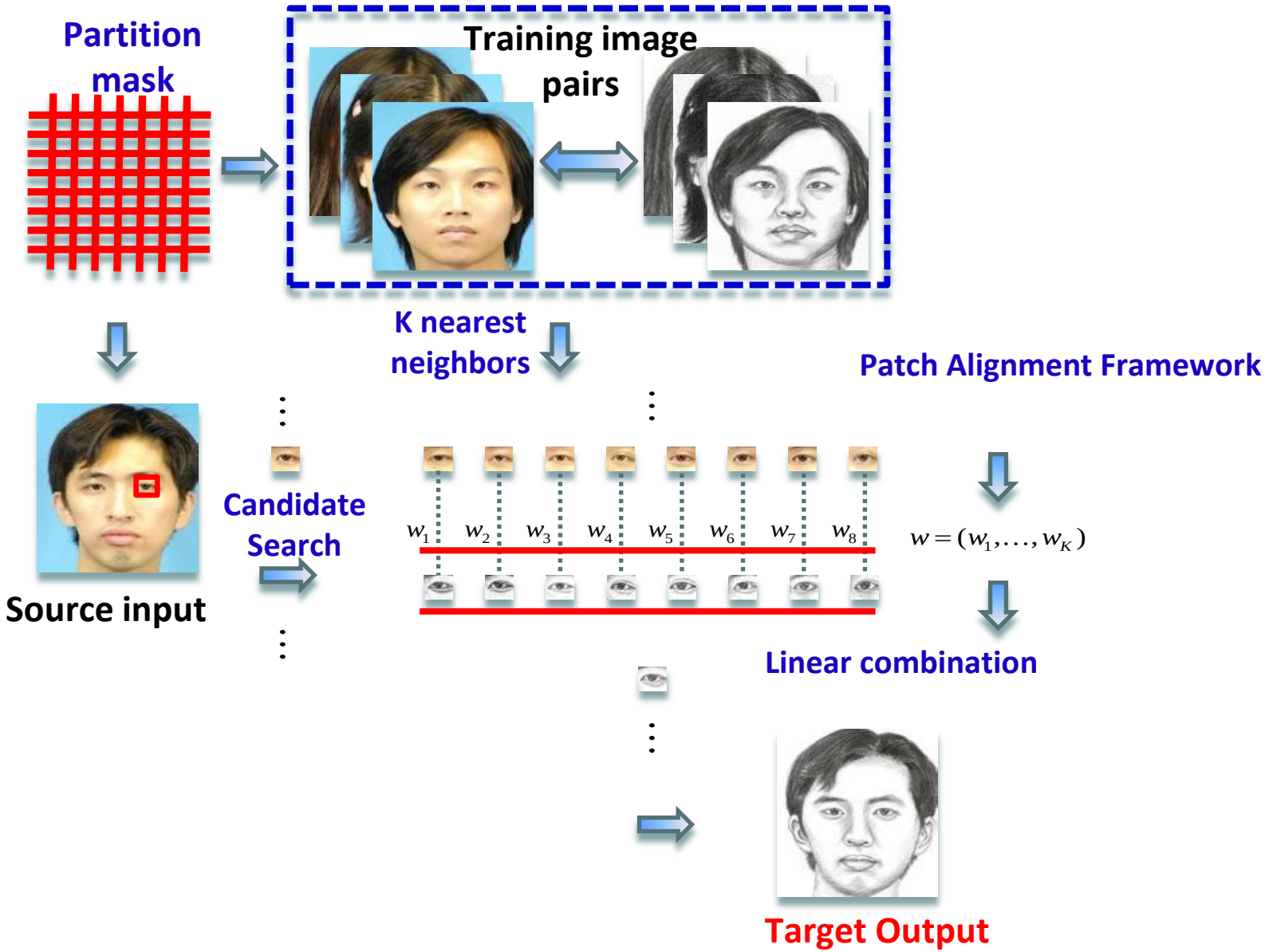
Performance Evaluation



KPCA: Kernel PCA

OTA: Offline Tensor Analysis

Face Sketch Synthesis Framework

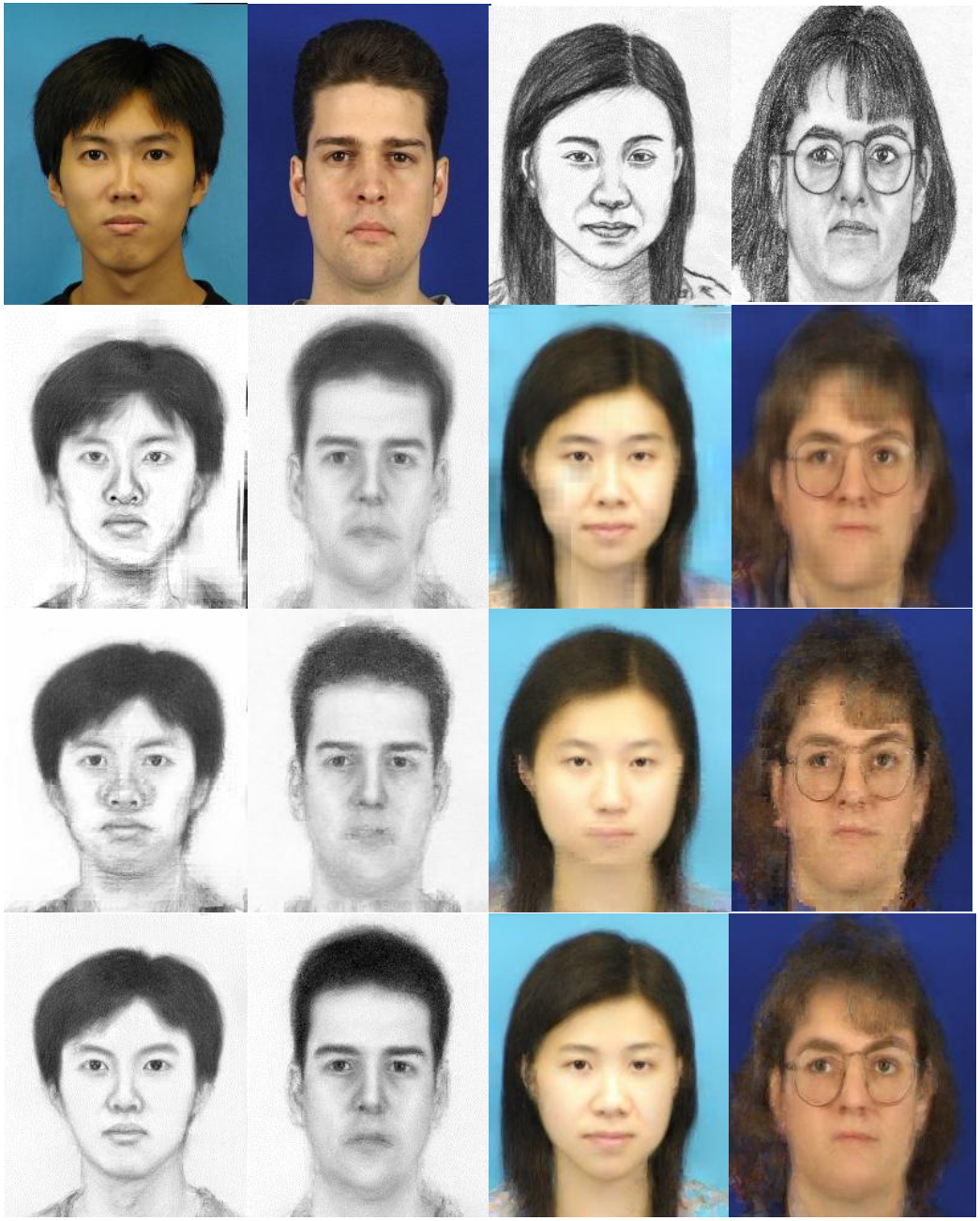


Input

LLE

MWF

Ours



Synthesized sketches Synthesized Photos

Face Recognition (Sketch)

Method	FR(%)
Direct Match	47.8
LLE	84
MWF	88
Ours	**93**

Face Recognition (Photo)

Method	FR(%)
Direct Match	47.8
LLE	87
MWF	92.5
Ours	**94.5**

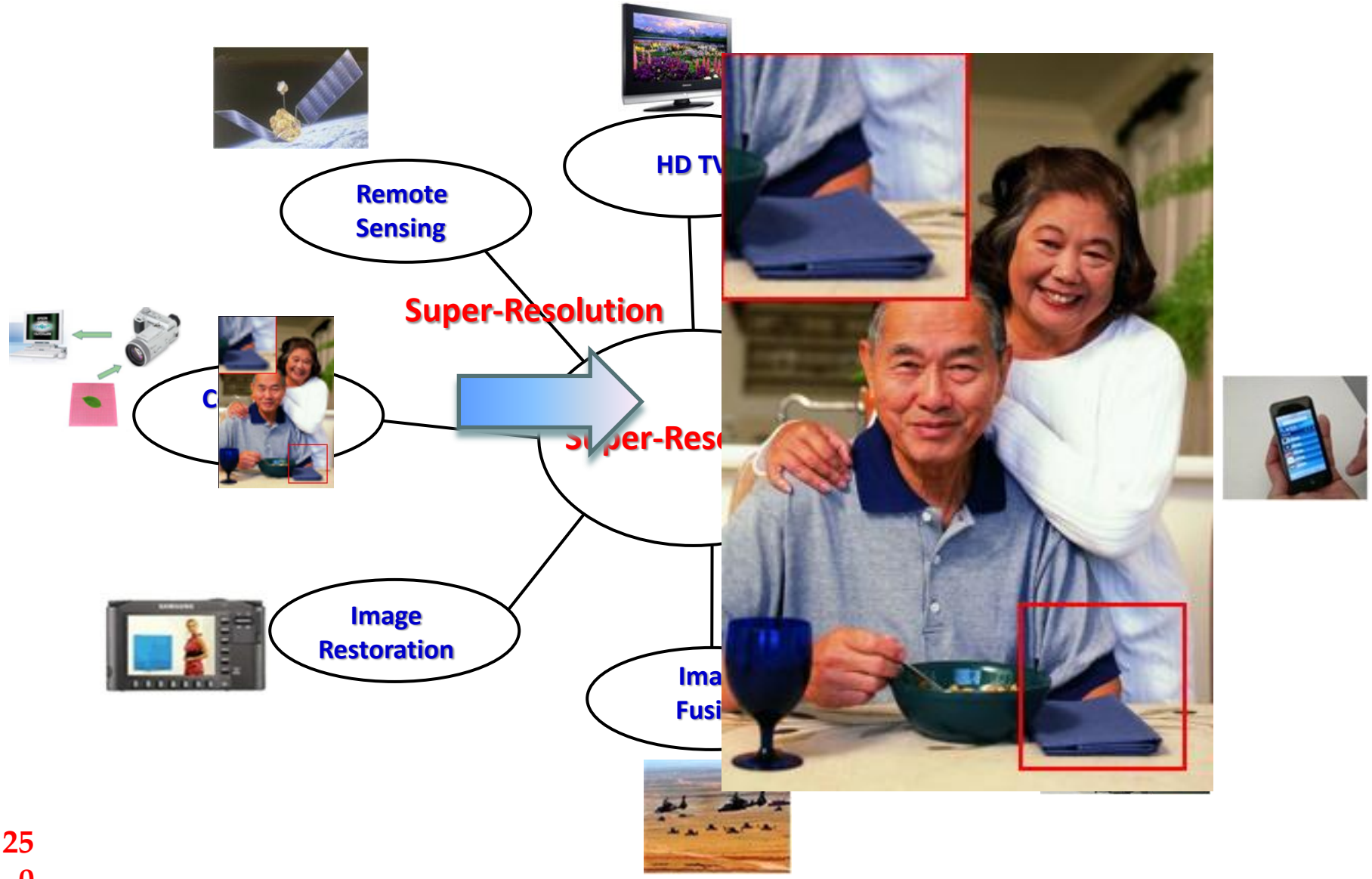
IMAGE SUPER-RESOLUTION

Applications, Challenges and Future Directions

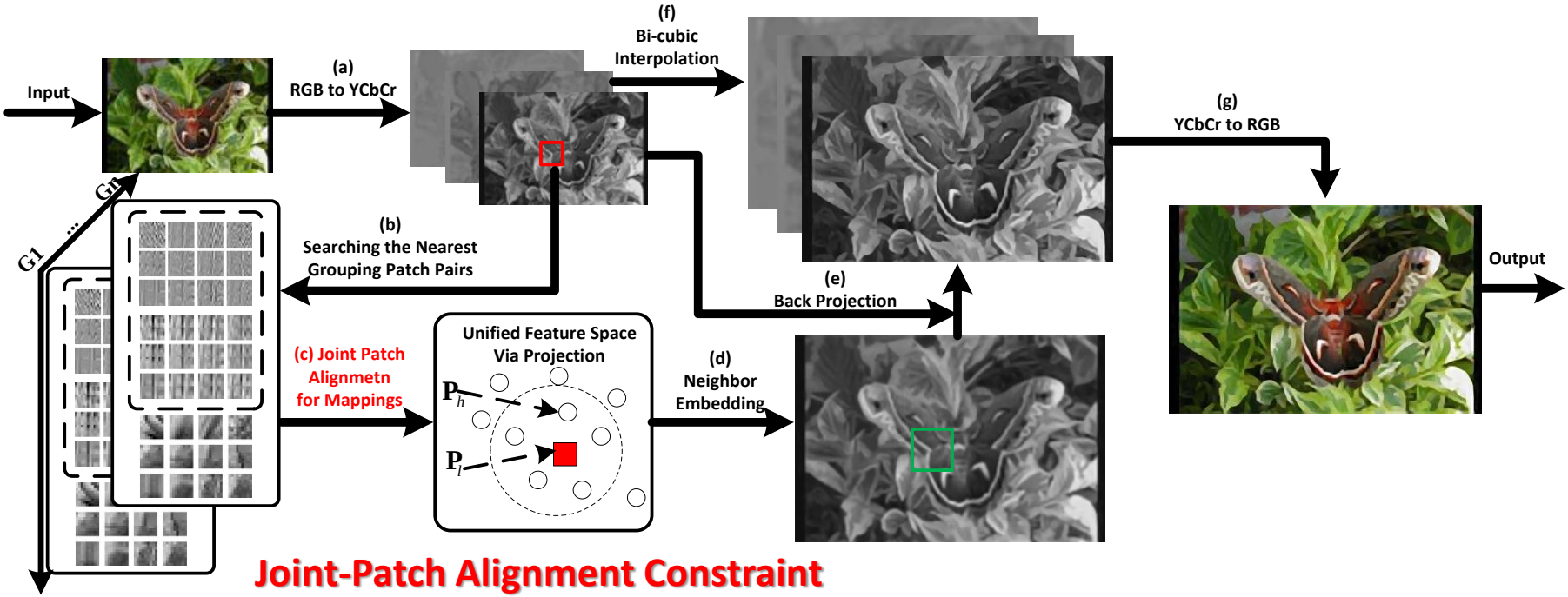
PART FIVE

UTS:QCIS
QUANTUM COMPUTATION & INTELLIGENT SYSTEMS

Super-Resolution



Super-Resolution Framework



Super-Resolution Results

(a) Ground Truth

(b) SC

(c) SR

(d) TV

(e) [16]

(f)ASDS

(g) [10]

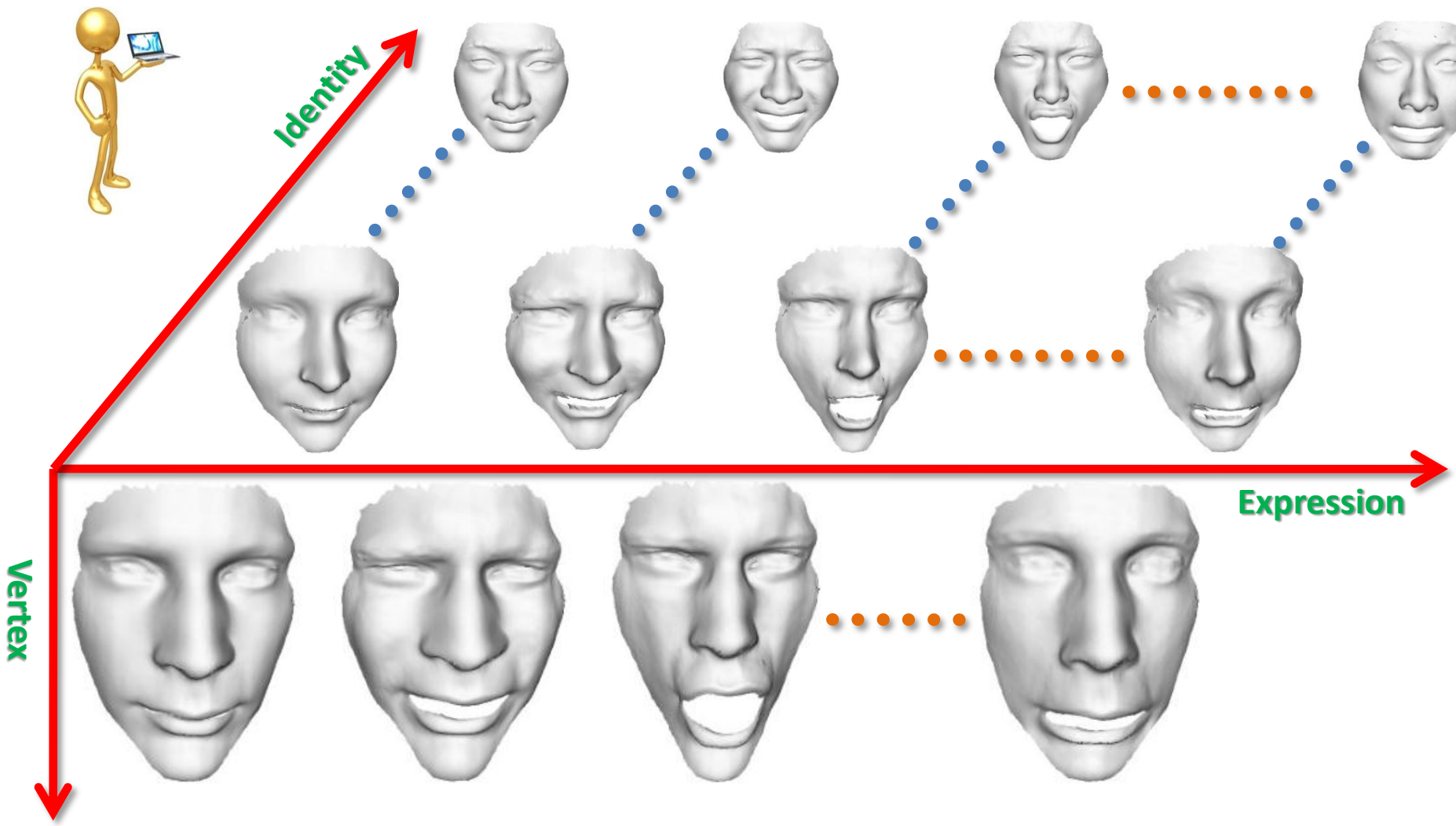
(h) Ours

Super-Resolution Results

PSNR and SSIM (bottom) Value

No.	Methods				
	Bicubic	NE	NeedFS	PSNE	**Ours**
1	25.495 0.7758	27.076 0.7963	27.362 0.8036	27.239 0.7973	**28.291** **0.8404**
2	27.842 0.8116	28.887 0.8222	29.122 0.8317	29.02 0.8286	**29.848** **0.8557**
3	28.788 0.7996	26.936 0.7845	28.296 0.8358	28.899 0.8381	**29.115** **0.8630**
4	23.549 0.6562	24.587 0.6582	25.235 0.7028	25.041 0.6965	**25.502** **0.7402**
5	23.359 0.6352	23.747 0.6478	24.099 0.6670	24.13 0.6742	**24.662** **0.7051**
6	26.082 0.8134	29.974 0.8165	29.808 0.8023	29.919 0.819	**31.231** **0.8623**
7	32.534 0.7933	31.158 0.7876	30.219 0.7731	31.755 0.7885	**31.93** **0.7912**
8	30.039 0.8494	29.286 0.8558	25.953 0.8337	29.33 0.8601	**29.358** **0.8677**
Average	27.211 0.7668	27.706 0.7711	27.512 0.7813	28.167 0.7878	**28.742** **0.8157**

3D FACIAL EXPRESSION RETARGETING



3D FACE RECONSTRUCTION

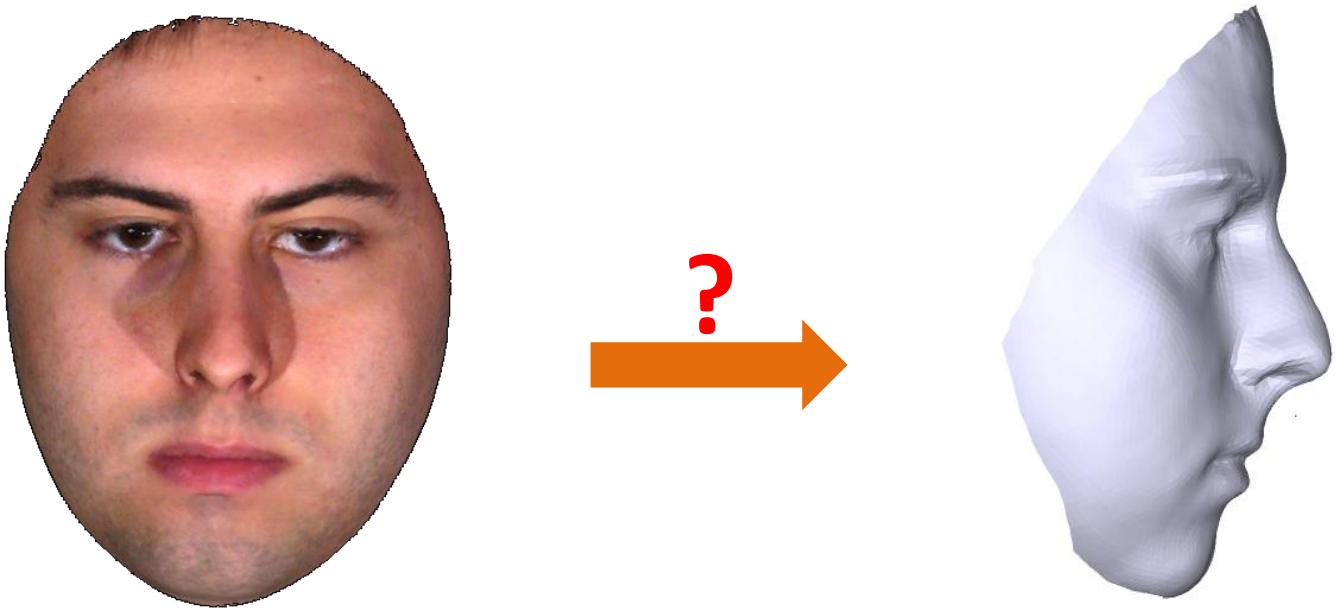
Applications, Challenges and Future Directions

PART FIVE

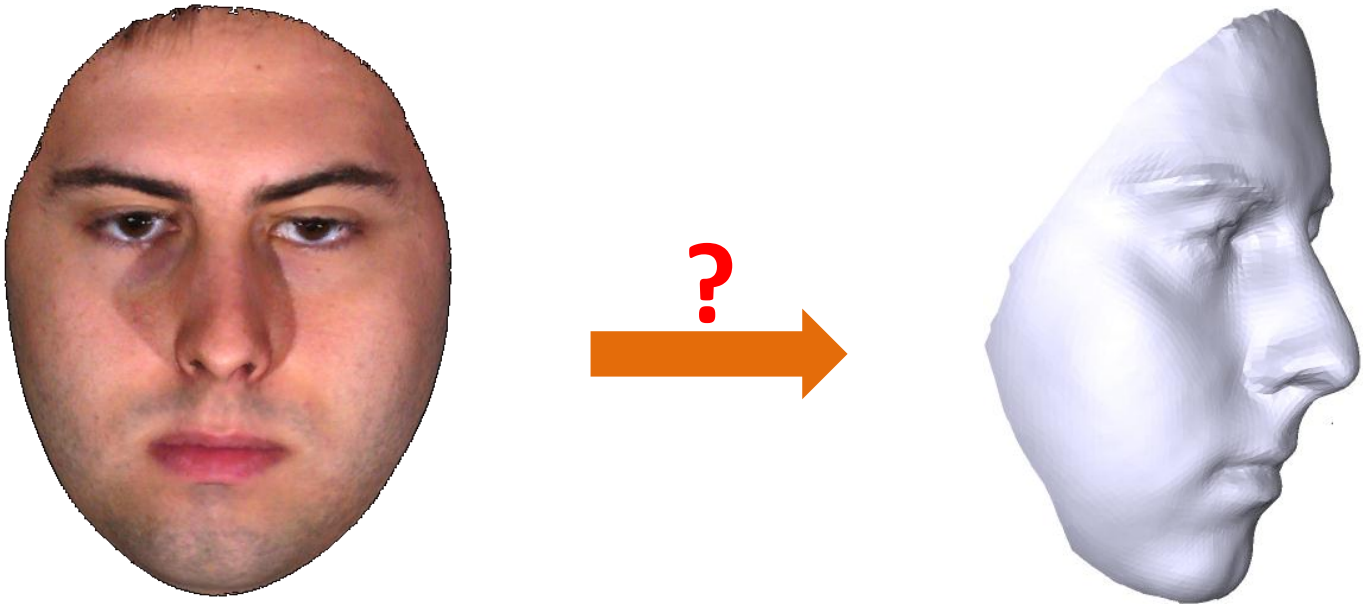
UTS:QCIS
QUANTUM COMPUTATION & INTELLIGENT SYSTEMS

3D Face Reconstruction

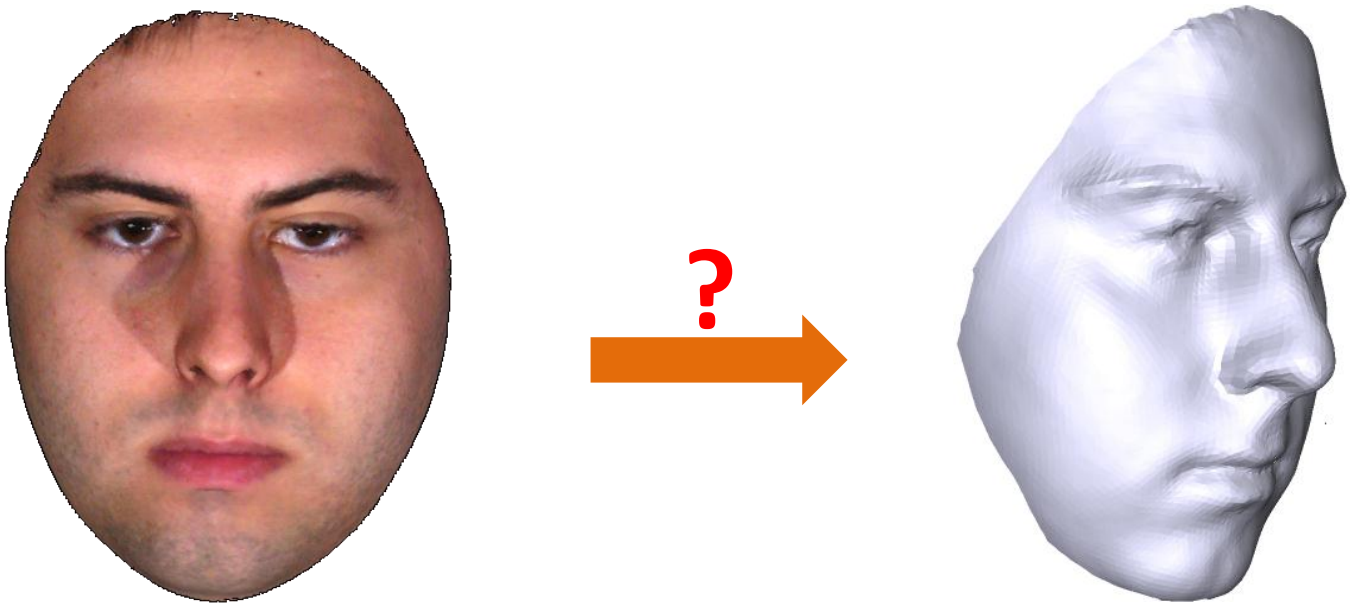
3D Face Reconstruction



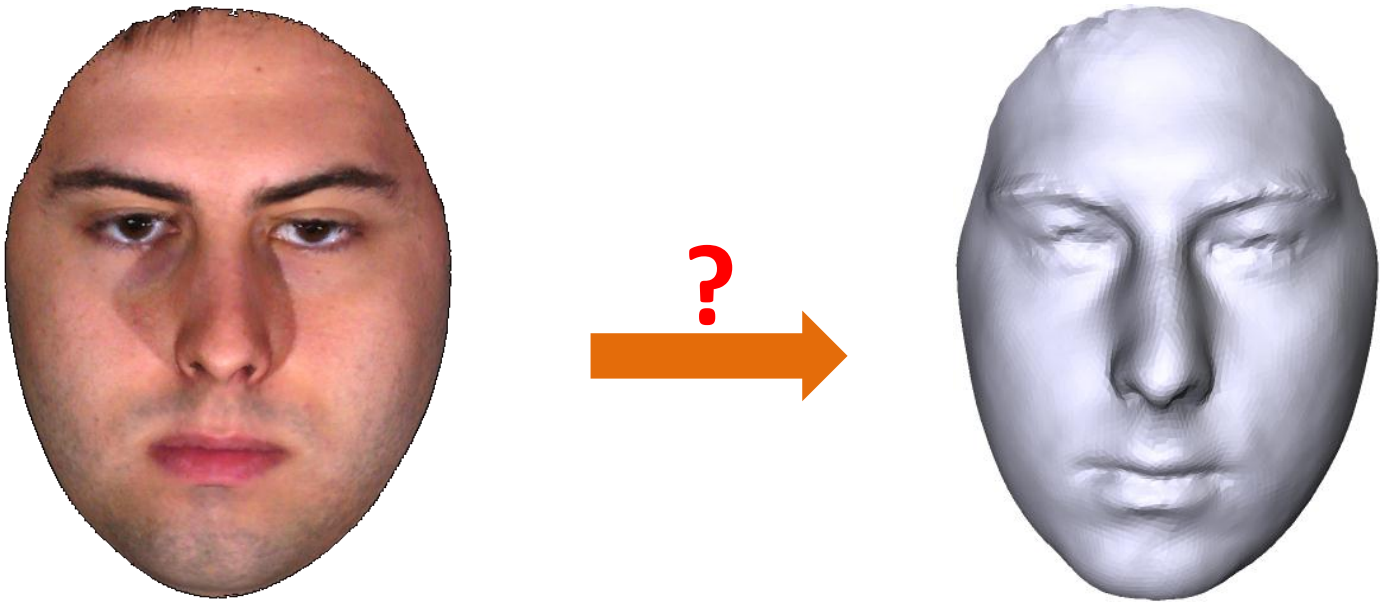
3D Face Reconstruction



3D Face Reconstruction



3D Face Reconstruction



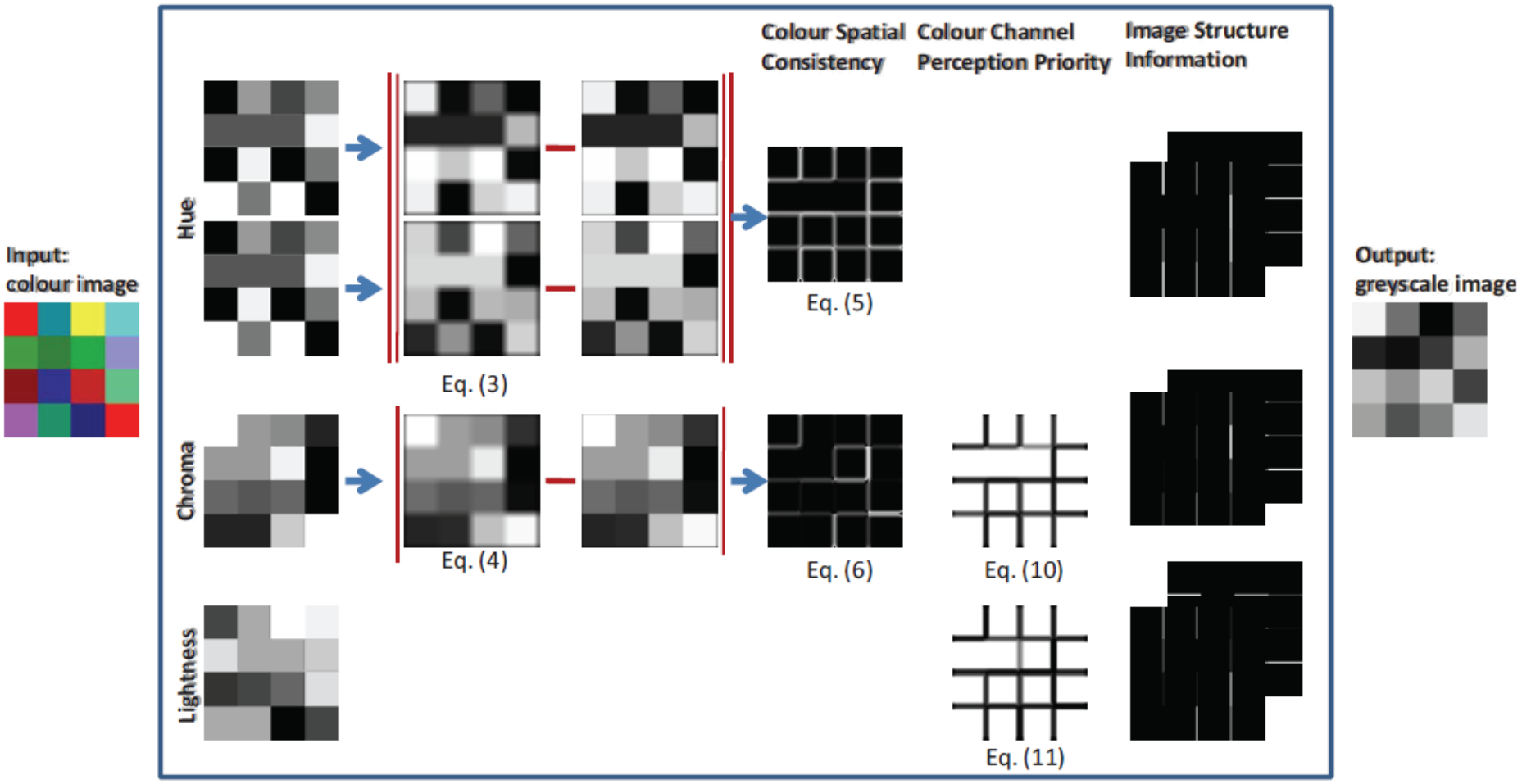
COLOR2GRAY

Applications, Challenges and Future Directions

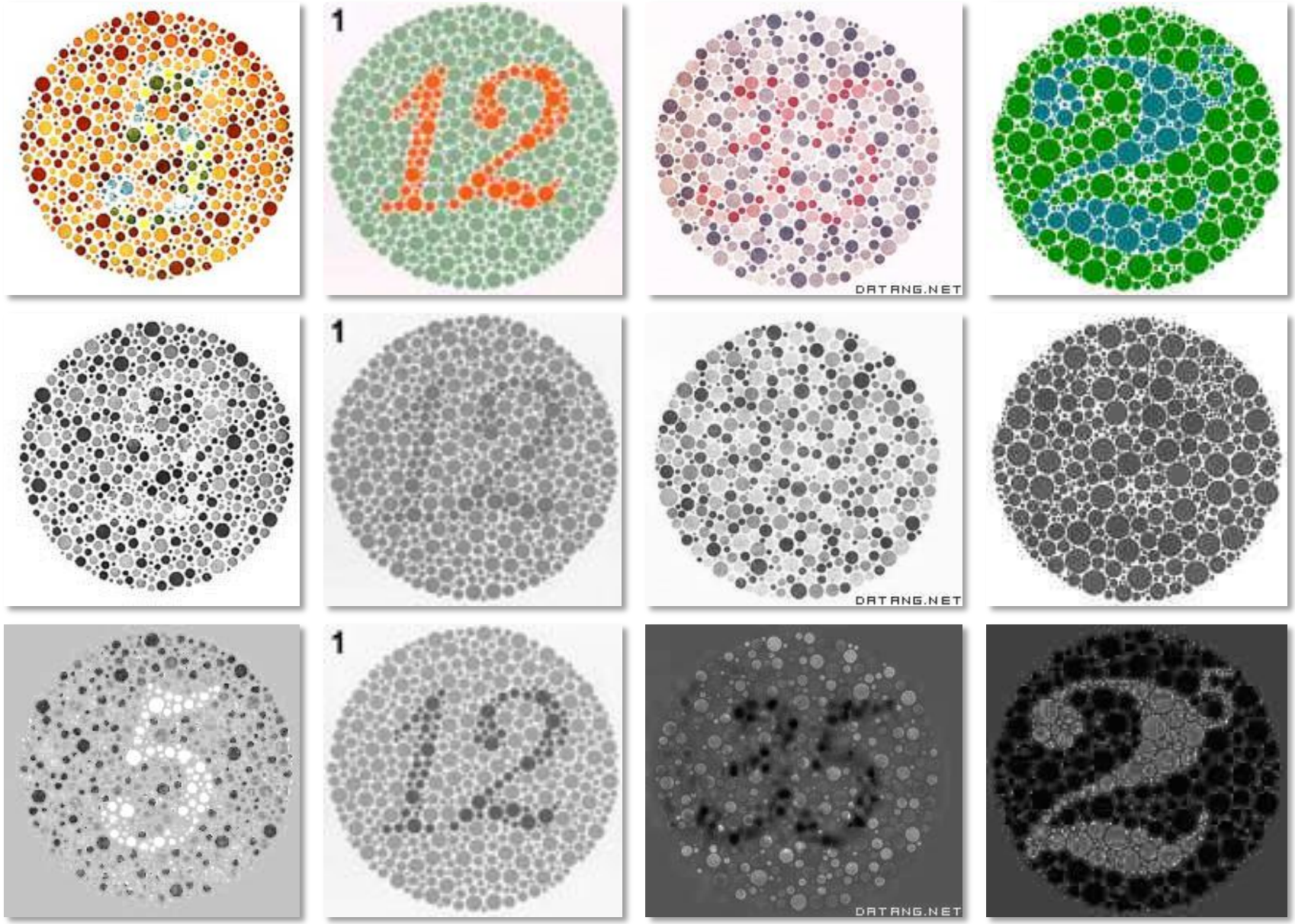
PART FIVE

UTS:QCIS
QUANTUM COMPUTATION & INTELLIGENT SYSTEMS

Colour to Grey: Visual Cue Preservation



Ref: M. Song, D. Tao et al., "Colour to Grey: Visual Cue Preservation," ***IEEE Transactions on Pattern Analysis and Machine Intelligence***, to appear.


1
12
DATANG.NET
1
DATANG.NET
1
DATANG.NET

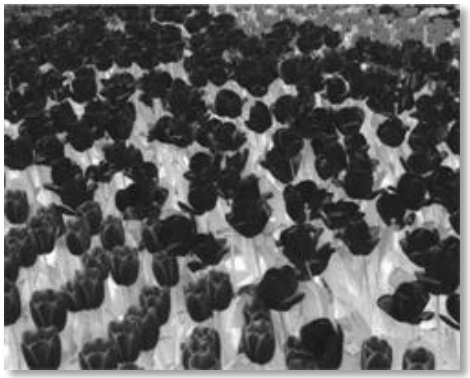










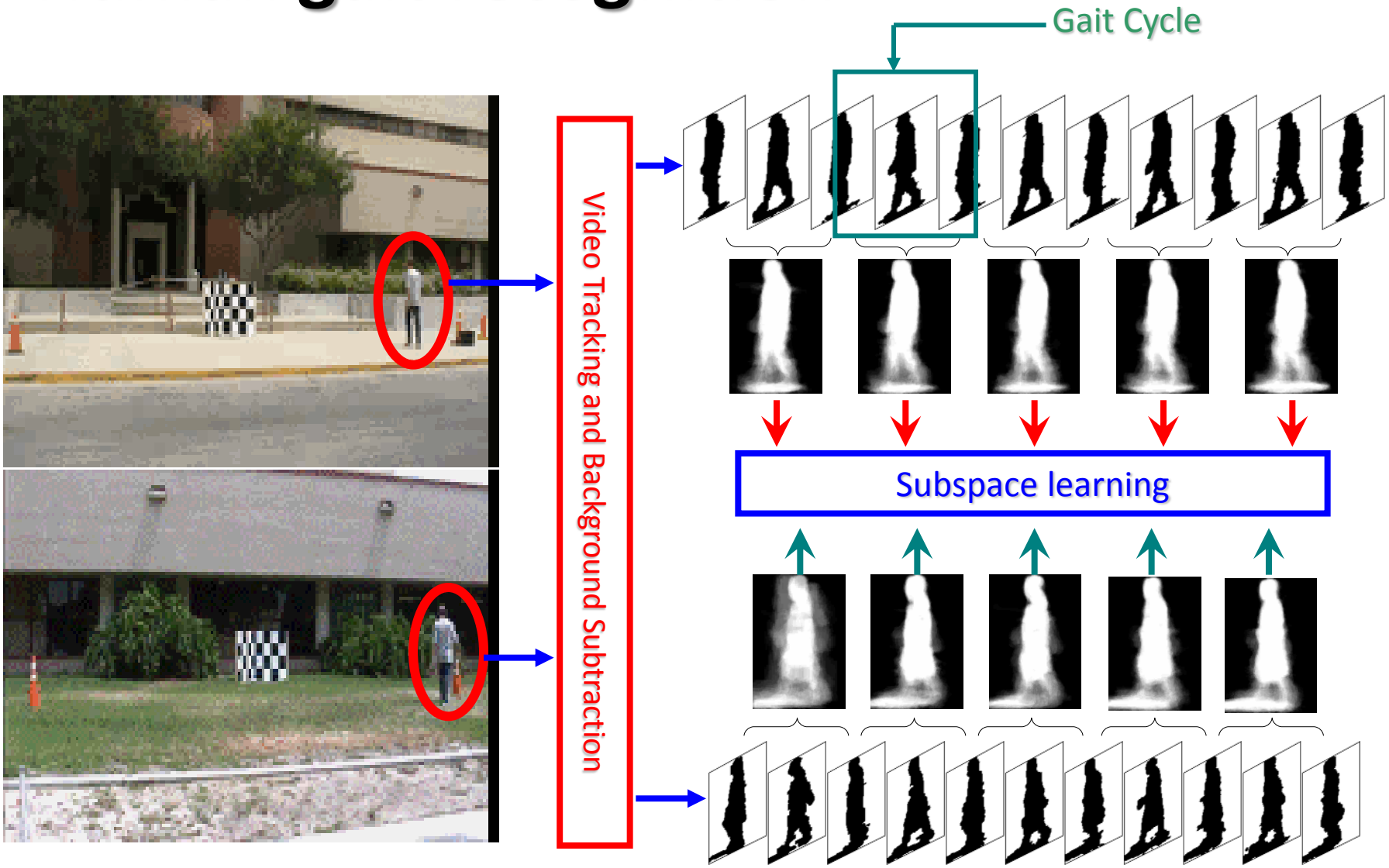



HUMAN GAIT RECOGNITION
ACTION RECOGNITION

Applications, Challenges and Future Directions

PART FIVE

Human gait recognition



Challenges and Future

- Theoretical aspect
  - Error bounds
  - Stability
- Algorithmic aspect
  - Speed
  - Scalability
- Application
- Data ?

- To ICME organizers: Dan, Jian, Qiang, etc.
- To Lei Zhang at MSRA, etc.
- To students in UTS, Xidian Univ, Peking Univ, USTC, etc.

References

- Tianhao Zhang, Dacheng Tao, Xuelong Li, Jie Yang: Patch Alignment for Dimensionality Reduction. IEEE Trans. Knowl. Data Eng. 21(9): 1299-1313 (2009)
- Naiyang Guan, Dacheng Tao, Zhigang Luo, Bo Yuan: Non-Negative Patch Alignment Framework. IEEE Transactions on Neural Networks 22(8): 1218-1230 (2011)
- Naiyang Guan, Dacheng Tao, Zhigang Luo, Bo Yuan: NeNMF: An Optimal Gradient Method for Nonnegative Matrix Factorization. IEEE Transactions on Signal Processing 60(6): 2882-2898 (2012)
- Naiyang Guan, Dacheng Tao, Zhigang Luo, Bo Yuan: Online Nonnegative Matrix Factorization With Robust Stochastic Approximation. IEEE Trans. Neural Netw. Learning Syst. 23(7): 1087-1099 (2012)
- Tianyi Zhou, Dacheng Tao, Xindong Wu: Manifold elastic net: a unified framework for sparse dimension reduction. Data Min. Knowl. Discov. 22(3): 340-371 (2011)
- Si Si, Dacheng Tao, Bo Geng: Bregman Divergence-Based Regularization for Transfer Subspace Learning. IEEE Trans. Knowl. Data Eng. 22(7): 929-942 (2010)
- Tian Xia, Dacheng Tao, Tao Mei, Yongdong Zhang: Multiview Spectral Embedding. IEEE Transactions on Systems, Man, and Cybernetics, Part B 40(6): 1438-1446 (2010)
- Xinmei Tian, Dacheng Tao, Xian-Sheng Hua, Xiuqing Wu: Active Reranking for Web Image Search. IEEE Transactions on Image Processing 19(3): 805-820 (2010)

References

- TbA


BIG
DATA

Thanks!

UTS:QCIS
QUANTUM COMPUTATION & INTELLIGENT SYSTEMS