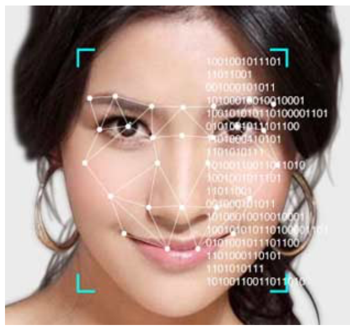
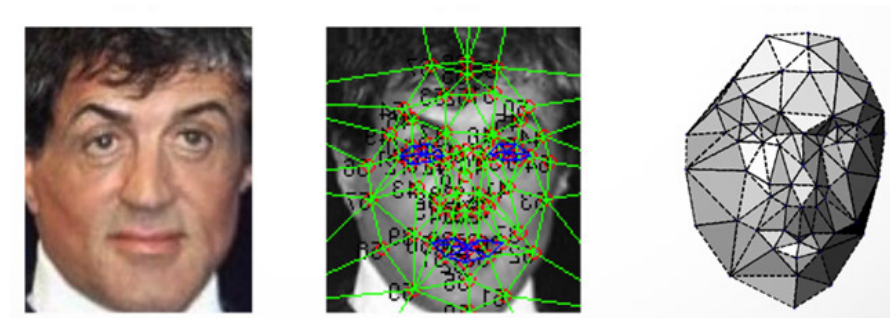
Face Recognition using Multi-modal Low-rank Dictionary Learning

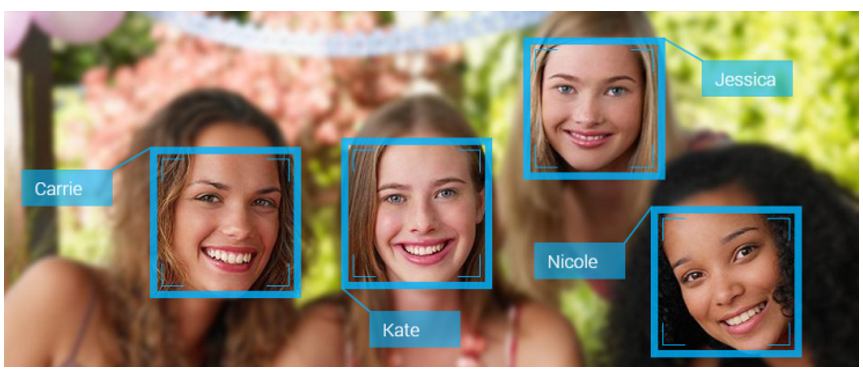
Homa Foroughi, **Moein Shakeri**, Nilanjan Ray, Hong Zhang

Department of Computing Science
University of Alberta
Edmonton, Canada

Face Recognition

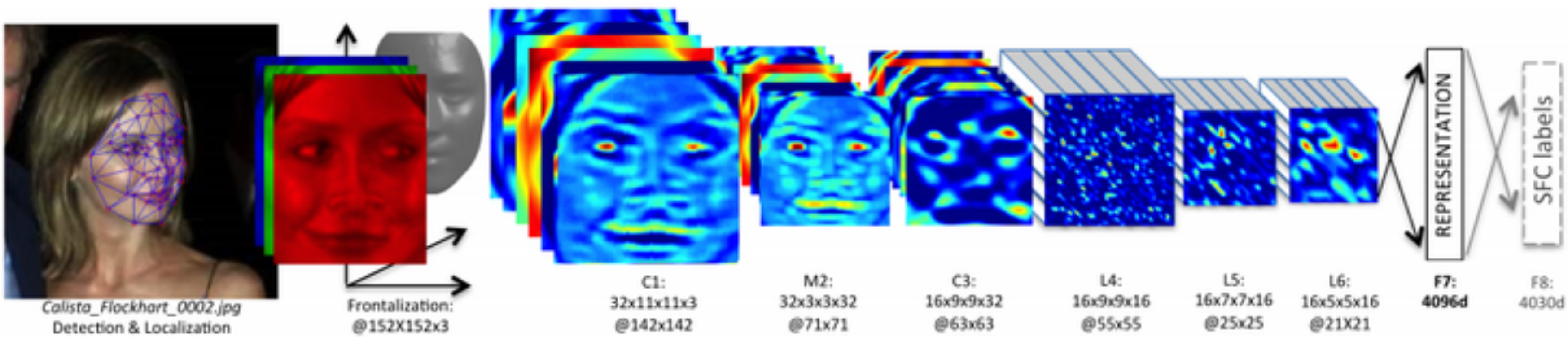




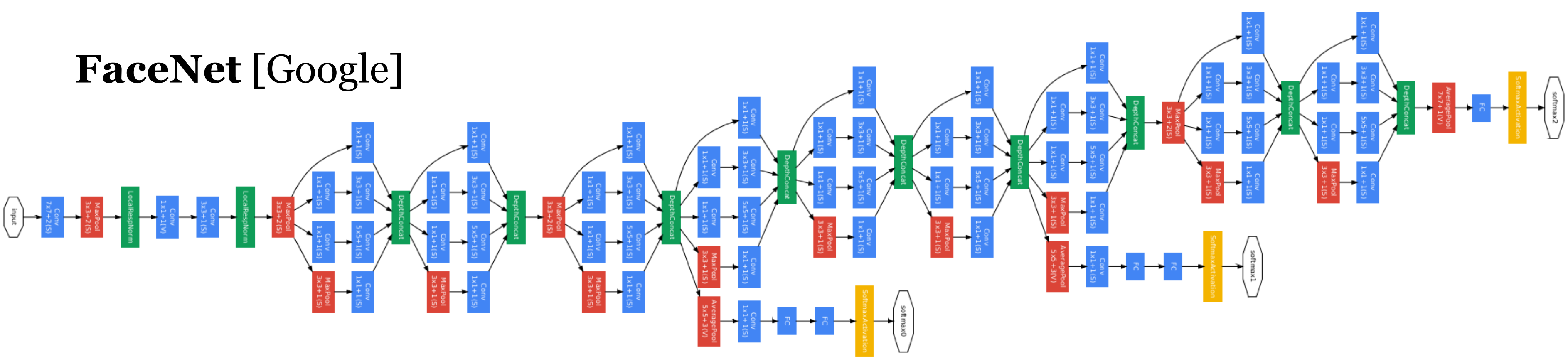


The Winner?!

DeepFace[Facebook]



FaceNet [Google]



Success of CNNs

Dataset	# Images	# Identities
MegaFace	1.02 M	690 K
VGGFace (Oxford)	2.6 M	2,622
DeepFace (Facebook)	4.4 M	4,030
FaceNet (Google)	200 M	8 M
LFW	**13,233**	**5,749**

- Data Augmentation
- Transfer Learning
  - CNN as a fixed feature extractor
  - Fine-tune CNN

Yosinski, Jason, et al. "How transferable are features in deep neural networks?" NIPS. '14

Face Recognition

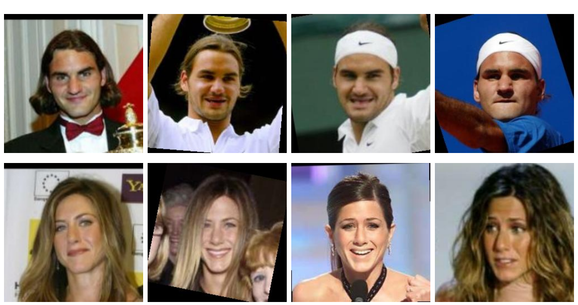
- ❖ 38 subjects
- ❖ Training: 20 out of 64 images per class

Large intra-class variation

- ❖ 100 subjects
- ❖ Training: 8 out of 26 images per class

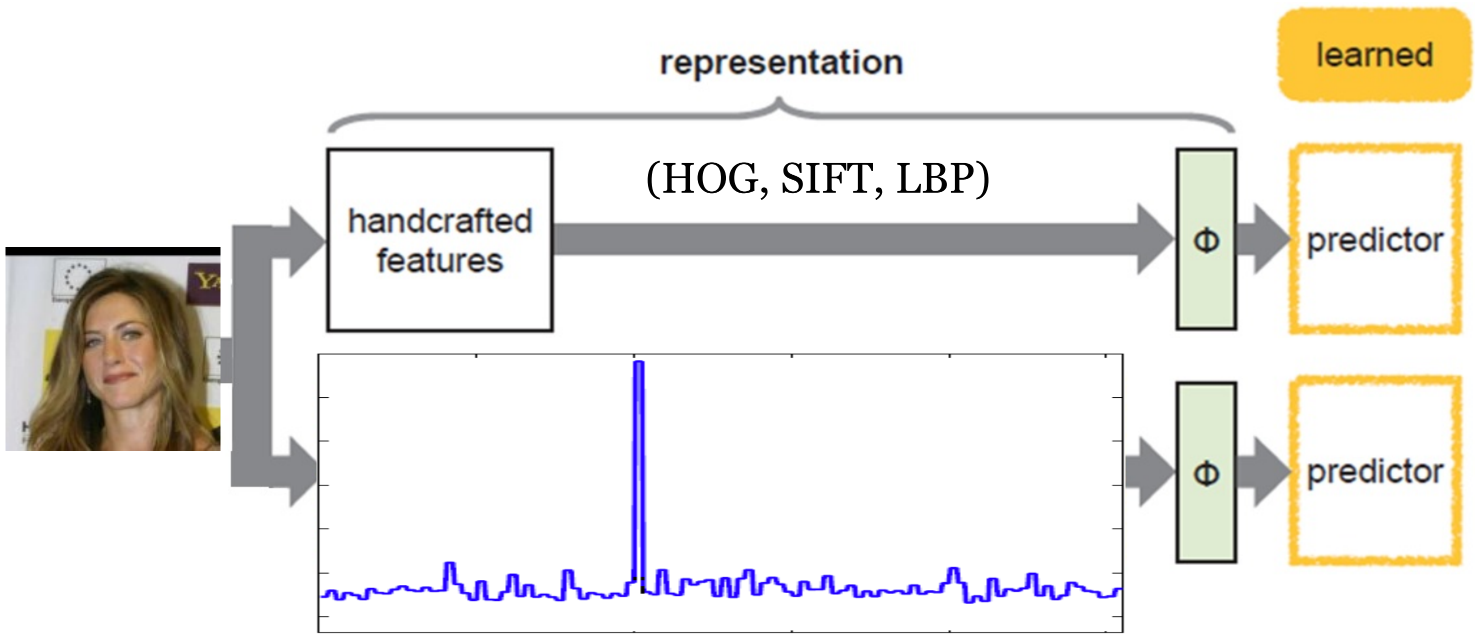
Small-sized training set

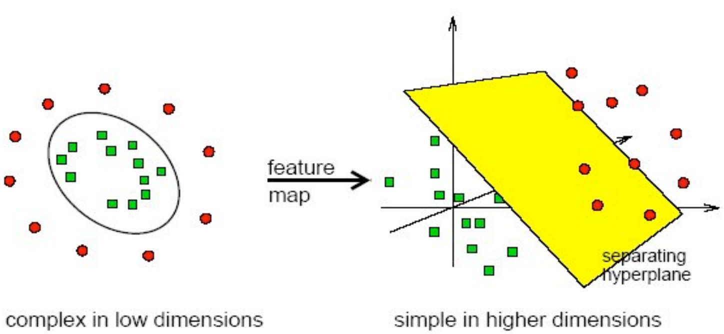
Low-rank Dictionary Learning

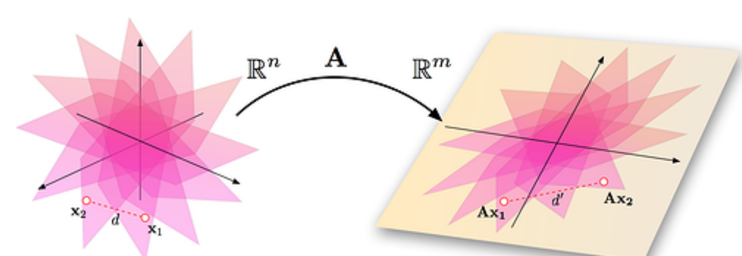


- ❖ 143 subjects
- ❖ Training: 10 out of >11 images per class
  (Min: 11, Max: 500)

Shallow Methods for Face Recognition





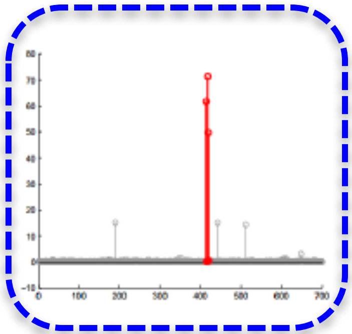


$$d^2(1-\delta_{2k}) \leq (d')^2 \leq d^2(1+\delta_{2k})$$

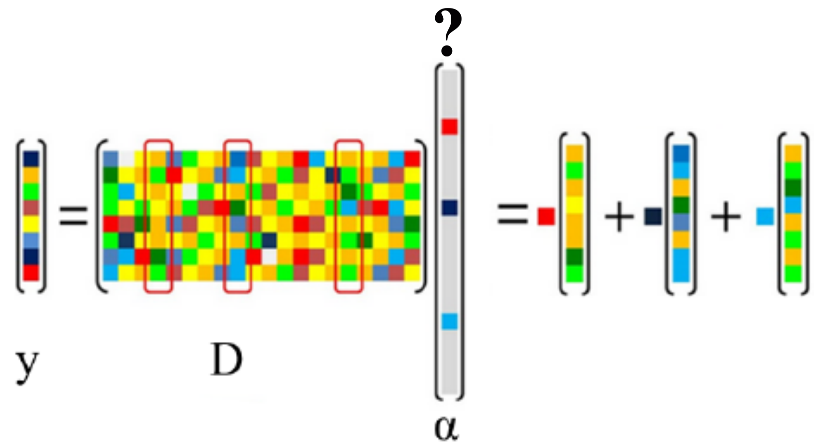
Sparse Representation Theory

=

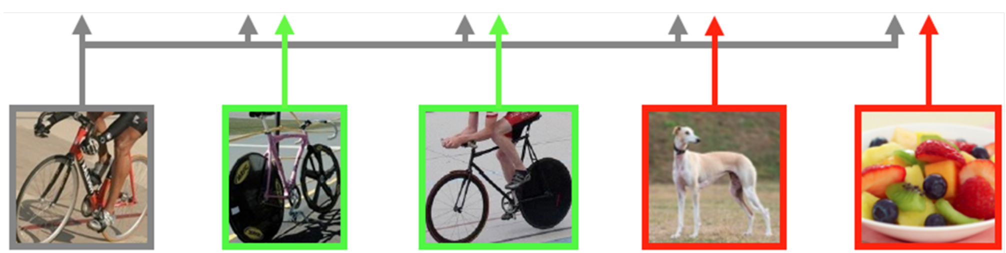
×



Sparse Code α



$$y \;=\; \alpha_1\, D_{:,1} + \;\alpha_2\, D_{:,2} + \;\alpha_3\, D_{:,3} + \;\alpha_4\, D_{:,4} + \cdots$$

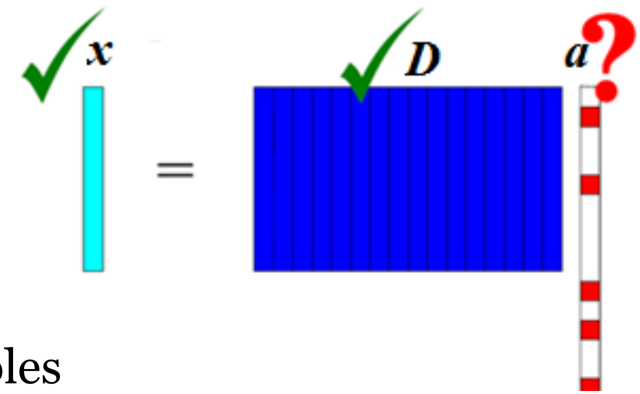


- ❖ Simple
- ❖ Succinct
- ❖ Rich

Across-class representation : Collaboration & Competition

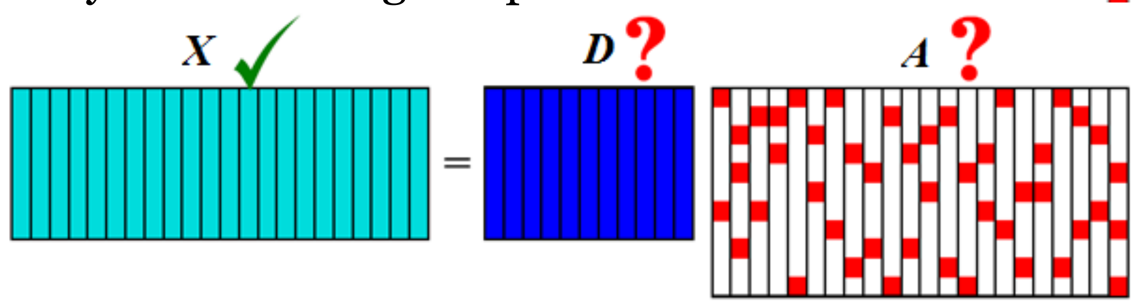
Building Dictionary

- Naïve way : Using training samples as dictionary without any training



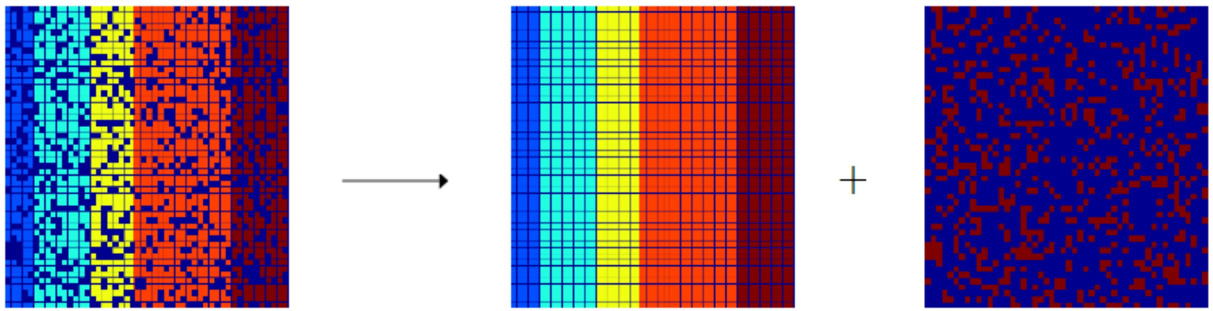
- Better way : Learning dictionary from training samples

Dictionary Learning (DL)



$$< D, A >= \underset{D,A}{\operatorname{argmin}} \|X - D\,A\|_F^2 \quad s.t. \quad \begin{cases} Prior(A) \\ Prior(D) \\ \|a_i\|_1 \leq \epsilon \quad \forall i \end{cases}$$

Low-rank Matrix Recovery

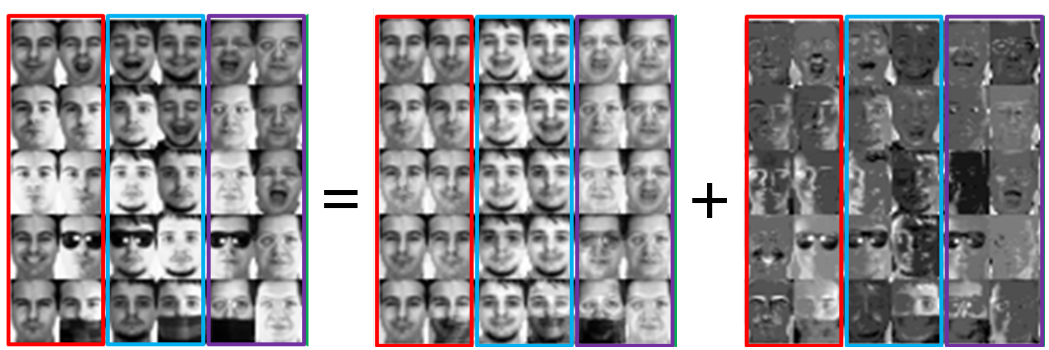


Matrix of corrupted observations *Underlying low-rank matrix* *Sparse error matrix*

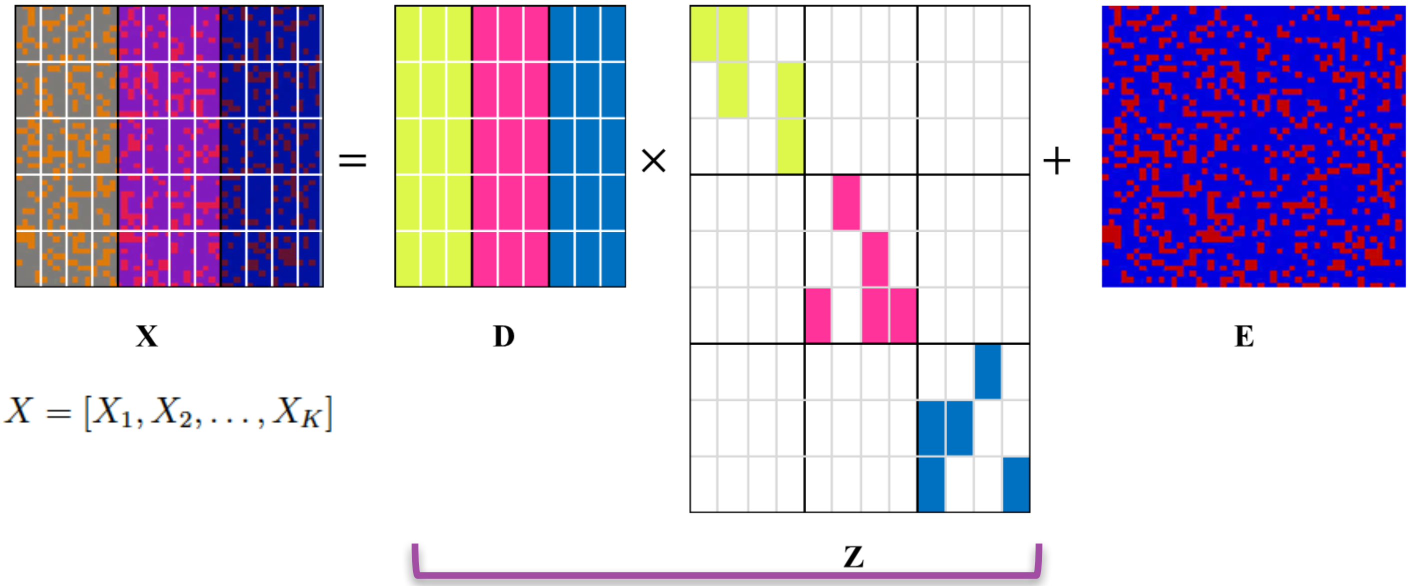
$X = [X_1, X_2, \ldots, X_K] \in R^{m \times N}$

$$\min_{L_i, E_i} \|L_i\|_* + \lambda \|E_i\|_1 \quad s.t. \quad X_i = L_i + E_i \quad \forall i = 1 \ldots K$$

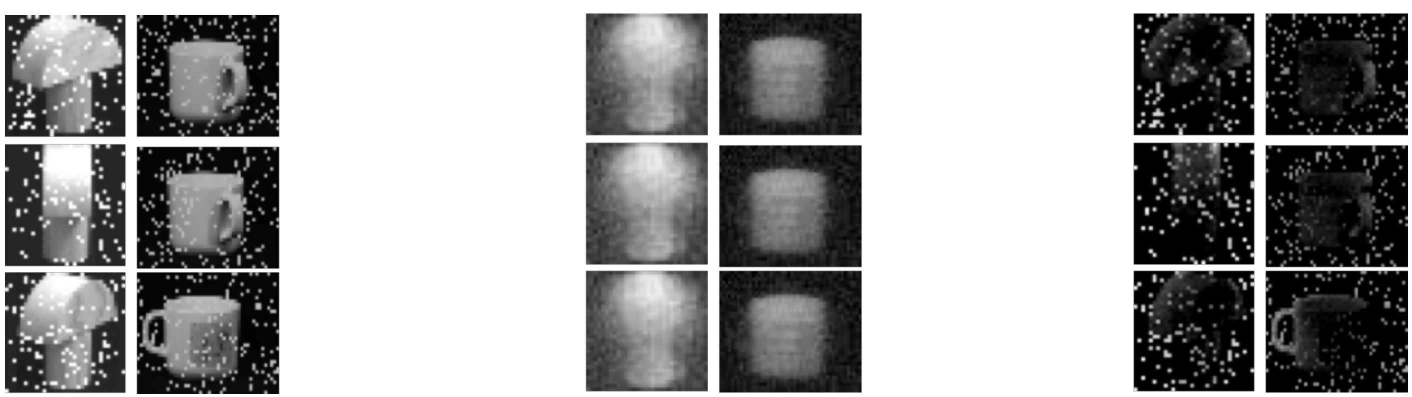
X = L + E



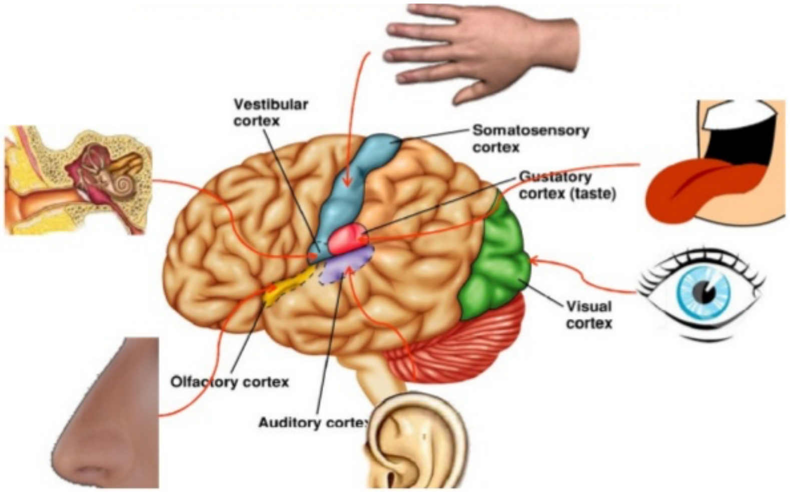
Low-rank Dictionary Learning

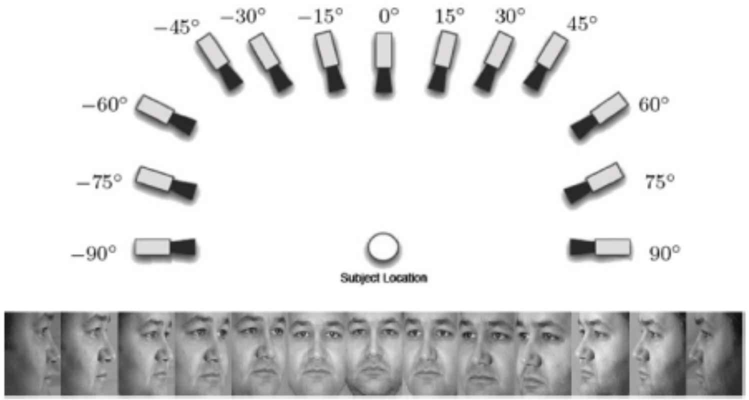


$X = [X_1, X_2, \ldots, X_K]$

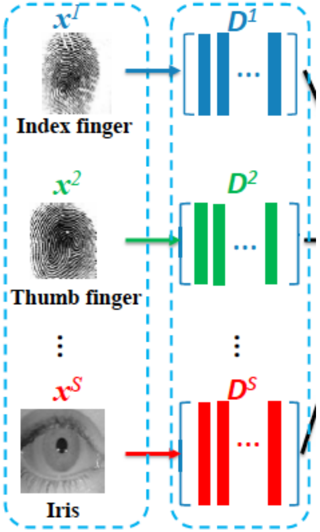


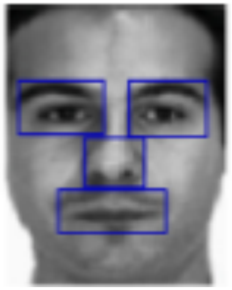
Information Fusion





Multi-Modal Dictionary Learning





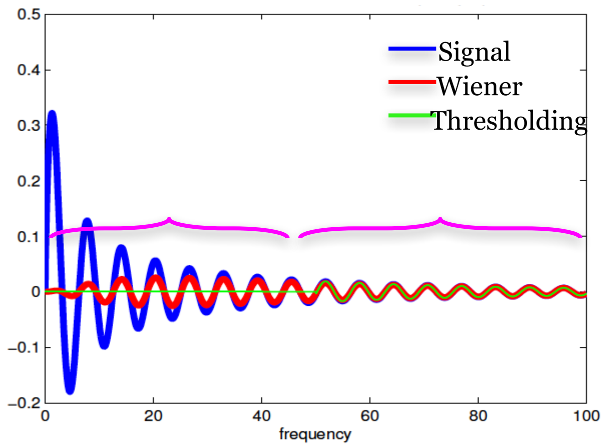
- **Goals:**
  - No overhead, Second modality derived from first one
  - Applicable on single-modality captured data
  - Improve recognition rate

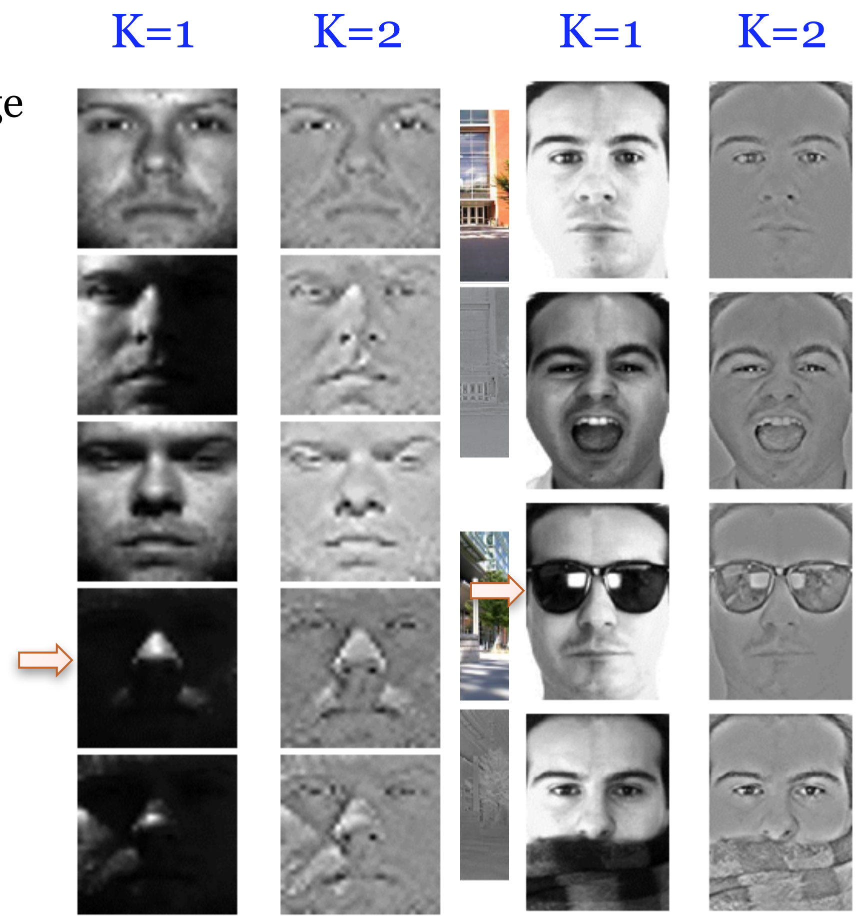
Our Modalities

- ✧ Two modalities
  - ○ K=1; Gray-scale image
  - ○ K=2; Illumination invariant representation of image

$$I(x,y) = R(x,y)L(x,y)$$

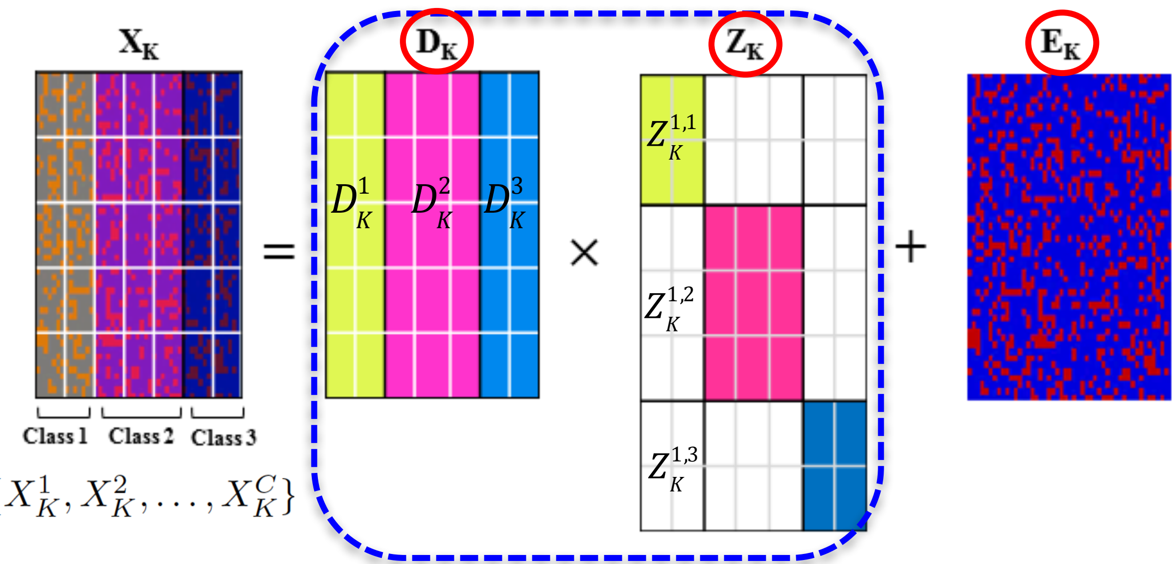
Reflectance Luminance





Moein Shakeri, et al. "Illumination invariant representation of natural images for visual place recognition". IROS 16

Multi-Modal Structured Low-rank DL



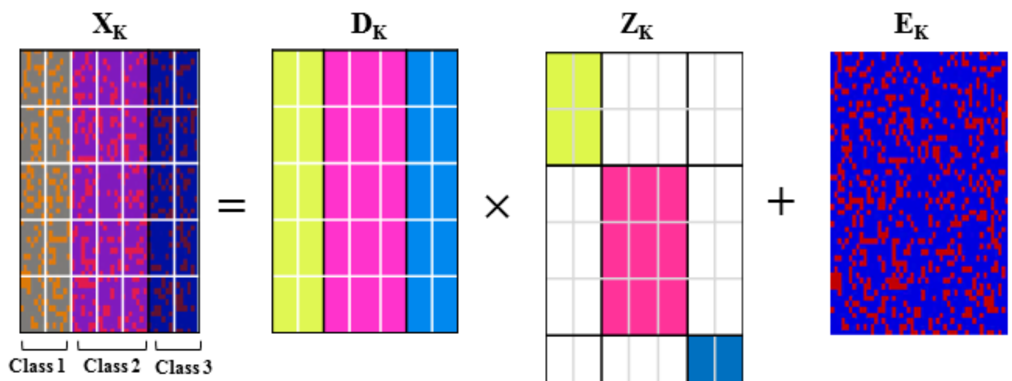
$$X_K = \{X_K^1, X_K^2, \ldots, X_K^C\}$$

- Dictionary: Discriminative & Reconstructive
- Noise: Sparse
- Coding coefficients: Sparse, Block-diagonal, Low-rank

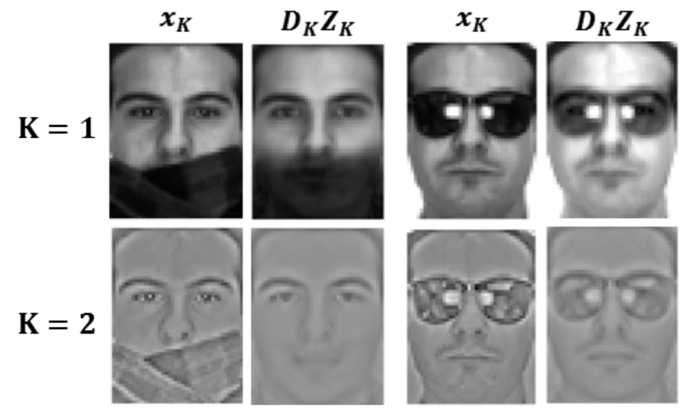
$$Z_K^* = \begin{bmatrix} {Z^*}_K^1 & 0 & 0 & 0 \\ 0 & {Z^*}_K^2 & 0 & 0 \\ 0 & 0 & \ldots & 0 \\ 0 & 0 & 0 & {Z^*}_K^C \end{bmatrix}$$

$$Z_K = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 \end{bmatrix}$$

Multi-Modal Structured Low-rank DL



$$X_K = \{X_K^1, X_K^2, \ldots, X_K^C\}$$



$$\min_{D_K, Z_K, E_K} \sum_{K=1}^{2} (\|Z_K\|_* + \beta\|Z_K\|_1 + \lambda\|E_K\|_1) + \alpha\|Z_1 Z_2^T - Q\|_F^2$$

$$s.t. \quad X_K = D_K Z_K + E_K \quad K = 1, 2$$

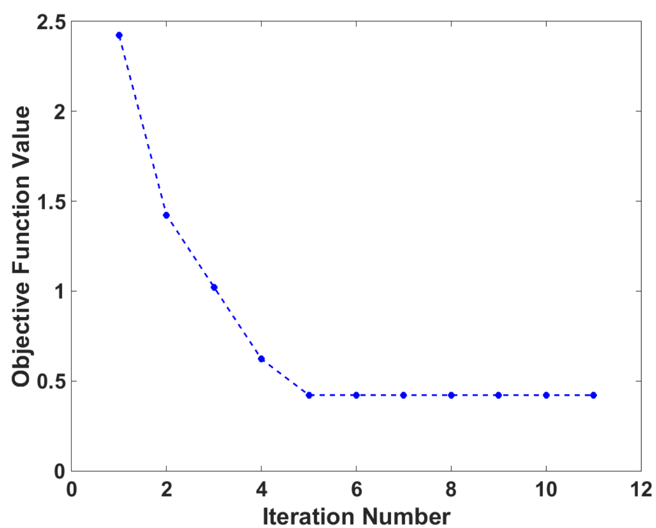
Q: ideal representation

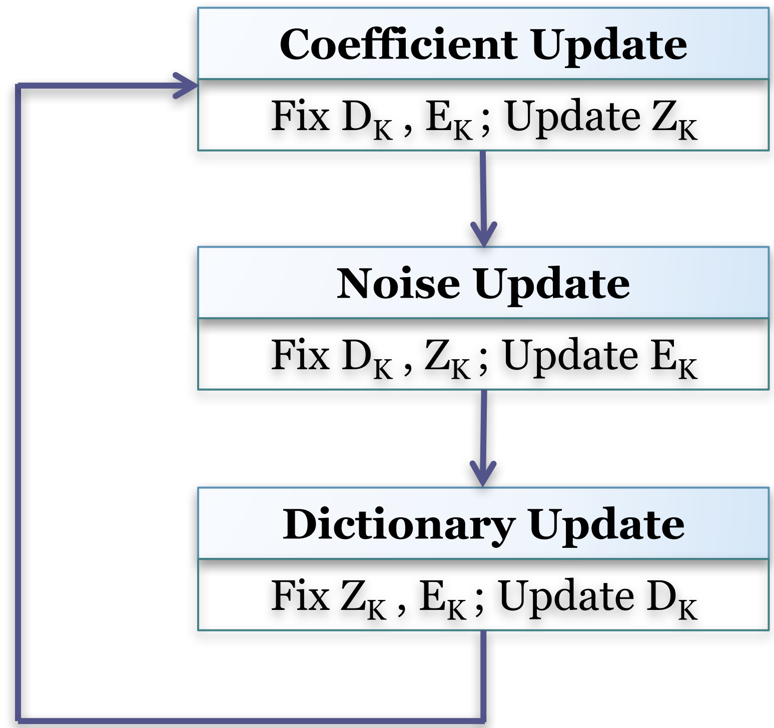
$$\begin{bmatrix} p_1 & p_1 & 0 & 0 & 0 & 0 & 0 \\ p_1 & p_1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & p_2 & p_2 & p_2 & 0 & 0 \\ 0 & 0 & p_2 & p_2 & p_2 & 0 & 0 \\ 0 & 0 & p_2 & p_2 & p_2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & p_3 & p_3 \\ 0 & 0 & 0 & 0 & 0 & p_3 & p_3 \end{bmatrix}$$

- Training images of the same class have the same representation code
- Collaboration of two modalities; affect on each other

Optimization of MM-SLDL

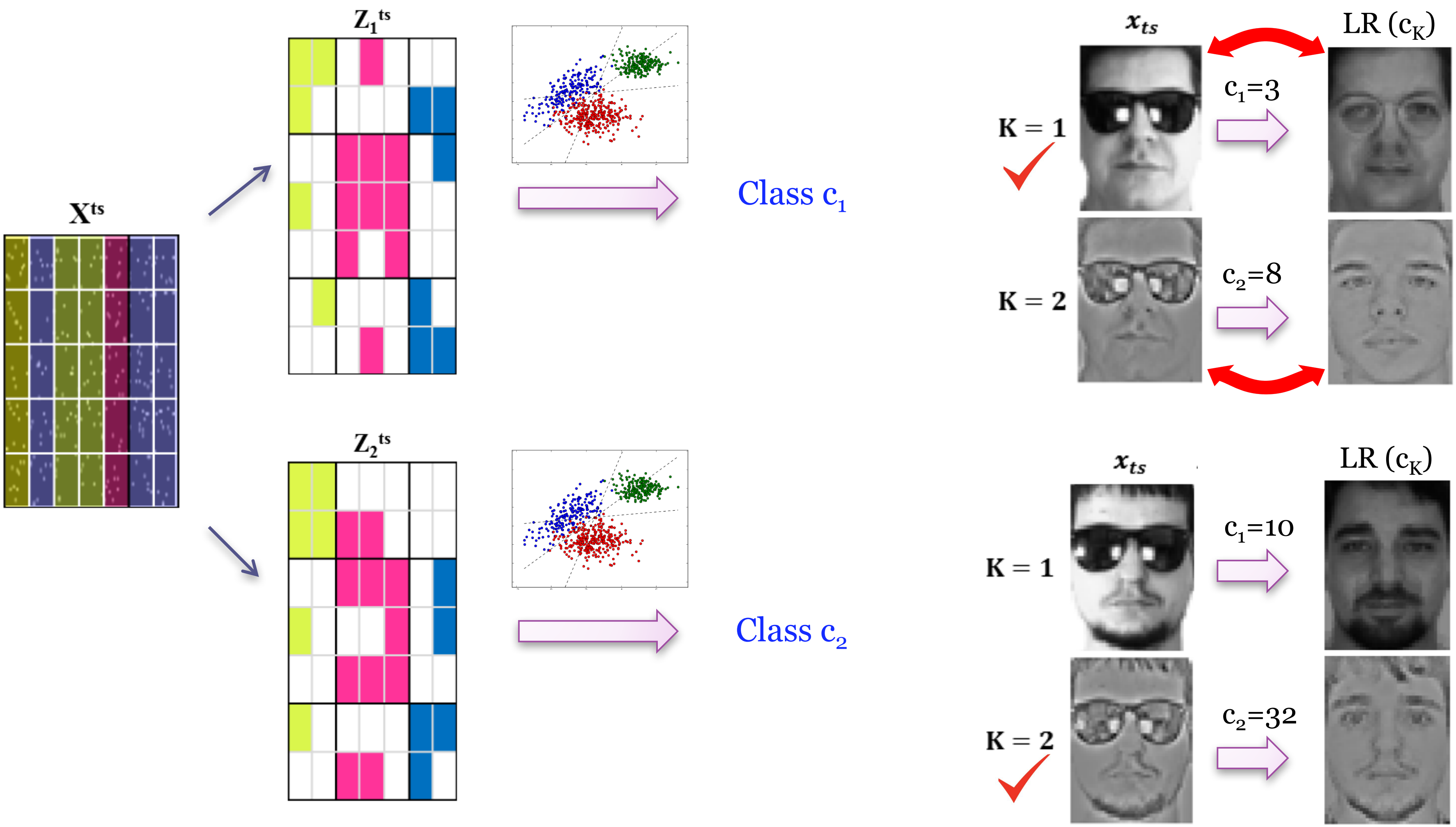
- Convexity





Inexact ALM

Find Winner Modality for Classification

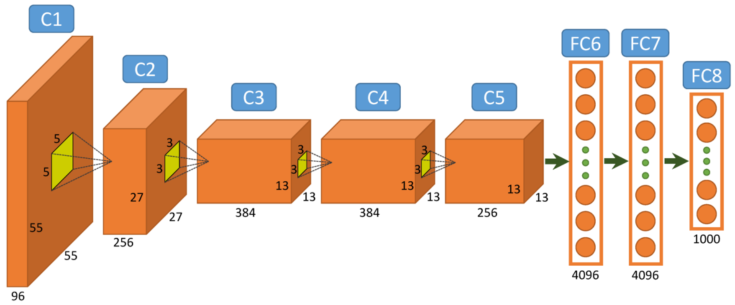


Competitors

- ✧ Shallow methods:
  - Single-modality low-rank dictionary learning
  - Multi-modal dictionary learning
- ✧ Deep methods:

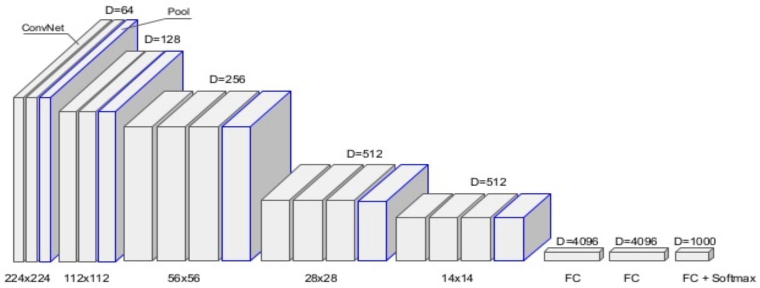
Fine-tuned

Alex-Net: trained on 1.2M natural images



Extract CCN features + NN Classifier

VGG-Face: trained on 2.6M face images



Uncontrolled Face Recognition - LFW

- Subset: 143 subjects with >11 images per class, First 10 for train
- 65×40 images = 2600 features
- Occlusion, variations in pose, expression, illumination, clothing

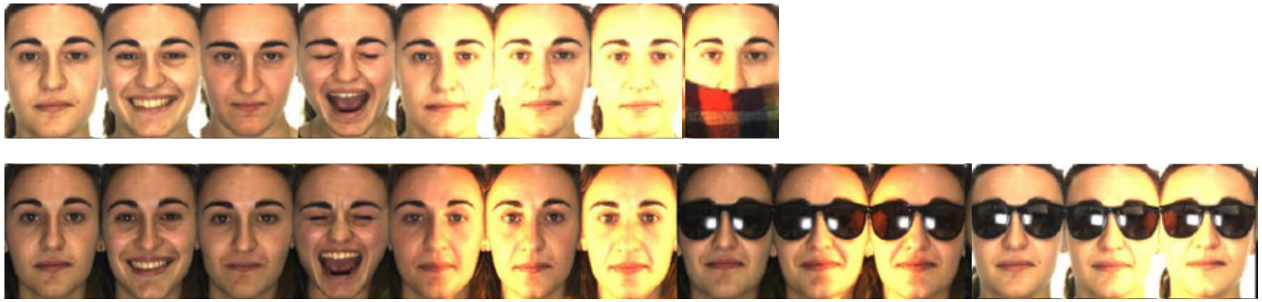
Method	Rec. Rate	Method	Rec. Rate
MLDL [9]	74.10	UMD^{2}L [8]	70.43
MSDL [15]	64.25	D^{2}L^{2}R^2 [4]	75.20
SLRDL [5]	74.20	JP-LRDL [3]	79.87
AlexNet [16]	40.31	VGG-Face [17]	90.01
SLDL-Mod2	76.77	**MM-SLDL**	**88.04**

Controlled Face Recognition - AR

- 100 subjects
- 55*40 images =2200 features
- 8 out of 26 images per class for train

Train

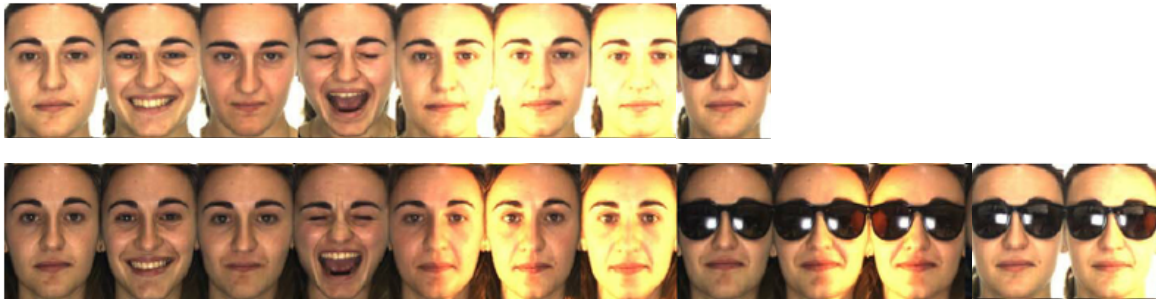
Test



Method	Sunglasses
MLDL [9]	90.51
UMD^{2}L [8]	88.26
MSDL [15]	83.20
D^{2}L^{2}R^2 [4]	92.20
SLRDL [5]	87.35
JP-LRDL [3]	93.20
AlexNet [16]	30.33
VGG-Face [17]	85.90
MM-SLDL	**96.70**

Train

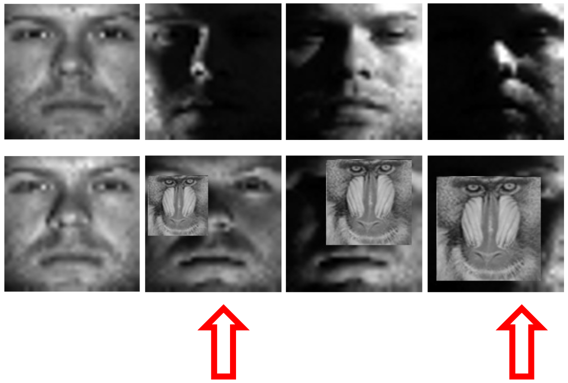
Test

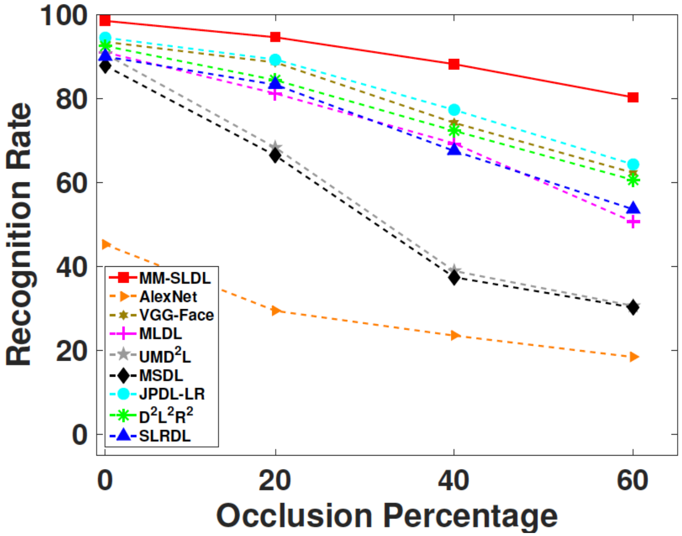


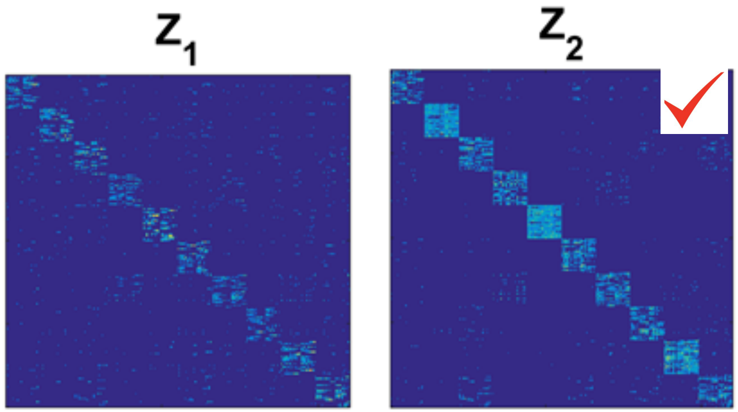
Method	Misc.
MLDL [9]	76.33
UMD^{2}L [8]	71.30
MSDL [15]	68.44
D^{2}L^{2}R^2 [4]	75.30
SLRDL [5]	72.30
JP-LRDL [3]	78.23
AlexNet [16]	25.55
VGG-Face [17]	79.83
MM-SLDL	**85.30**

Controlled Face Recognition – Extended YaleB

- 38 subjects, 55*48 images =2640 features
- 20 out of 64 images per class for train







Conclusions

- ✧ A novel multi-modal structured low-rank dictionary learning method
- ✧ Adopting illumination invariant representation as a modality
- ✧ Applicable to millions of images captured under single-modality
- ✧ Superior performance on small training sets with large intra-class variation

Thank You!