Submodular Dictionary Learning for Sparse Coding

Zhuolin Jiang, Guangxiao Zhang, Larry S. Davis

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

Goals

- Motivations

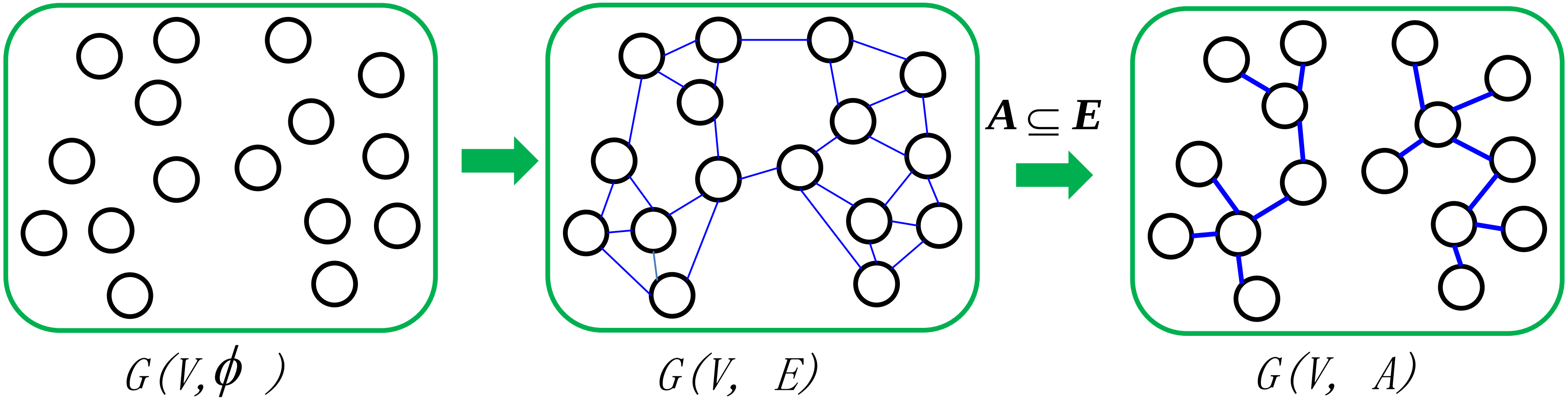
  Most of recent dictionary learning techniques are iterative batch procedures, it is relatively slow close to the minimum.

- Goals

  - Learn a discriminative and representational dictionary for sparse representation efficiently using a greedy algorithm for a submodular objective set function.

Approaches

- Approaches
  - A dataset is mapped into an undirected k-nearest neighbor graph G=(V, E). The dictionary learning is modeled as a graph topology selection problem. A subset of edges A is selected from initial edge set E such that the resulting graph G=(V, A), contains exactly K connected components or clusters.



Approaches

- Approaches
  - A monotonic and submodular objective function for dictionary learning consists of two terms: the entropy rate of a random walk on a graph and a discriminative term
  - The objective function is optimized by a highly efficient greedy algorithm
  - This simple greedy algorithm gives a near-optimal solution with a (1/2)-approximation bound [5].

Related Work

- Sparse Coding has been successfully applied to a variety of problems such as face recognition [1]. The 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 it does not consider discrimination.
- Discriminative dictionary learning approaches:
  - Constructing a separate dictionary for each class.
  - Adding discriminative terms into the objective function of dictionary learning [3].
- The diminishing return property of a submodular function has been employed in applications such as sensor placement, clustering and superpixel segmentation [4].

Preliminaries

- Submodular Set Function

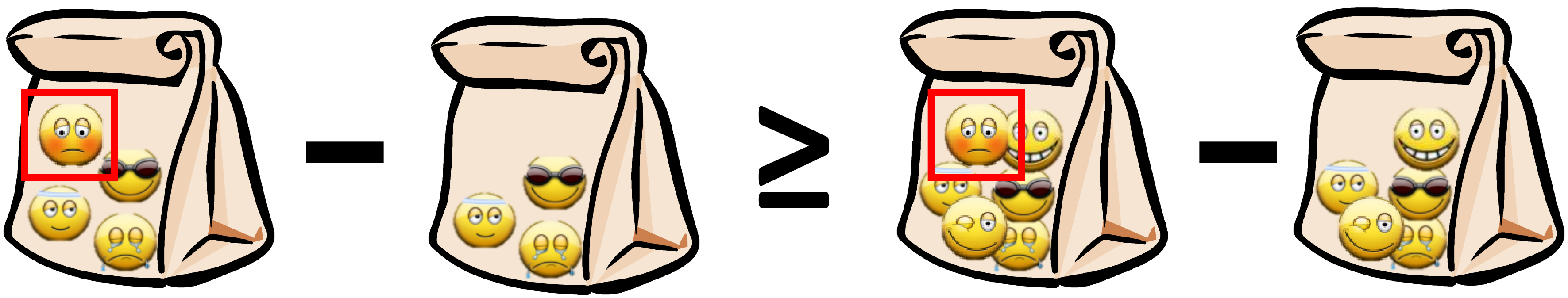
A set function $F : 2^E \rightarrow R$ is submodular if

$$F(A_1 \cup \{a\}) - F(A_1) \geq F(A_2 \cup \ a) - F(A_2)$$

for all $A_1 \subseteq A_2 \subseteq E$ and $a \in E \backslash A_2$

diminishing return property

$F(\ A_1\ U\ \{a\}\ )\ -\ F(\ A_1\ )\ \geq\ F(\ A_2\ U\ \{a\}\ )\ -\ F(\ A_2\ )$



Submodular Dictionary Learning

- Monotonic and Submodular Objective Set Function
  - It consists of an entropy rate term $\mathcal{H}(A)$ and a discriminative term $\mathcal{Q}(A)$:

$$\max_{A} \mathcal{F}(A) = \mathcal{H}(A) + \lambda \mathcal{Q}(A) \ s.t. \ A \subseteq E \ and \ N_A \geq K,$$

where

A: selected subset of edge set *E*;

N_A: number of connected components induced by *A*

Submodular Dictionary Learning

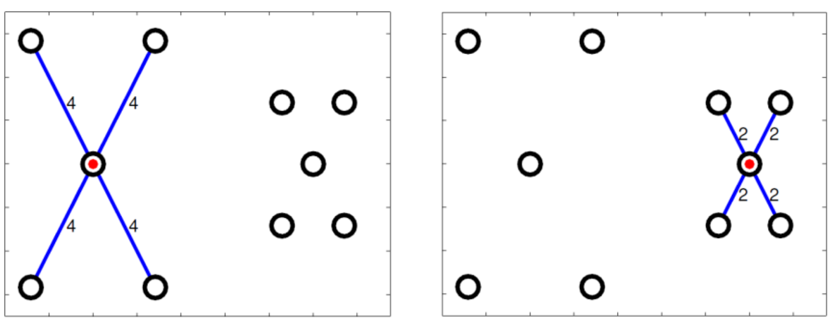
- Entropy Rate of a Random Walk

$$\mathcal{H}(A) = -\sum_i \mu_i \sum_j P_{i,j}(A) \log P_{i,j}(A)$$

μ_i: Stationary probability of vertex v_i

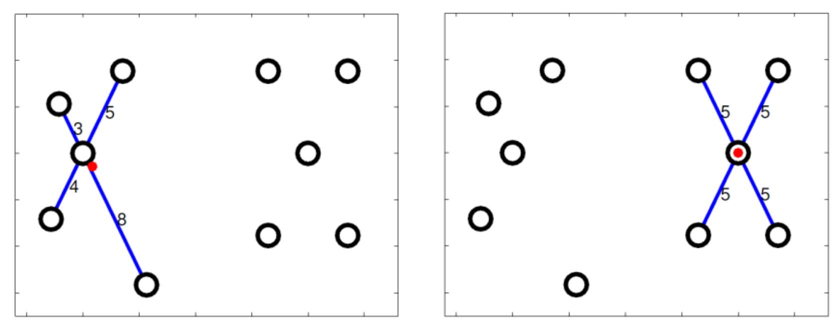
$P_{i,j}$: Transition probability from v_i to v_j

Compactness



(a) Entropy Rate = 0.03 (b) Entropy Rate = 0.43

Homogeneity



(c) Entropy Rate = 0.22 (d) Entropy Rate = 0.24

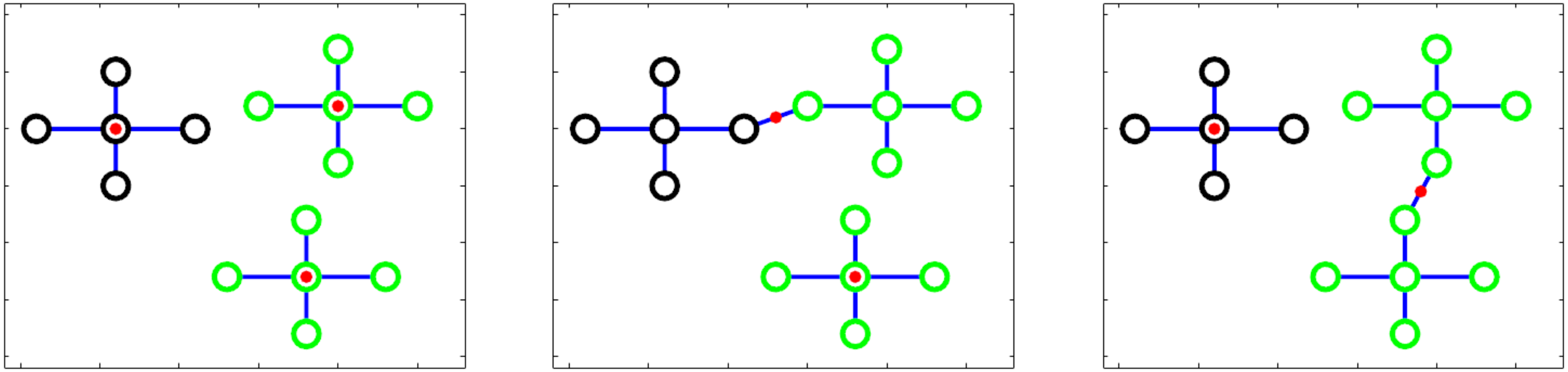
Submodular Dictionary Learning

- Discriminative Term

$$\mathcal{Q}(A) = \frac{1}{C}\sum_{i=1}^{N_A} \max_y N_y^i - N_A$$

N_y^i: Number of elements from class *y* in cluster *i*

Class Pure & A Smaller Number of Clusters



(a) Disc. Fun. = −2.00 (b) Disc. Fun. = −1.33 (c) Disc. Fun. = −1.00

Submodular Dictionary Learning

- Optimization
  - A simple greedy gives a (1/2)-approximation to the optimal solution.

Algorithm 1 Submodular Dictionary Learning (SDL)

Input: $G = (V, E)$, w, K, λ and $\mathcal{N}$
Output: D
Initialization: $A \leftarrow \emptyset$, $D \leftarrow \emptyset$
for $N_A > K$ **do**
  $\tilde{e} = \underset{A \cup \{e\} \in \mathcal{I}}{\operatorname{argmax}} \mathcal{F}(A \cup \{e\}) - \mathcal{F}(A)$
  $A \leftarrow A \cup \{\tilde{e}\}$
end for
for each subgraph S_i in $G = (V, A)$ **do**
  $D \leftarrow D \cup \{\frac{1}{|S_i|} \sum_{j: v_j \in S_i} v_j\}$
end for

Classification

- Object and Face
  - For a test image y_i, first compute its sparse representation:

$$z_i = \arg\min_{z_i} \|y_i - Dz_i\|_2^2 \; s.t. \; \|z_i\|_0 \leq s$$

  - Then the label of y_i is the index i corresponding to the largest element of a class label vector $l = W z_i$.

Multivariate ridge regression

- Human Actions
  - Dynamic time warping is employed to align two sequences in the sparse representation domain; next a K-NN classifier is used

Experimental Results

- Evaluation Datasets
  - Extended YaleB Database (Face database)
  - Keck Gesture Dataset (Gesture)
  - Caltech101 Dataset (Object)
- Experimental Setup
  - Random face-based features

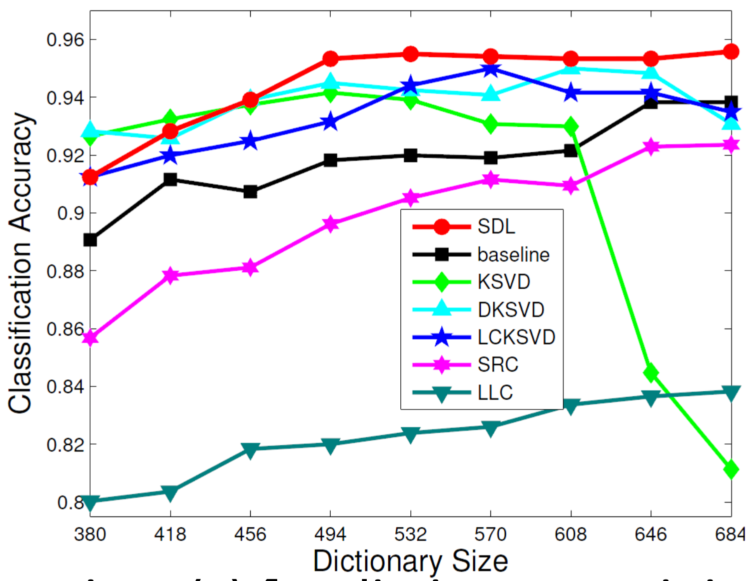
    - dims: 504 (Extended YaleB)
  - Joint Shape and Motion features

    - dims: 512 (Keck Gesture)
  - Spatial pyramid features

    - dims: 3000 (Caltech101)

Experiment Results

- Extended YaleB
  - Classification accuracy comparison

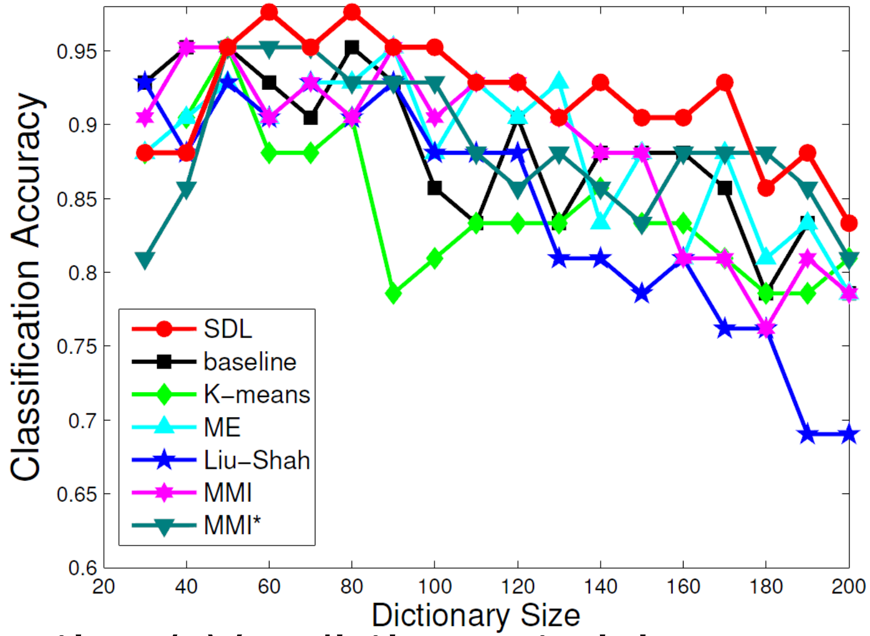


  - Computation time (s) for dictionary training

Dict. size	418	456	494	532	570	608	646	684
SDL	0.9	1.0	0.9	0.9	0.9	1.0	0.9	0.9
K-SVD [1]	52.6	56.1	59.8	64.9	67.9	72.2	76.2	78.0
D-KSVD [35]	53.1	56.9	60.5	65.8	68.1	74.9	77.6	79.2
LC-KSVD [12]	67.2	72.6	78.3	86.5	90.7	97.8	104.4	112.3

Experiment Results

- Keck Gesture Dataset
  - Classification accuracy comparison



  - Computation time (s) for dictionary training

Dict. size	40	60	80	100	120	140	160	180
SDL	1.0	1.0	1.1	1.0	1.0	1.1	1.0	1.0
K-means	1.2	1.1	1.6	1.4	1.8	2.1	2.1	2.2
ME [10]	48.5	57.2	70.2	84.6	91.5	113.1	118.9	130
LiuShah [18]	599.2	597.9	597.2	596.1	593.9	590.3	587.4	582
MMI [26]	64.6	92.6	115.5	140.3	150.1	164.1	184.4	201

Experimental Results

- Caltech101
  - Classification accuracy comparison

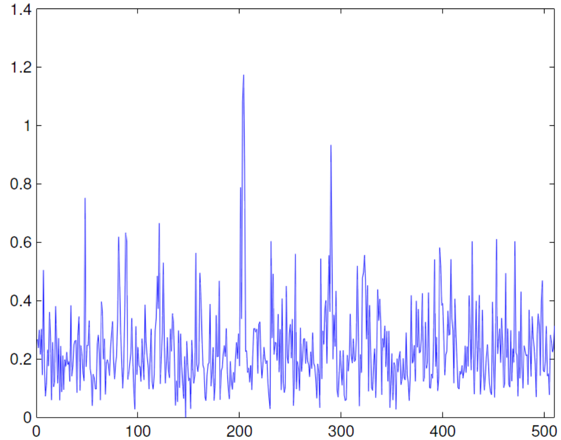
Training Images	5	10	15	20	25	30
Malik [34]	46.6	55.8	59.1	62.0	-	66.20
Lazebnik [15]	-	-	56.4	-	-	64.6
Griffin [9]	44.2	54.5	59.0	63.3	65.8	67.60
Irani [2]	-	-	65.0	-	-	70.40
Grauman [11]	-	-	61.0	-	-	69.10
Venkatesh [25]	-	-	42.0	-	-	-
Gemert [7]	-	-	-	-	-	64.16
Yang [31]	-	-	67.0	-	-	73.20
Wang [29]	51.15	59.77	65.43	67.74	70.16	73.44
SRC [30]	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 [35]	49.6	59.5	65.1	68.6	71.1	73.0
LC-KSVD [12]	54.0	63.1	**67.7**	70.5	72.3	73.6
SDL	**55.3** ± 0.5	**63.4** $\pm$ 0.5	67.5 $\pm$ 0.3	**70.7** $\pm$ 0.3	**73.1** $\pm$ 0.4	**75.3** $\pm$ 0.4

  - Computation time (s) for dictionary training

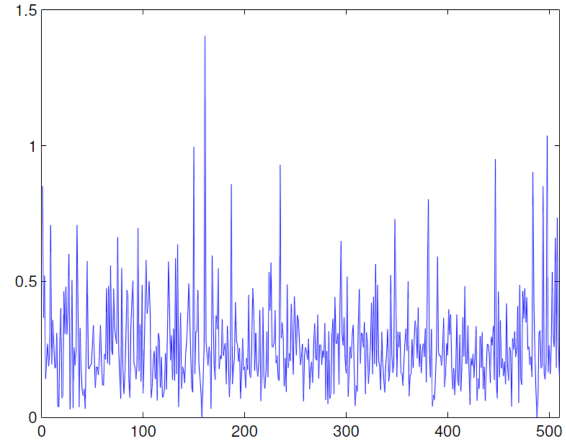
Dict. size	306	510	714	918	1122	1326	1530
SDL	37.5	36.7	36.6	36.9	37.1	36.7	36.7
K-SVD [1]	578.3	790.1	1055	1337	1665	2110	2467
D-KSVD [35]	560.1	801.3	1061	1355	1696	2081	2551
LC-KSVD [12]	612.1	880.6	1182	1543	1971	2496	3112

Experiment Results

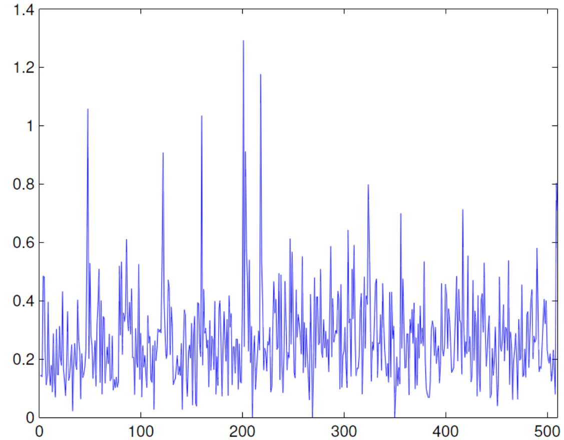
- Examples of sparse codes



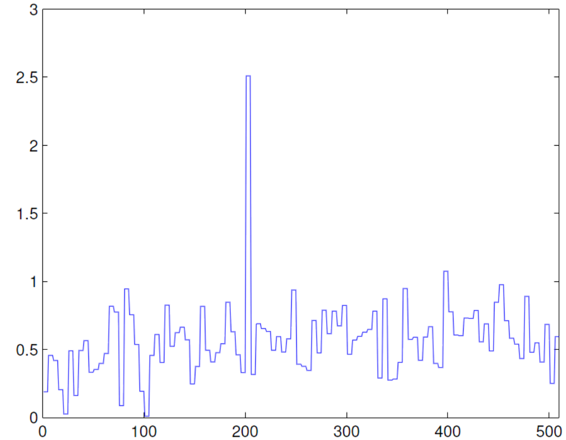
SRC



K-SVD



D-KSVD

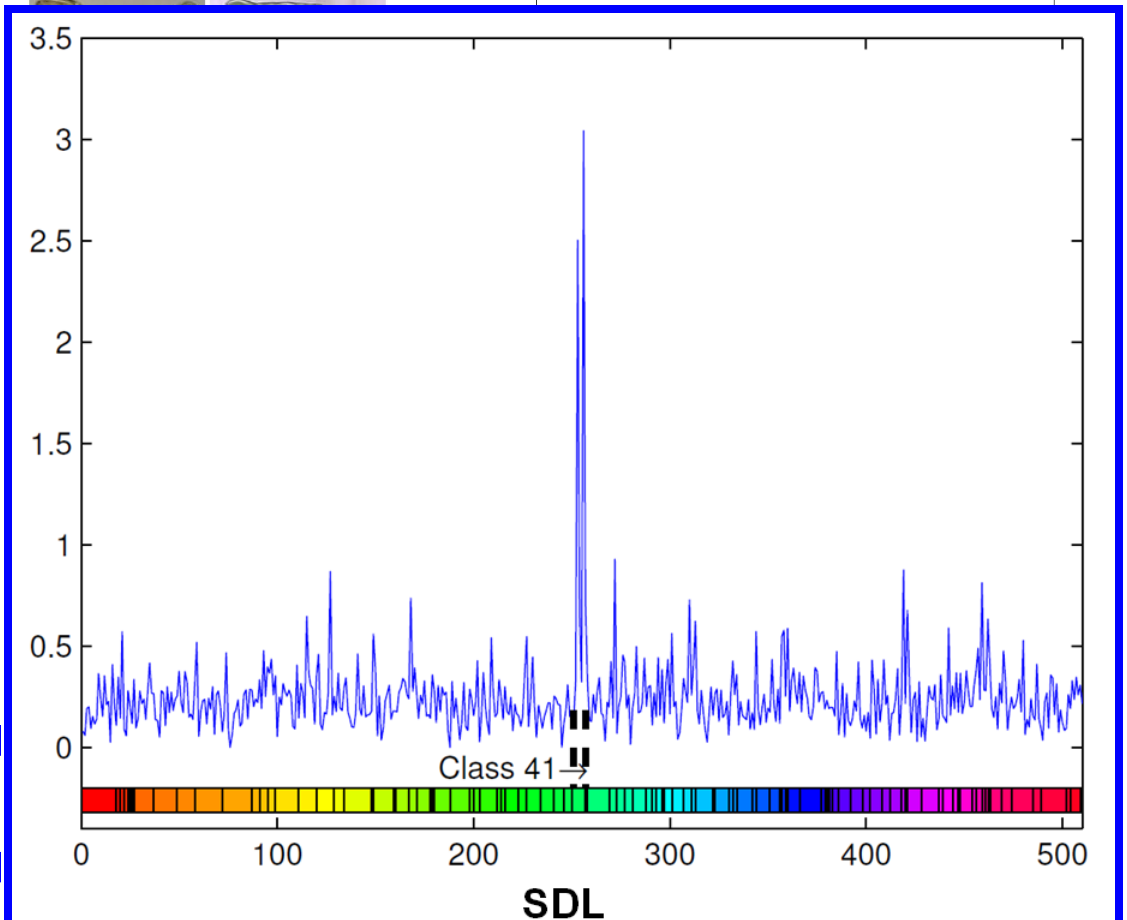


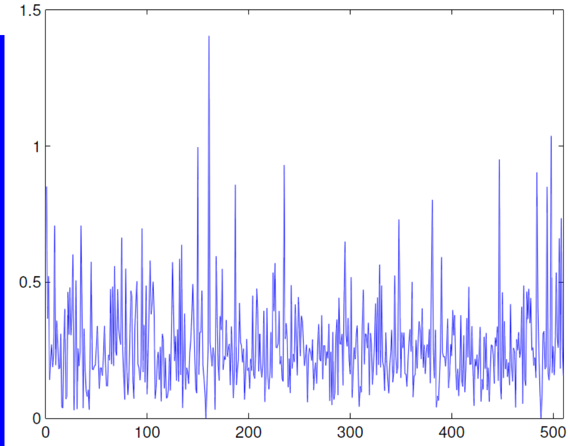
LC-KSVD

- Class 41 in Caltech101 (55 test images).
- Y axis indicates a sum of absolute sparse codes.

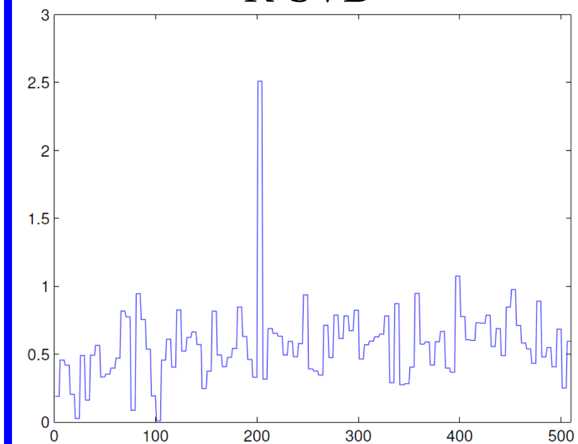
Experiment Results

- Examples of sparse codes





K-SVD



LC-KSVD

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 over-complete dictionaries for sparse representation. Sig. Proc., 2006.
3. Q. Zhang and B. Li. Discriminative k-svd for dictionary learning in face recognition, CVPR 2010.
4. M. Liu, O. Tuzel, S. Ramalingam, and R. Chellappa. Entropy rate superpixel segmentation, CVPR 2011.
5. G. Nemhauser, L. Wolsey, and M. Fisher. An analysis of the approximations for maximizing submodular set functions. Mathematical Programming, 1978