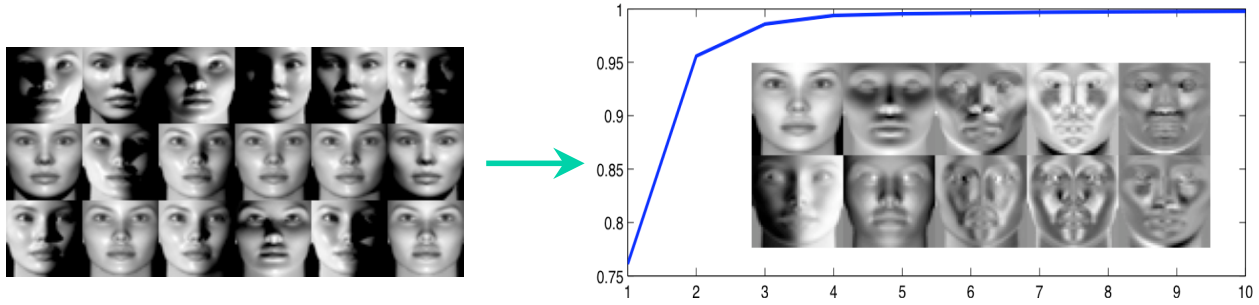
Subspace-based Learning with Grassmann Kernels

Jihun Hamm and Daniel. D. Lee
University of Pennsylvania
July 8, 2008

Subspace structure in data

- Image data



  - Illumination variation is low-dimensional
    - empirical [Hallinan94,Epstein95], theoretical [Belhumeur98,Ramamoorthi01,Ramamoorthi02,Basri03]
  - Pose, expression, etc also modeled well by subspaces: "Eigenface" [Sirovich87,Kirby90,Turk91]

Set of illumination-subspaces

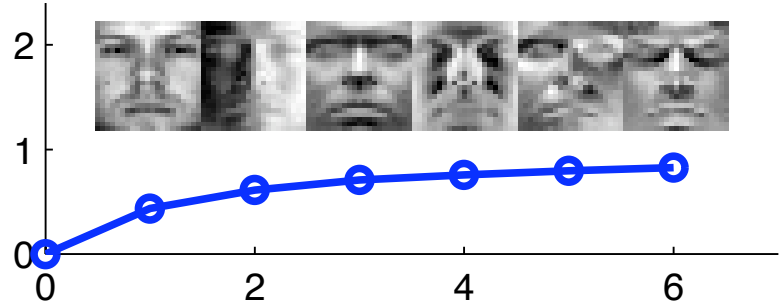
sets

subspaces

X_1

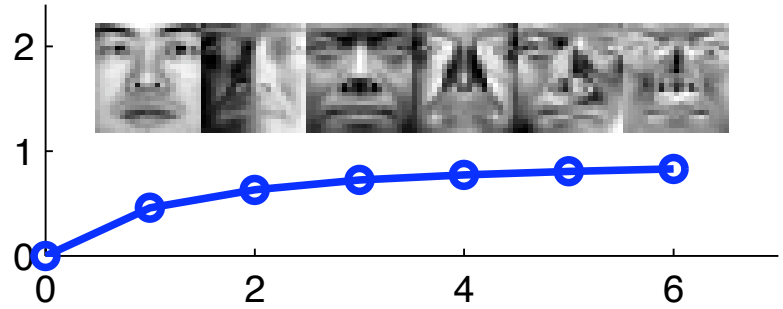
PCA

$Y_1 \in \mathbb{R}^{D \times m}$



X_2

Y_2

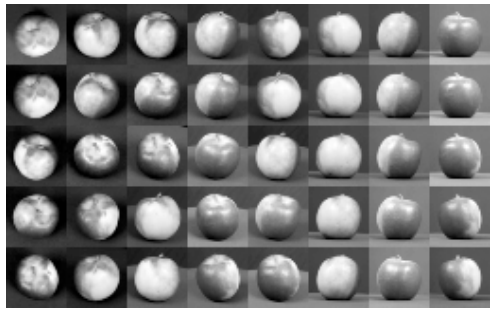


Set of pose-subsapces

sets

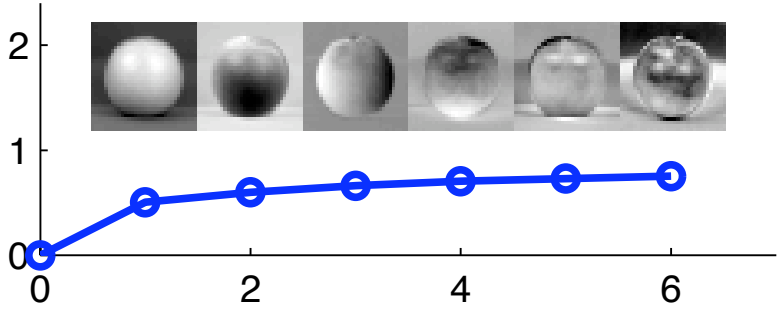
subspaces

X_1

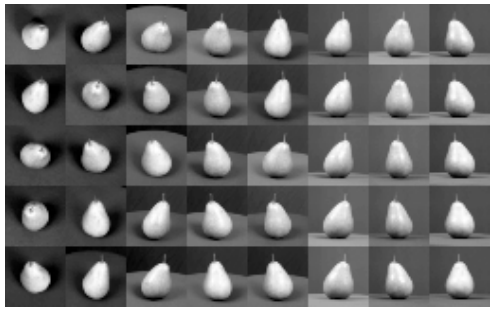


PCA

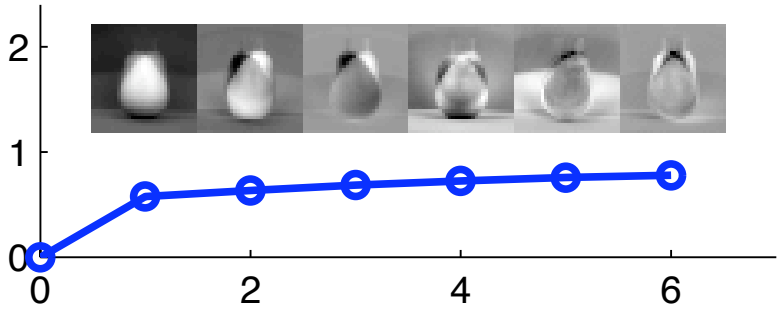
$Y_1 \in \mathbb{R}^{D \times m}$



X_2



Y_2



Linear dynamical model

- ARMA model of a sequential data

$$\begin{aligned} x(t+1) &= Ax(t) + v(t) \\ y(t) &= Cx(t) + w(t) \end{aligned}$$

$x(t)$: internal state, $y(t)$: observed output, $v(t), w(t)$: noise

  - e.g., dynamic textures, human actions
    [Doretto03, Veeraraghavan05, Turaga08]

- Observability matrix [Cock02]

$$O_{(A,C)} = [C \; ; \; CA \; ; \; CA^2 \; ; \; ...] \in \mathbb{R}^{D \times m}$$

  - Each ARMA model spans a unique subspace

Set of observability subspaces

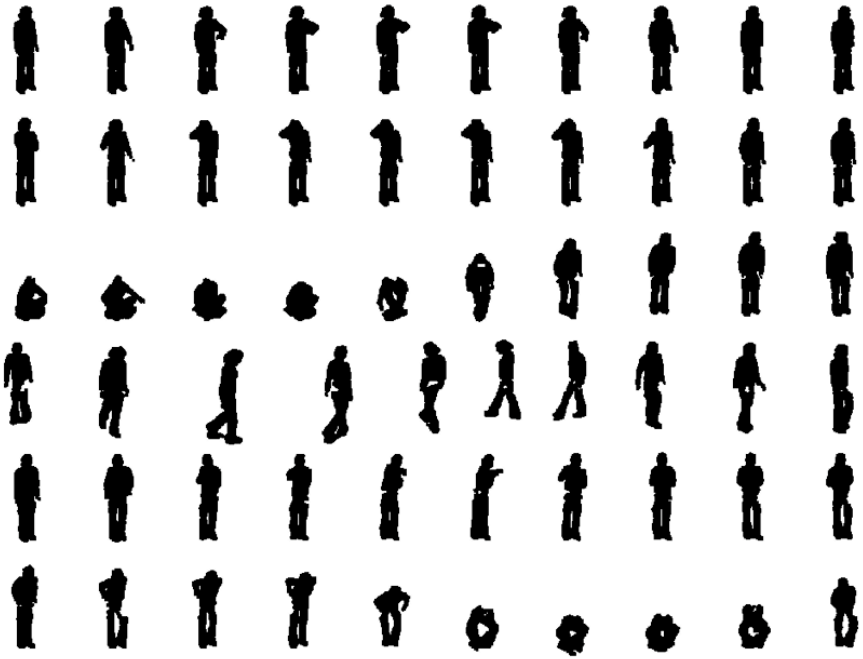
image sequences

observability matrices

check watch

scratch head

...



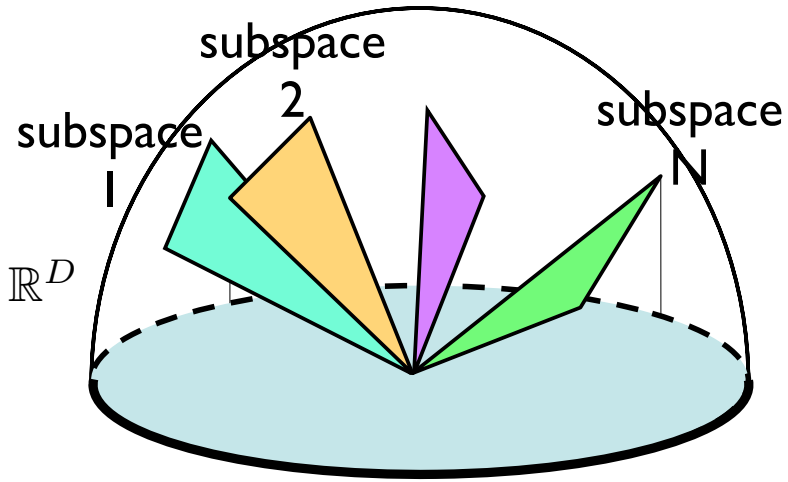
O_{A_1,C_1}

O_{A_2,C_2}

...

Subspace-based learning

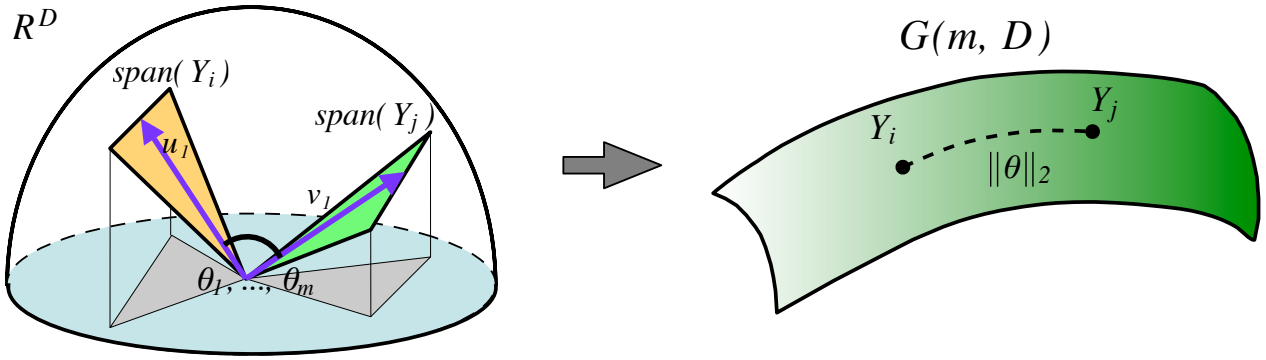
- Assumption: data consists of linear subspaces



  - Model-out the known (undesired) variability with linear subspaces
  - Learn the unknown (interesting) variability *between* subspaces

Framework for subspace-based learning

- The Grassmann manifold $\mathcal{G}(m, D)$ is the set of m-dimensional linear subspaces of the $\mathbb{R}^D$.



- Applications in signal processing and control [Srivastava00,Henkel05,Baumann07], optimization [Edelman99], and computer vision [Liu04,Lin06,Chang06,Turaga08].

Representation

- Quotient space representations of $\mathcal{G}(m, D)$

$$\mathcal{G}(m, D) = \mathcal{O}(D)/\mathcal{O}(m) \times \mathcal{O}(D - m)$$

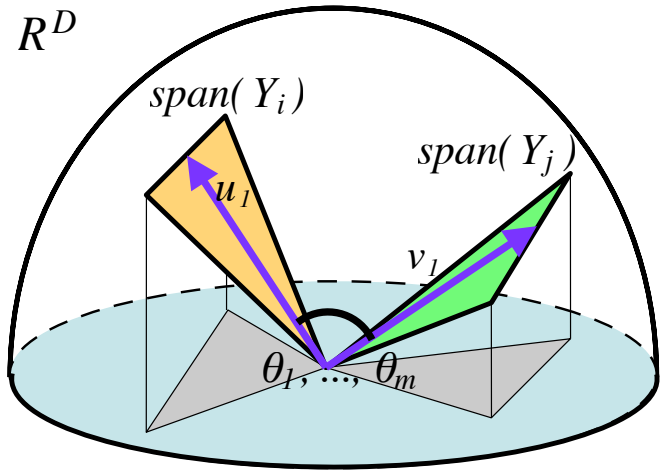
- Basis representation of an element

An element of $\mathcal{G}(m, D)$ is represented by a $D \times m$ matrix such that $Y'Y = I_m$, with the equivalence relation:

$$\begin{aligned} & Y_1 \cong Y_2 \\ \iff \quad & \text{span}(Y_1) = \text{span}(Y_2) \\ \iff \quad & \exists R_m \in \mathcal{O}(m), \text{ such that } Y_1 = Y_2 R_m \end{aligned}$$

Principal angle/canonical corr

Let Y_1 and Y_2 be two orthonormal matrices of size D by m, and let $u \in \mathrm{span}(Y_1)$ and $v \in \mathrm{span}(Y_2)$ be unit vectors.



The first principal angle/canoncial corr between $\mathrm{span}(Y_1)$ and $\mathrm{span}(Y_2)$ is

$$\cos\theta_1 = \max_{u \in \mathrm{span}(Y_1)} \max_{v \in \mathrm{span}(Y_2)} \boldsymbol{u}'\boldsymbol{v}, \quad \text{subject to} \quad \|\boldsymbol{u}\| = \|\boldsymbol{v}\| = 1.$$

k-th principal angle

The k-th principal angle/cannonical correlation is:

$$\cos\theta_k = \max_{\boldsymbol{u_k}\in \mathrm{span}(Y_1)} \max_{\boldsymbol{v_k}\in \mathrm{span}(Y_2)} \boldsymbol{u_k}'\boldsymbol{v_k}, \quad \text{subject to}$$

$$\boldsymbol{u_k}'\boldsymbol{u_k} = 1,\ \boldsymbol{v_k}'\boldsymbol{v_k} = 1,$$

$$\boldsymbol{u_k}'\boldsymbol{u_i} = 0,\ \boldsymbol{v_k}'\boldsymbol{v_i} = 0,\ (i = 1, ..., k-1).$$

- $0 \le \theta_1 \le \cdots \le \theta_m \le \pi/2$ and $1 \ge \cos\theta_1 \ge \cdots \ge \cos\theta_m \ge 0$

- Use SVD for computation: [Golub96]
  $Y_1'Y_2 = USV'$, where $U = [\boldsymbol{u_1}\ ...\ \boldsymbol{u_m}]$, $V = [\boldsymbol{v_1}\ ...\ \boldsymbol{v_m}]$,
  and S is the diagonal matrix $S = \mathrm{diag}(\cos\theta_1\ ...\ \cos\theta_m)$.

Principal angles and distance

- Given: two subspaces Y_1 and Y_2, and principal angles $\theta_1, \ldots, \theta_m$ from SVD
- How to define a good subspace distance from $\theta_1, \ldots, \theta_m$?
- A canonical distance [Edelman99]
  - Arc-length: $d^2_{\text{Arc}}(Y_1, Y_2) = \sum_i \theta_i^2$
  - Is this the only distance?

Grassmann distances

- Projection distance [Edelmann99,Wang06]

$$d^2_{\text{Proj}}(Y_1, Y_2) = \sum_{i=1}^{m} \sin^2 \theta_i = 2^{-1} \, \|Y_1 Y_1' - Y_2 Y_2'\|_F^2$$

- Binet-Cauchy distance [Wolf03,Vishwanathan04]

$$d^2_{\text{BC}}(Y_1, Y_2) = 1 - \prod_i \cos^2 \theta_i = 1 - \det(Y_1' Y_2)^2$$

  - Martin distance between two ARMA models [Martin00]

- Max Corr, Min Corr, Procrustes1/2

Comparison

	Arc Length	Projection	Binet-Cauchy	
$d^2(Y_1, Y_2)$	$\cdot$	$2^{-1}\|Y_1Y_1' - Y_2Y_2'\|_F^2$	$1 - \det(Y_1'Y_2)^2$	
In terms of θ	$\sum \theta_i^2$	$\sum \sin^2 \theta_i$	$1 - \prod \cos^2 \theta_i$	
Is a metric?	Yes	Yes	Yes	
	Max Corr	Min Corr	Procrustes 1	Procrustes 2
$d^2(Y_1, Y_2)$	$2 - 2\|Y_1'Y_2\|_2^2$	$\|Y_1Y_1' - Y_2Y_2'\|_2^2$	$\|Y_1U - Y_2V\|_F^2$	$\|Y_1U - Y_2V\|_2^2$
In terms of θ	$2\sin^2 \theta_1$	$\sin^2 \theta_m$	$4\sum \sin^2(\theta_i/2)$	$4\sin^2(\theta_m/2)$
Is a metric?	No	Yes	Yes	Yes

- Characteristics
  - $d^2_{\text{MaxCor}} = 2\sin^2 \theta_1$ is a rough measure
  - $d^2_{\text{MinCor}} = 2\sin^2 \theta_m,\ d^2_{\text{Proc2}} = 4\sin^2(\theta_m/2)$ too sensitive
  - $d^2_{\text{Arc}} = \sum \theta_i^2,\ d^2_{\text{Proj}} = \sum \sin^2 \theta_i,\ d^2_{\text{Proc1}} = 4\sum \sin^2(\theta_i/2)$, $d^2_{\text{BC}} = 1 - \prod_{i=1}^m \cos^2 \theta_i$ intermediate

Applications

- Distance-based: e.g., k-NN
  - Mutual Subspace Method [Yamaguchi98]
- Beyond k-NN: subspace-based discriminant analysis
  - Previous methods
    - Constrained Mutual Subspace Method [Fukui03]
    - Discriminant Analysis of Canonical Correlations [Kim06]

Some complications

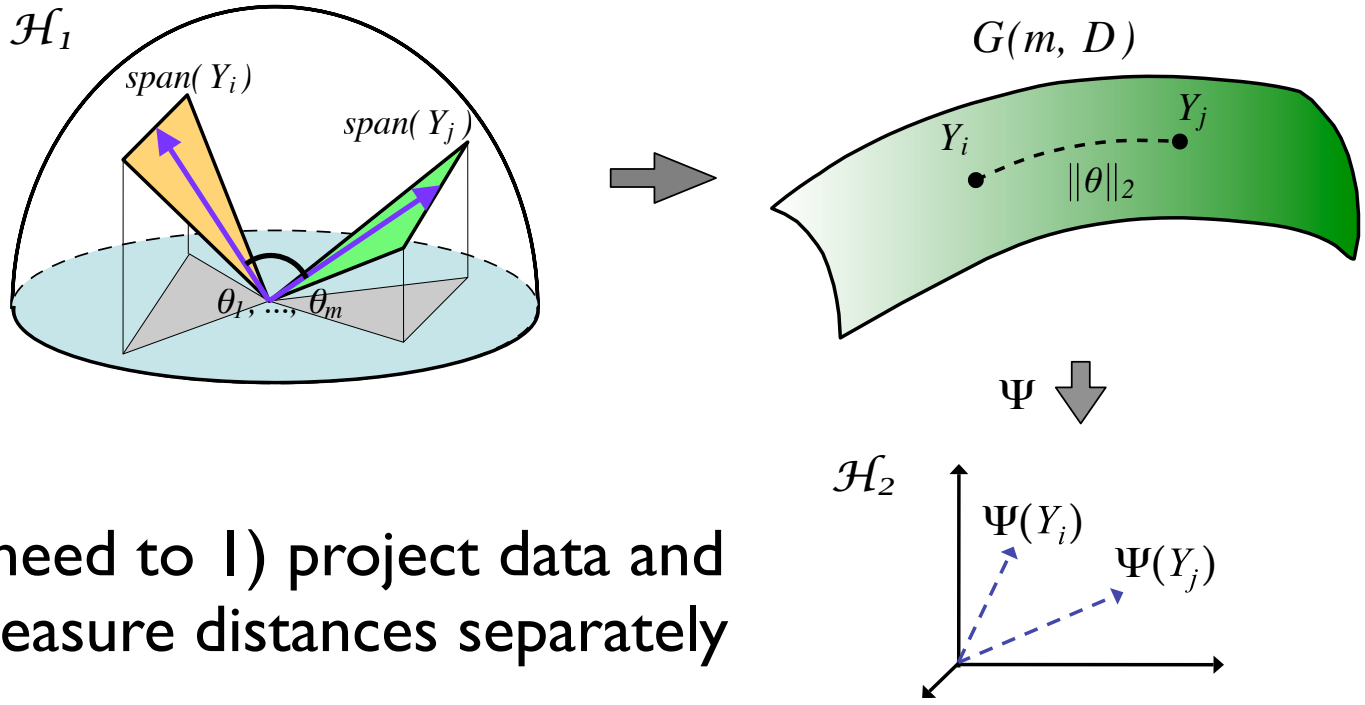
- Find a discriminative direction $w \in \mathbb{R}^D$, so that

$$d(w'X_i, w'X_j) \text{ becomes } \begin{cases} \text{small} & \text{if } y_i = y_j, \\ \text{large} & \text{if } y_i \neq y_j, \end{cases}$$

  - A difficult optimization problem
  - Procedures often iterative, and not well justified
- Inconsistency: projection is performed in image space, but distances are computed on Grassmann space

Easier solution

- Use kernel-induced Hilbert space



No need to 1) project data and
2) measure distances separately

Grassmann kernels

- Let $k : \mathbb{R}^{Dm} \times \mathbb{R}^{Dm} \to \mathbb{R}$ be a real-valued symmetric function, $k(x_1, x_2) = k(x_2, x_1)$
  - Invariance:

    $$k(Y_1, Y_2) = k(Y_1 R_1, Y_2 R_2),\ \forall R_1, R_2 \in \mathcal{O}(m)$$

  - Positive definiteness

$$\sum_{i,j} c_i c_j k(x_i, x_j) \geq 0, \quad \forall (x_1, ..., x_n),\ \forall (c_1, ..., c_n),\ n \in \mathcal{N}$$

- $d_{\text{Proj}},\ d_{\text{BC}}$ have corresponding Grassmann kernels

Projection kernel

- Projection embedding [Chikuse06]

  The map $\Psi : \mathcal{G}(m, D) \to \mathbb{R}^{D \times D}, \quad \mathrm{span}(Y) \mapsto YY'$
  is an isometric embedding from $(\mathcal{G}, d_{\mathrm{Proj}})$ to $(\mathbb{R}^{D \times D}, \|\cdot\|_F)$.

  - Natural inner product in $\mathbb{R}^{D \times D}$: $\mathrm{tr}(Y_1 Y_1' Y_2 Y_2')$

- Projection kernel

  - $k_{\mathrm{Proj}}(Y_1, Y_2) = \mathrm{tr}(Y_1 Y_1' Y_2 Y_2') = \|Y_1' Y_2\|_F^2$ is a Grassmann kernel

  - Has a very simple form and requires only $O(Dm)$ multiplications to evaluate.

Binet-Cauchy kernel

- **Binet-Cauchy identity** [Horn85]

  Suppose we choose m rows from a $D \times m$ matrix A.
  Then, there are $n = {}_D C_m$ square subsmatrices $A^{(s_1)}, ..., A^{(s_n)}$

  $$\det(A'B) = \sum_s \det A^{(s)} \det B^{(s)}$$

- **Binet-Cauchy embedding**

$\Psi : \mathcal{G}(m, D) \to \mathbb{R}^n, \quad \mathrm{span}(Y) \mapsto \left(\det Y^{(s_1)}, ..., \det Y^{(s_n)}\right)$
is an embedding. It is also an isometry from $(\mathcal{G}, d_{\mathrm{BC}})$ to $(\mathbb{R}^n, \|\cdot\|_2)$.

- **Binet-Cauchy kernel** [Wolf03,Vishwanathan04]

  - $k_{\mathrm{BC}}(Y_1, Y_2) = (\det Y_1' Y_2)^2$ **is a Grassmann kernel**

Advantages of Grassmann kernel

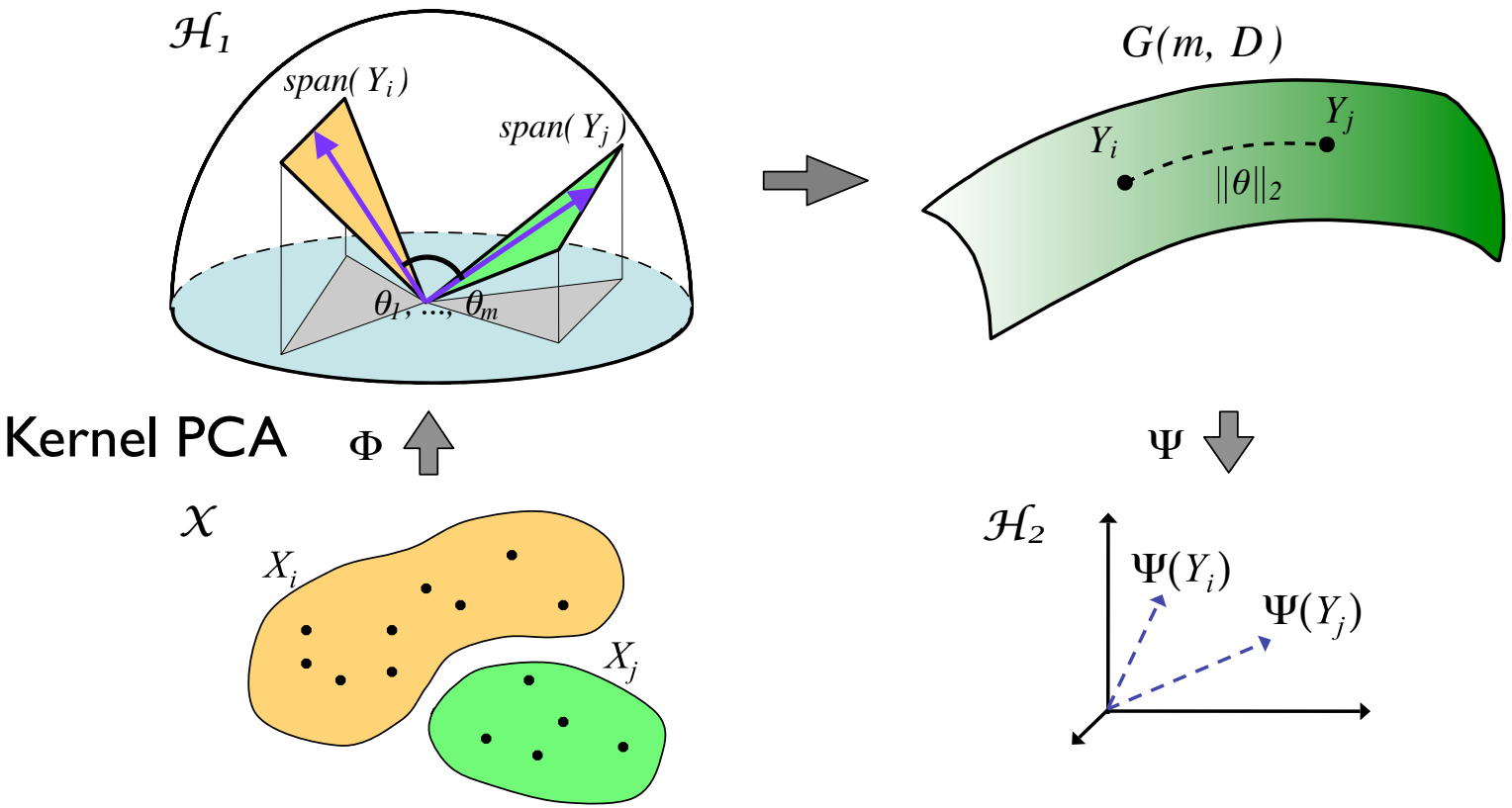
- Access to all the kernel-based algorithms for Hilbert spaces!
- Can generate a family of kernels:

  If $k_1(x, y)$ and $k_2(x, y)$ are PD kernels, then so are

  1. $\alpha_1 k_1(x, y) + \alpha_2 k_2(x, y),\ (\alpha_1, \alpha_2 > 0)$,
  2. $k_1(x, y) k_2(x, y)$,
  3. $\int k_1(x, z) k_1(y, z)\ dz$,
  4. $f(x) k_1(x, y) f(y)$

Extension to nonlinear subspace

- `Doubly kernel' method [Wolf03,Wang06]



Kernel Fisher Discriminant Analysis

Training:

1. *Compute the matrix* $[K_{\text{train}}]_{ij} = k_P(Y_i, Y_j)$ *or* $k_{BC}(Y_i, Y_j)$ *for all* Y_i, Y_j *in the training set.*
2. *Solve* $\max_{\boldsymbol{\alpha}} L(\boldsymbol{\alpha})$ *by eigen-decomposition.*
3. *Compute the* $(C-1)$*-dimensional coefficients* $F_{\text{train}} = \boldsymbol{\alpha}' K_{\text{train}}$.

Testing:

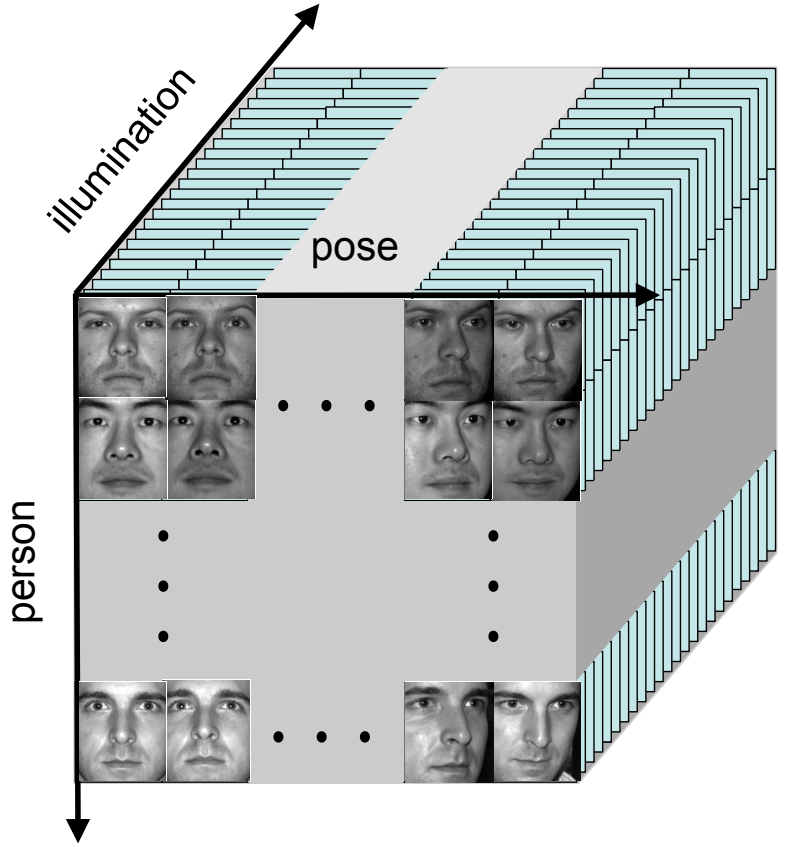
1. *Compute the matrix* $[K_{\text{test}}]_{ij} = k_P(Y_i, Y_j)$ *or* $k_{BC}(Y_i, Y_j)$ *for all* Y_i *in training set and* Y_j *in the test set.*
2. *Compute the* $(C-1)$*-dim coefficients* $F_{\text{test}} = \boldsymbol{\alpha}' K_{\text{test}}$.
3. *Perform 1-NN classification from the Euclidean distance between* F_{train} *and* F_{test}.

Discriminant Analysis Algorithms

- Baseline: Euclidean FDA
- Grassmann Discriminant Analysis (GDA):
  - kernel FDA + Proj / BC kernel
- Others
  - MSM : no dim reduction + MaxCor [Yamaguchi98]
  - cMSM : heur dim reduction + MaxCor [Fukui03]
  - DCC : iterating between 1.NDA 2. Proc1 [Kim07]

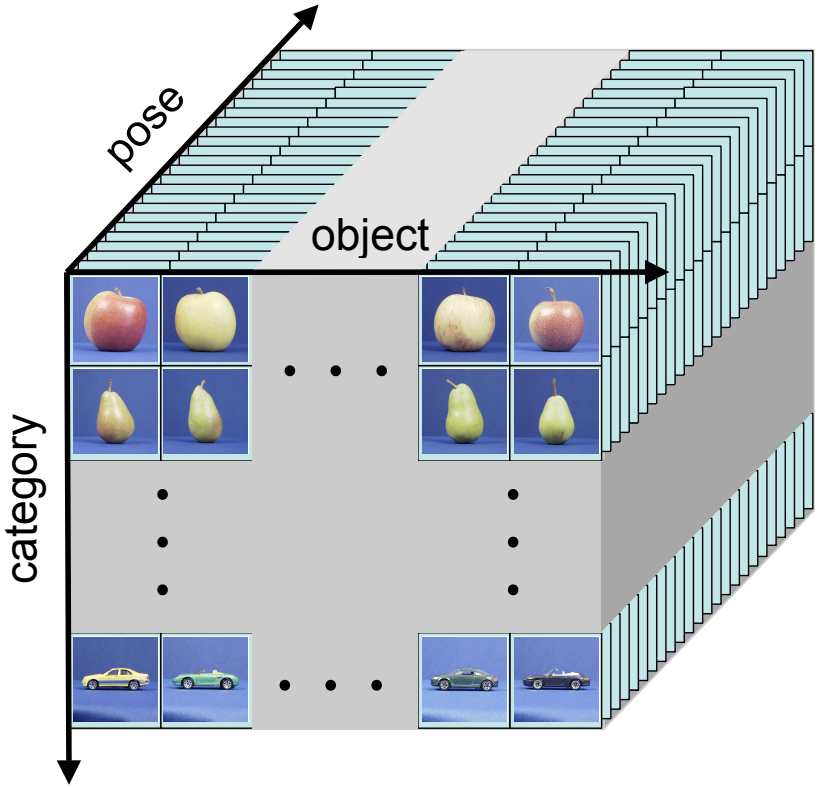
Illum-invariant face recognition

- Yale face database [Georghiades01]
  - 38 persons x 9 poses x 45 illums
  - PCA along illum axis
  - 9-fold cross validation
  - Illum-invariant face recognition
- CMU-PIE database [Sim03]
  - 68 persons x 7 poses x 43 illums
  - 7-fold cross validation



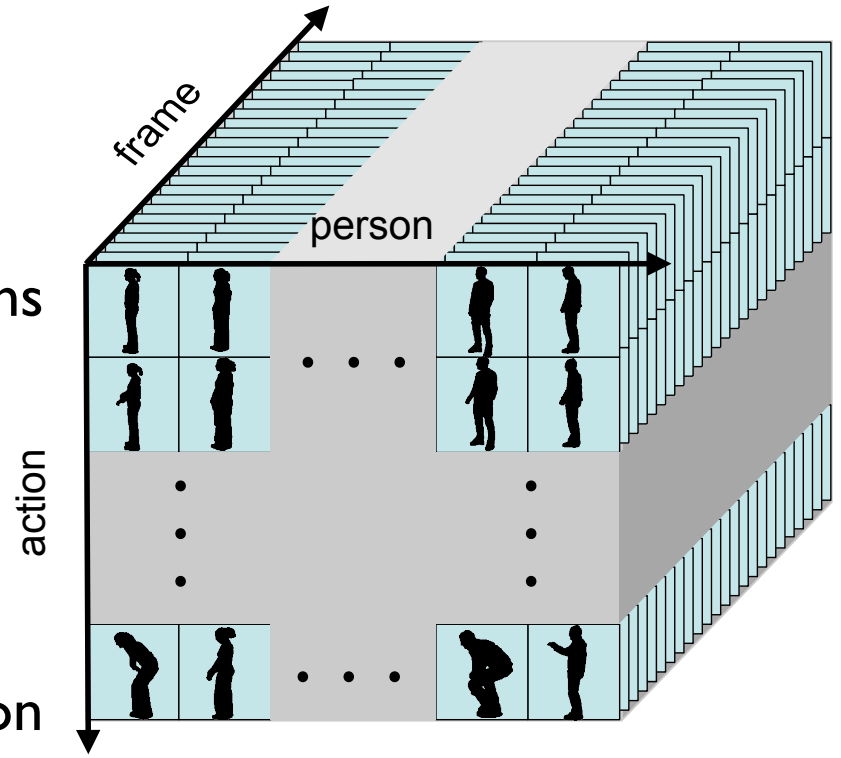
Pose-inv. object categorization

- ETH-80 database [Leibe02]
  - 8 cats x 10 objects x 41 poses
  - PCA along pose axis
  - 10-fold cross validation
  - Pose-invariant object categorization

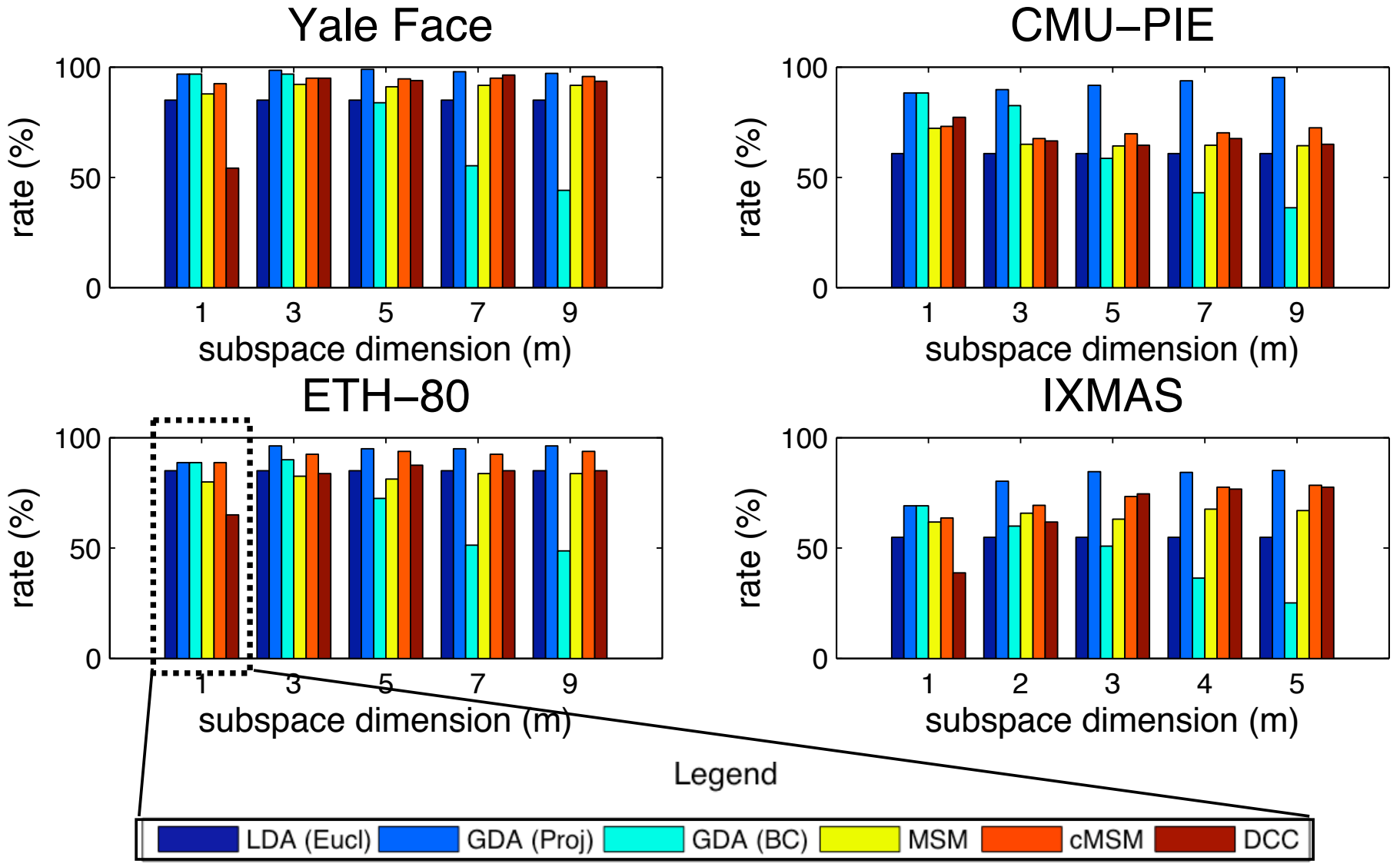


Video-based action recognition

- IXMAS database [Weinland06]
  - 11 actions x 11 actors x T frms (x 3 trials)
  - ARMA model with T frms
  - 11-fold cross validation
  - Video-based action recognition



Results


Yale Face
CMU–PIE
ETH–80
IXMAS
rate (%)
100
50
0
1
3
5
7
9
2
4
subspace dimension (m)
Legend
LDA (Eucl)
GDA (Proj)
GDA (BC)
MSM
cMSM
DCC

Conclusion

- Subspace-based learning: new paradigm for exploiting inherent linear structures in data
- Grassmann manifold as a framework: Projection distances and kernels
- Experiments: superior classification performance with proposed method
- Not limited to image data/FDA method/ classification task
- An open question