Learning A Discriminative Dictionary for Sparse Coding via Label Consistent K-SVD

Zhuolin Jiang[1] *Zhe Lin*[2] *Larry S. Davis*[1]

[1]University of Maryland at College Park, MD, 20742

[2]Adobe Systems Incorporated, San Jose, CA, 95110

1. Overview

- **Goal**
  – To learn a dictionary with discriminative and representational power for sparse representation.
- **Approach**
  – 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 (for all categories) are learned simultaneously.

2. Related Work

- Sparse Coding has been successfully applied to a variety of problems in computer vision such as face recognition [1]. SRC algorithm [1] employs the entire set of training samples to form a dictionary.
- K-SVD [2]: Efficiently learn an over-complete dictionary with a small size. It focuses on representational power, but does not consider discriminative capability.
- Discriminative dictionary learning approaches:
  - ❑ Constructing a separate dictionary for each class.
  - ❑ Unifying the dictionary learning and classifier training into a mixed reconstructive and discriminative formulation [3,4].

3. Dictionary Learning

- Dictionary Learning for Reconstruction and Sparse Coding

Let Y be a set of n-dimensional N input signals, $Y = [y_1 ... y_N] \in R^{n \times N}$, Dictionary D is learned:

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

Given D, the sparse representation X of Y is:

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

- 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:

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

3. Label Consistent K-SVD

- **LC-KSVD1**

❑ Objective function

$$< D, A, X >= \arg\min_{D,A,X} \|Y - DX\|_2^2 + \alpha \|Q - AX\|_2^2 \; s.t. \forall i, \|x_i\|_0 \le T$$

❑ An example of Q

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

A : a linear transformation matrix

Q: discriminative sparse codes of input signals Y for classification

- **LC-KSVD2**

❑ Objective function:

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

Assume $X' = AX$, then $D' = DA^{-1}, W' = WT^{-1}$.The above objective function is rewritten as

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

- **Optimization**

We rewrite the objective function of LC-KSVD2 as:

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

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

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

- **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: $W = (XX^t + \lambda_1 I)^{-1} XH^t$

- **Classification**

❑ $\hat{D}, \hat{A}, \hat{W}$

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

$$\hat{D} = \{\frac{d_1}{\|d_1\|_2}, \frac{d_2}{\|d_2\|_2} \cdots \frac{d_K}{\|d_K\|_2}\}$$

$$\hat{A} = \{\frac{a_1}{\|d_1\|_2}, \frac{a_2}{\|d_2\|_2} \cdots \frac{a_K}{\|d_K\|_2}\}$$

$$\hat{W} = \{\frac{w_1}{\|d_1\|_2}, \frac{w_2}{\|d_2\|_2} \cdots \frac{w_K}{\|d_K\|_2}\}$$

❑ **Classification**

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

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

Then the classification result (i.e. the label j of y_i) is given by

$$j = \arg\max_j (l = \hat{W} x_i)$$

4. Experiments

- **Experimental Setup**

❑ **Random face-based feature**

- dims: 504 (Extended Yale), 540 (AR Face)

❑ **Spatial pyramid feature**

- 1024 bases
- dims: 3000 (Caltech101)

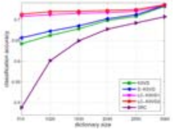
- **Caltech101**

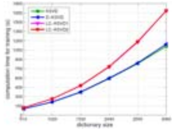
-102 classes

- The number of images per category: 31~800

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





- **Extended Yale**

- (Randomly selected) half of the images (training) + the other half (testing).

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

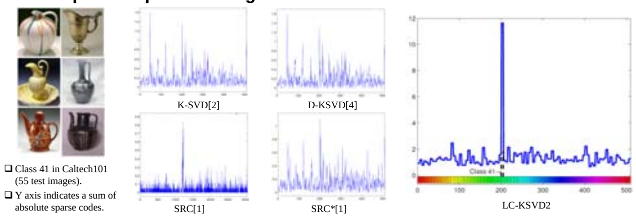
- **AR Face**

- (Randomly selected) 20 images (training) + 6 (testing)

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

- **Examples of sparse coding**



❑ Class 41 in Caltech101 (55 test images).

❑ Y axis indicates a sum of absolute sparse codes.

5. Key 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.