Learning A Discriminative Dictionary for Sparse Representation

Zhuolin Jiang†, Zhe Lin‡, Larry S. Davis†

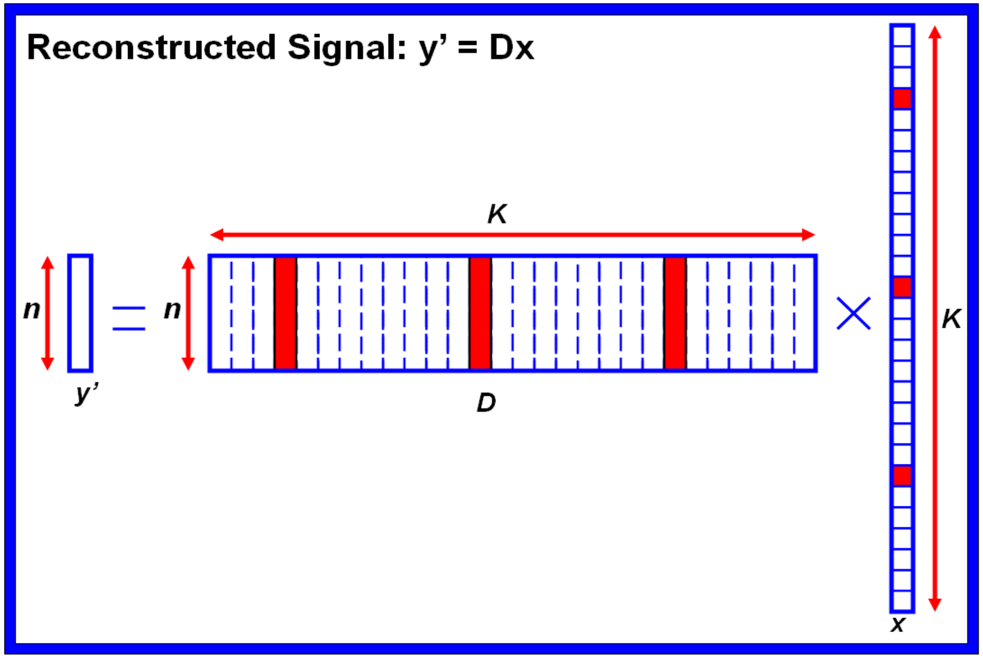
†Computer Vision Laboratory
University of Maryland, College Park
{zhuolin, lsd}@umiacs.umd.edu

‡Advanced Technology Labs
Adobe Systems Incorporated, San Jose
zlin@adobe.com

Sparse Paradigm

- Given a signal $y \in R^n$ and the dictionary $D \in R^{n \times K}$, the sparse representation $x \in R^K$ of y is estimated by:

$$x = \min_{x} \underbrace{\left\{ \|y - Dx\|_F^2 \right\}}_{\text{reconstruction error}} \text{ subject to } \underbrace{\|x\|_0}_{\text{sparsity constraint}} \leq T$$



Dictionary Learning Techniques

- SRC algorithm employs the entire set of training samples to form a dictionary. (face recognition) [Wright, TPAMI2009]
- K-SVD: Efficiently learn an over-complete dictionary with a small size. It focuses on representational power, but does not consider discriminative capability. (restoration, compression) [Aharon, TSP2006]
- An online algorithm for learning dictionaries is proposed and faster than batch alternatives such as K-SVD on large datasets. (restoration) [Mairal, JMLR 2010]
- A tree-structured sparse regularization is proposed to learn a dictionary embedded in a tree efficiently. (restoration) [Jenatton, ICML2010]

Dictionary Learning Techniques

- Discriminative dictionary learning approaches:
  - Constructing a separate dictionary for each class [Mairal CVPR2008]
  - Unifying the dictionary learning and classifier training into a mixed reconstructive and discriminative formulation [Pham,CVPR2008][Zhang,CVPR2010][Mairal,NIPS2009][Yang,CVPR2010]
    - Solve the problem to alternate between the two variables, minimizing over one while keeping the other one fixed
    - Learn simultaneously an over-complete dictionary and multiple classification models for each class

Goals

- Learn a dictionary with representational and discriminative power for sparse representation;
  - Representational power, which helps for achieving lower reconstruction error;
  - Discriminative power, which is good for object classification.
- Learn a universal multiclass linear classifier;
- Develop an efficient way of finding the optimal solution to the dictionary learning problem

Approaches

- A new label consistency constraint called 'discriminative sparse-code error' is introduced and combined with reconstruction error and classification error to form a unified objective function for dictionary learning.

- The optimal solution is efficiently obtained using the K-SVD algorithm.
  - A single compact discriminative dictionary and a universal multiclass linear classier are learned simultaneously.

Dictionary Learning

- Dictionary for Reconstruction and Sparse Coding

  Let Y be a set of n-dim input signals, $Y = [y_1 ... y_N] \in R^{n \times N}$ dictionary *D* in $R^{n\times K}$ $(K>n)$ is learned by:

$$< D, X >= \arg\min_{D,X} \underbrace{\|Y - DX\|_2^2}_{\text{reconstruction error}} \quad \underbrace{s.t. \forall i, \|x_i\|_0 \le T}_{\text{sparsity constraint}}$$

- Sparse Coding

  Given D, the sparse representation *X* in $R^{K\times N}$ of *Y* is:

$$X = \arg\min_{X} \|Y - DX\|_2^2 \quad s.t. \forall i, \|x_i\|_0 \le T$$

Dictionary Learning

- Dictionary Learning for Classification
  - A good classifier f(x) can be obtained by determining its model parameters *W*:

$$W = \arg\min_{W} \sum_i \mathcal{L}\{h_i, f(x_i, W)\} + \lambda_1 \|W\|_F^2$$

  - *D* and *W* can be learned jointly:

$$< D, W, X >= \arg\min_{D,W,X} \|Y - DX\|_2^2$$
$$+ \sum_i \mathcal{L}\{h_i, f(x_i, W)\} + \lambda_1 \|W\|_F^2 \; s.t. \forall i, \|x_i\|_0 \leq T$$

Label-Consistent K-SVD

- LC-KSVD1
  - Objective function

$$< D, A, X > = \arg\min_{D,A,X} \underbrace{\|Y - DX\|_2^2}_{\text{reconstruction error}} + \alpha \underbrace{\|Q - AX\|_2^2}_{\text{discriminative sparse code error}} \; s.t. \forall i, \|x_i\|_0 \leq T$$

where:
A is a linear transformation matrix;
Q are discriminative sparse codes for input training signals Y.

Label-Consistent K-SVD

- LC-KSVD1
  - Objective function

$< D, A, X >=$ … $i, \|x_i\|_0 \leq T$

An example of Q

Assuming D = [$d_1 \ldots d_6$] and Y = [$y_1 \ldots y_6$], where y_1, y_2, d_1 and d_2 are from class1, y_3, y_4, d_3 and d_4 are from class2, y_5, y_6, d_5 and d_6 are from class3, Q can be defined:

$$Q \equiv \begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 \end{bmatrix}$$

where:

A is a linear tra

Q are discrimi

Label-Consistent K-SVD

- LC-KSVD2

- Objective function

$$< D, W, A, X >= \arg \min_{D,W,A,X} \|Y - DX\|_2^2$$
$$+\alpha \underbrace{\|Q - AX\|^2}_{\text{discriminative sparse code error}} + \beta \underbrace{\|H - WX\|_2^2}_{\text{classification error}} \; s.t. \forall i, \|x_i\|_0 \leq T$$

Label-Consistent K-SVD

- LC-KSVD2
  - Objective function

$$< D, W, A, X >= \arg \min_{D,W,A,X} \|Y - DX\|_2^2$$
$$+\alpha\underbrace{\|Q - AX\|^2}_{\text{discriminative sparse code error}} + \beta\underbrace{\|H - WX\|_2^2}_{\text{classification error}} \; s.t. \forall i, \|x_i\|_0 \leq T$$

Assume $X' = AX$, then $D' = DA^{-1}, W' = WT^{-1}$.

The objective function is rewritten as

$$< D', W', X' >= \arg \min_{D',W',X'} \|Y - D'X'\|_2^2$$
$$+\alpha\|Q - X'\|_2^2 + \beta\|H - W'X'\|_2^2 \; s.t. \forall i, \|x_i\|_0 \leq T$$

Label-Consistent K-SVD

- Optimization

We rewrite the objective function of LC-KSVD2 as:

$$< D, W, A, X >= \arg \min_{D,W,A,X} \| \begin{pmatrix} Y \\ \sqrt{\alpha}Q \\ \sqrt{\beta}H \end{pmatrix} - \begin{pmatrix} D \\ \sqrt{\alpha}A \\ \sqrt{\beta}W \end{pmatrix} X \|_2^2 \quad s.t. \forall i, \|x_i\|_0 \leq T$$

Let $D_{new} = (D^t, \sqrt{\alpha}A^t, \sqrt{\beta}W^t)^t$, $Y_{new} = (Y^t, \sqrt{\alpha}Q^t, \sqrt{\beta}H^t)^t$. The optimization is equivalent to

$$< D_{new}, X >= \arg \min_{D_{new},X} \{\|Y_{new} - D_{new}X\|_2^2\} s.t. \forall i, \|x_i\|_0 \leq T$$

- Initialization

D_0: K-SVD is employed within each class and the outputs of each K-SVD are combined;

A_0: $A = (XX^t + \lambda_2 I)^{-1} XQ^t$ W_0: $W = (XX^t + \lambda_1 I)^{-1} XH^t$

Multivariate ridge regression

Label-Consistent K-SVD

- Classification

❑ $\hat{D}, \hat{A}, \hat{W}$

In general, D should be L2-normalized column wise, i. e. $\|(d_k^t, \sqrt{\alpha}a_k^t, \sqrt{\beta}w_k^t)^t\|_2 = 1$

$$\hat{D} = \{\frac{d_1}{\|d_1\|_2}, \frac{d_2}{\|d_2\|_2} \ldots \frac{d_K}{\|d_K\|_2}\}$$
$$\hat{A} = \{\frac{a_1}{\|d_1\|_2}, \frac{a_2}{\|d_2\|_2} \ldots \frac{a_K}{\|d_K\|_2}\}$$
$$\hat{W} = \{\frac{w_1}{\|d_1\|_2}, \frac{w_2}{\|d_2\|_2} \ldots \frac{w_K}{\|d_K\|_2}\}$$

❑ Classification

For a test image y_i, we first compute its sparse representation:

$$x_i = \arg\min_{x_i}\{\|y_i - \hat{D}x_i\|_2^2\} \quad s.t. \|x_i\|_0 \leq T$$

The classification result (i.e. the label j of y_i) is given by: $j = \arg\max_j (l = \hat{W}x_i)$

Experimental Results

- Evaluation Databases
  - Extended Yaleb
  - AR Face
  - Caltech101
- Experimental Setup
  - Random face-based feature
    - dims: 504 (Extended Yale), 540 (AR Face)
  - Spatial pyramid feature
    - 1024 bases
    - dims: 3000 (Caltech101)
    - Max pooling, L2 normalization

Experimental Results

- Extended Yale B dataset
  - 38 persons, 2414 frontal face images
  - Under varying illumination conditions, varying expressions
  - Randomly selected half of the images (training) + the other half (testing).

Experimental Results

- Extended Yale B dataset

Method	Acc.(%)	Acc.(%)
K-SVD(15 per person) [1]	93.1	98.0
D-KSVD(15 per person) [33]	94.1	98.0
SRC(all train. samp.) [28]	**97.2**	**99.0**
SRC*(15 per person) [28]	80.5	86.7
LLC(30 local bases) [27]	82.2	92.1
LLC(70 local bases) [27]	90.7	96.7
LC-KSVD1(15 per person)	94.5	98.3
LC-KSVD2(15 per person)	95.0	98.8
LC-KSVD2(all train. samp.)	96.7	**99.0**

Method	Avg. Time (ms)
SRC(all training samples) [28]	20.78
SRC*(15 per person) [28]	11.22
LC-KSVD1(15 per person)	0.52
LC-KSVD2(15 per person)	0.49

Experimental Results

- AR Face database
  - 2600 images (26 images per person), 50 male and 50 females
  - Different illumination conditions, different expressions and different facial disguises (with sunglasses and scarf respectively)
  - (Randomly selected) 20 images (training) + 6 (testing)

Experimental Results

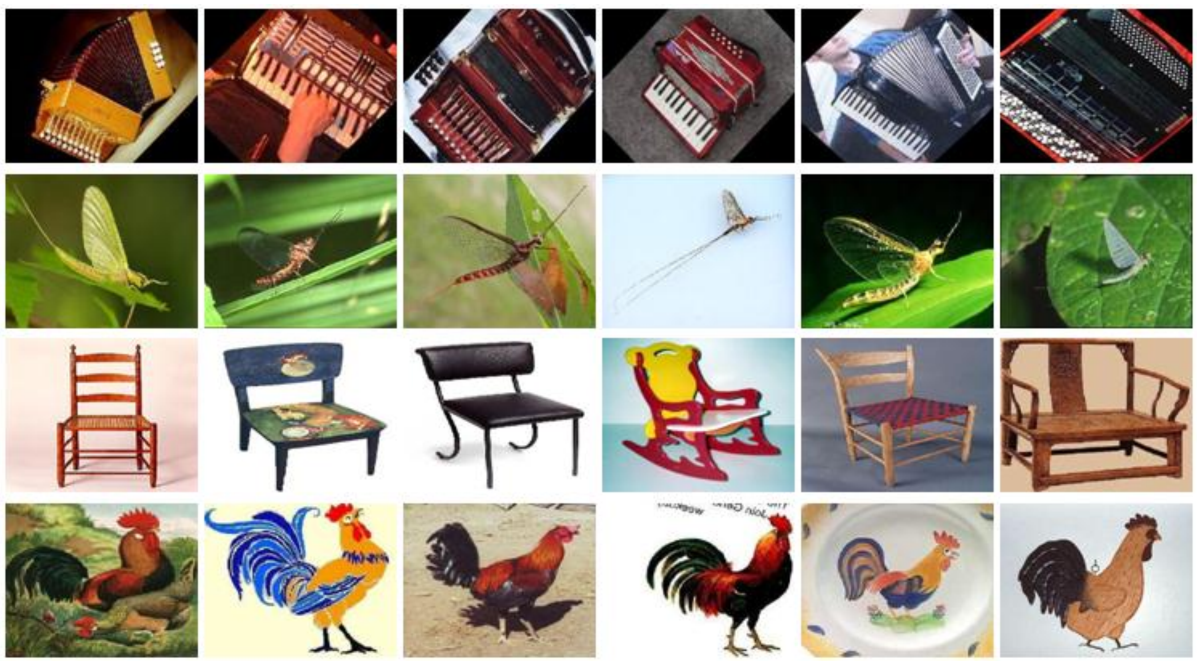
- AR Face database

Method	Acc. (%)
K-SVD(5 per person) [1]	86.5
D-KSVD(5 per person) [33]	88.8
SRC(all train. samp.) [28]	97.5
SRC*(5 per person) [28]	66.5
LLC(30 local bases) [27]	69.5
LLC(70 local bases) [27]	88.7
LC-KSVD1(5 per person)	92.5
LC-KSVD2(5 per person)	93.7
LC-KSVD2(all train. samp.)	**97.8**

Method	Avg. Time (ms)
SRC(all training samples) [28]	83.79
SRC*(5 per person) [28]	17.76
LC-KSVD1(5 per person)	0.541
LC-KSVD2(5 per person)	0.479

Experimental Results

- Caltech-101 Object Dataset
  - 9144 images, 102 classes (101 object classes and a ‘background’ class)
  - The number of images per category varies from 31 to 800
  - Significant variance in shape



Experimental Results

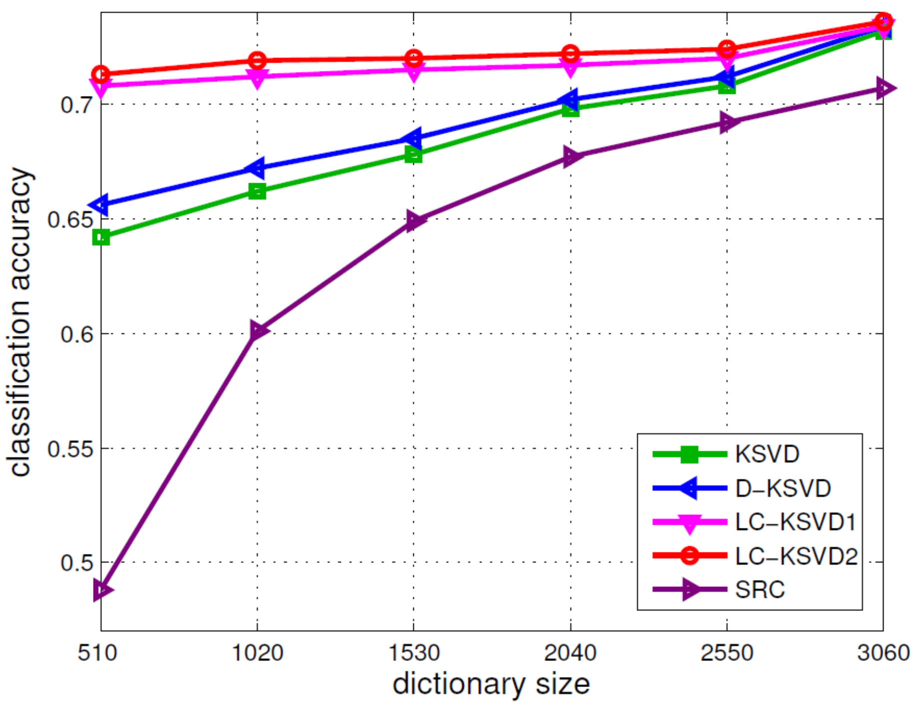
- Caltech-101 Object Dataset

number of train. samp.	5	10	15	20	25	30
Malik [32]	46.6	55.8	59.1	62.0	-	66.20
Lazebnik [15]	-	-	56.4	-	-	64.6
Griffin [11]	44.2	54.5	59.0	63.3	65.8	67.60
Irani [2]	-	-	65.0	-	-	70.40
Grauman [14]	-	-	61.0	-	-	69.10
Venkatesh [24]	-	-	42.0	-	-	-
Gemert [8]	-	-	-	-	-	64.16
Yang [29]	-	-	67.0	-	-	73.20
Wang [27]	51.15	59.77	65.43	67.74	70.16	73.44
SRC [28]	48.8	60.1	64.9	67.7	69.2	70.7
K-SVD [1]	49.8	59.8	65.2	68.7	71.0	73.2
D-KSVD [33]	49.6	59.5	65.1	68.6	71.1	73.0
LC-KSVD1	53.5	61.9	66.8	70.3	72.1	73.4
LC-KSVD2	**54.0**	**63.1**	**67.7**	**70.5**	**72.3**	**73.6**

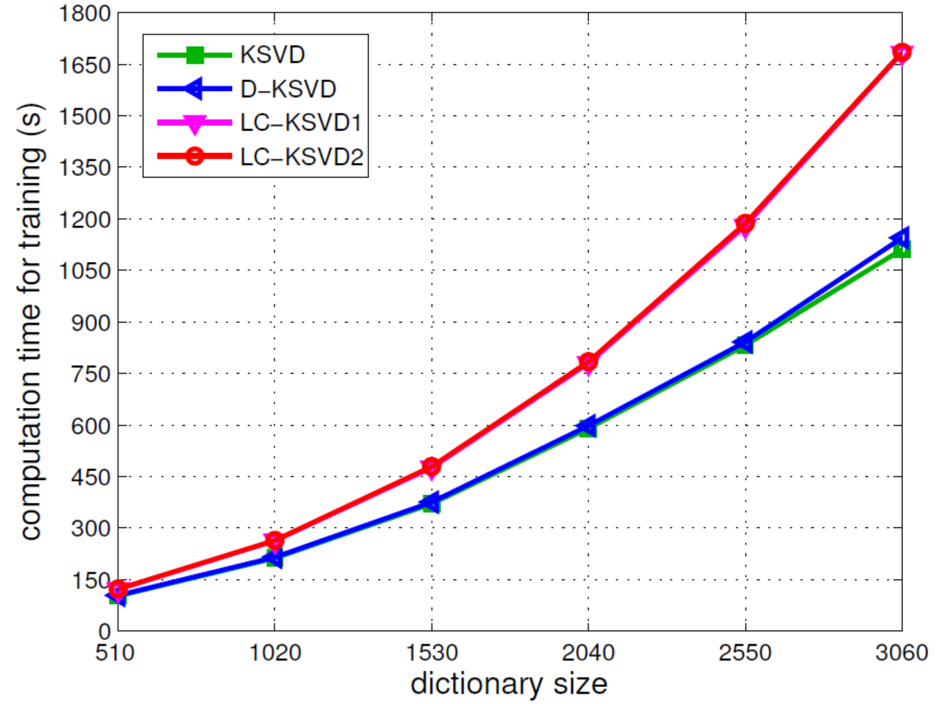
Dictionary size	510	1020	1530	2040	2550	3060
SRC [28]	173.44	343.12	520.88	662.40	835.34	987.55
LC-KSVD1	0.59	1.09	1.62	2.21	2.83	3.50
LC-KSVD2	0.54	0.98	1.44	1.94	2.50	3.17

Experimental Results

- Performance Comparisons



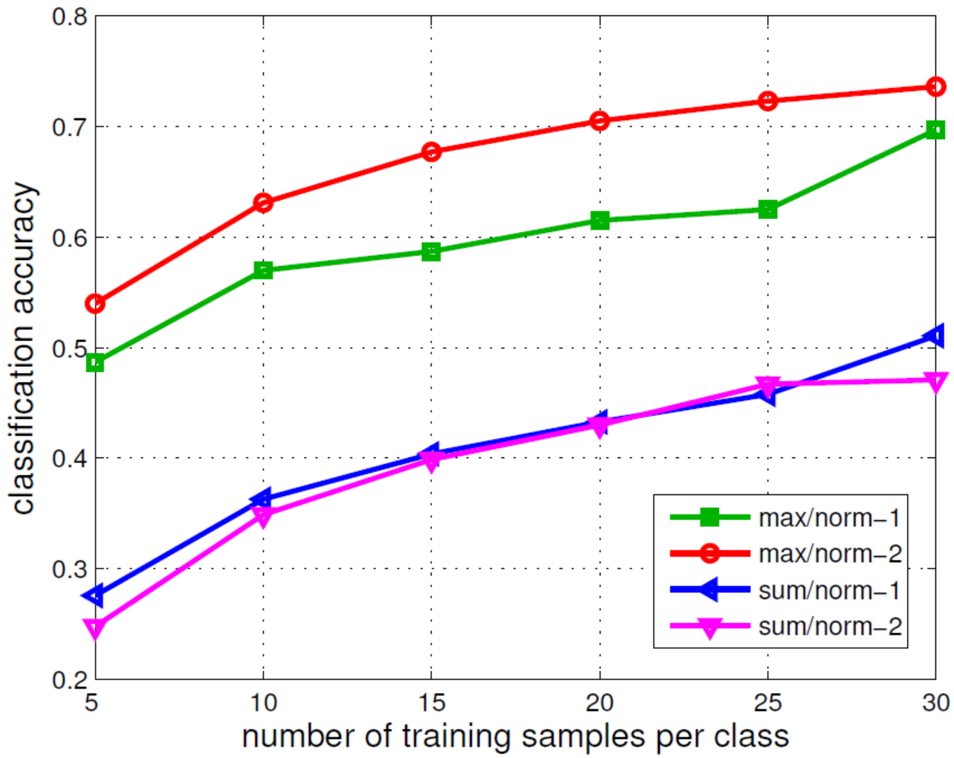
(a) Recognition accuracy
with varying different size



(b) Training time
with varying different size

Experimental Results

- Performance Comparisons



(c) Recognition accuracy with different spatial-pyramid-matching settings

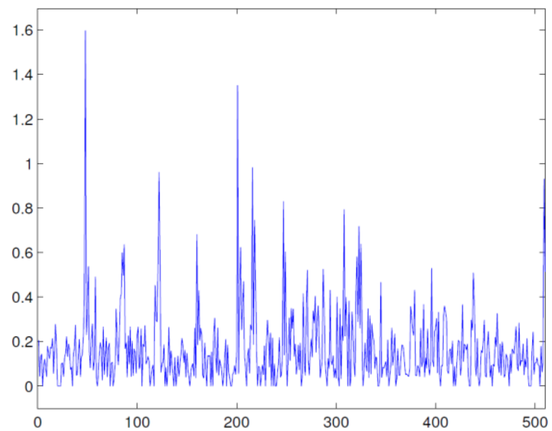
Sum: $x_{out} = x_1 +, ..., +x_n$

Max: $x_{out} = max(x_1, ..., x_n)$

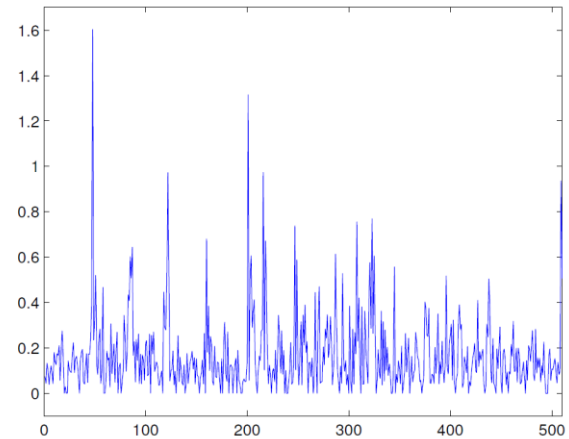
Norm-1: $x_{out} = x_{out} / \sum_i x_i$

Norm-2: $x_{out} = x_{out} / \|x_{out}\|_2$

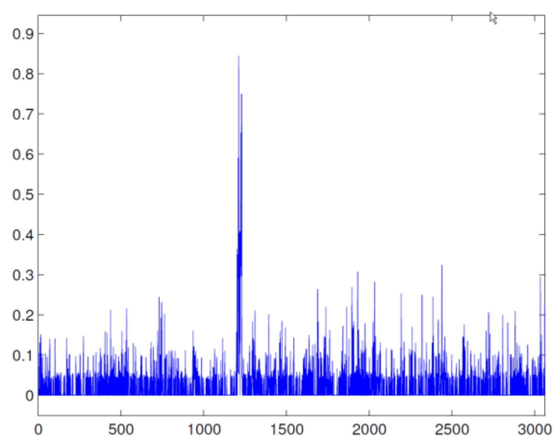
Examples of Sparse Representation



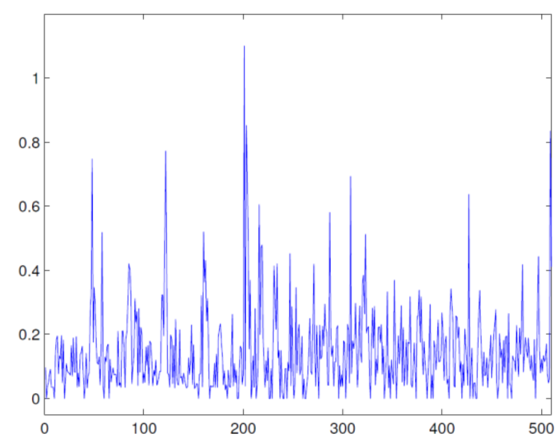
K-SVD [2]



D-KSVD [4]



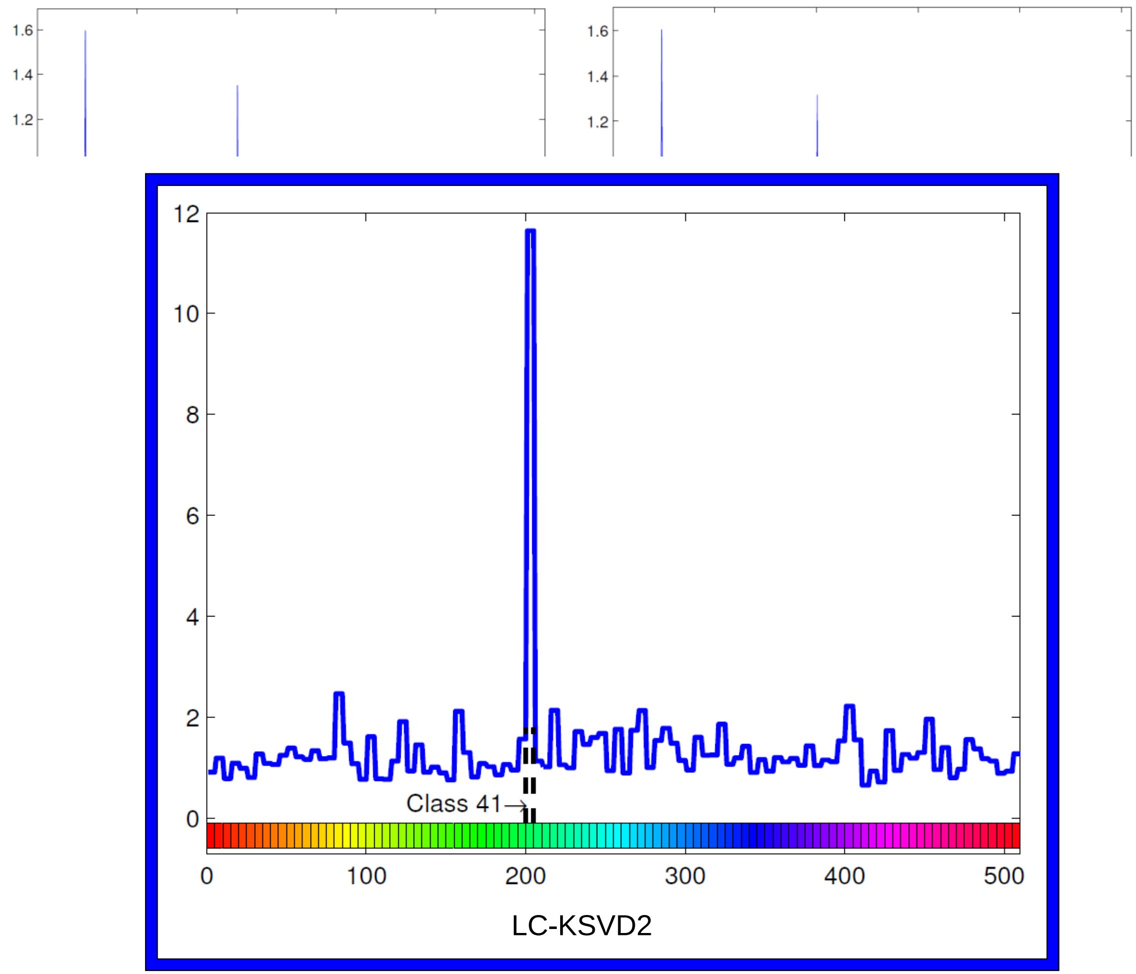
SRC [1]



SRC*[1]

- ❑ Class 41 in Caltech101 (55 test images).
- ❑ X axis means dictionary items; Y axis indicates a sum of absolute sparse codes.

Examples of Sparse Representation



- ❑ Class 41 in Caltech101 (55 test images).
- ❑ X axis means dictionary items; Y axis indicates a sum of absolute sparse codes.

Discussions and Conclusions

- Our approach is able to learn the dictionary, discriminative coding and classifier parameters simultaneously;
- Unlike approaches that learn multiple classifiers for categories to gain discrimination, our approach yields very good classification results with only one simple linear classifier;
- The basic reason for the good performance, is that the new constraint encourages the input signals from the same class to have similar sparse codes and those from different classes to have dissimilar sparse codes.
- Possible future work include: (a) How to learn the optimal dictionary size for the learned dictionary; (b) How to update the dictionary without retraining.

Important References

[1] J. Wright, A. Yang, A. Ganesh, S. Sastry and Y. Ma. **Robust face recognition via sparse representation**. TPAMI 2009.

[2] M. Aharon, M. Elad and A. Bruchstein. **K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation**. IEEE Trans. Sig. Proc., 2006.

[3] D. Pham and S. Venkatesh. **Joint learning and dictionary construction for pattern recognition**. CVPR 2008.

[4] Q. Zhang and B. Li. **Discriminative k-svd for dictionary learning in face recognition**. CVPR 2010.

[5] J. Wang, J. Yang and T. Huang. **Locality-constrained linear coding for image classification**. CVPR 2010.

[6] J. Mairal, F. Bach, J. Ponce, G. Sapiro and A. Zisserman. **Discriminative Learned Dictionaries for Local Image Analytics**. IEEE Conference on Computer Vision and Pattern Recognition. CVPR 2008.

[7] J. Mairal, F. Bach, J. Ponce and G. Sapiro. **Online learning for matrix factorization and sparse coding**. Journal of Machine Learning Research, 2010.

[8] R. Jenatton, J. Mairal, G. Obozinski, and F. Bach. **Proximal methods for sparse hierarchical dictionary learning**. ICML, 2010.

[9] J. Mairal, F. Bach, J. Ponce, G.Sapiro and A. Zissermann. **Supervised dictionary learning**. NIPS, 2009.

[10] J. Yang, K. Yu and T. Huang. **Supervised translation-invariant sparse coding**. CVPR, 2010.