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

- 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 $F : 2^E \rightarrow \mathbb{R}$ 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 $a \in E \setminus A_2$

diminishing return property

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

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

\[
\max_A \mathcal{F}(A) = \mathcal{H}(A) + \lambda Q(A) \text{ s.t. } A \subseteq E \text{ and } N_A \geq K,
\]

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

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

- \( \mu_i \): 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**

\[
Q(A) = \frac{1}{C} \sum_{i=1}^{N_A} \max_y N^i_y - N_A
\]

\(N^i_y\): 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: $G = (V, E)$, $w$, $K$, $\lambda$ and $\mathcal{N}$
Output: $D$
Initialization: $A \leftarrow \emptyset$, $D \leftarrow \emptyset$
for $N_A > K$ do
  $\tilde{e} = \arg\max_{A \cup \{e\} \in \mathcal{I}} \mathcal{F}(A \cup \{e\}) - \mathcal{F}(A)$
  $A \leftarrow A \cup \{\tilde{e}\}$
end for
for each subgraph $S_i$ in $G = (V, A)$ do
  $D \leftarrow D \cup \left\{ \frac{1}{|S_i|} \sum_{j : v_j \in S_i} v_j \right\}$
end for
Classification

- **Object and Face**
  - For a test image $y_i$, first compute its sparse representation:
    $$ z_i = \arg \min_{z_i} \| y_i - Dz_i \|_2^2 \text{ s.t. } \| z_i \|_0 \leq s $$
  - Then the label of $y_i$ is the index $i$ corresponding to the largest element of a class label vector $l = Wz_i$.

- **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
  
  ![Classification Accuracy Comparison Graph](image)

  - Computation time (s) for dictionary training

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

- Keck Gesture Dataset
  - Classification accuracy comparison

![Classification Accuracy Comparison](image)

- Computation time (s) for dictionary training

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

- **Caltech101**
  - Classification accuracy comparison

<table>
<thead>
<tr>
<th>Training Images</th>
<th>5</th>
<th>10</th>
<th>15</th>
<th>20</th>
<th>25</th>
<th>30</th>
</tr>
</thead>
<tbody>
<tr>
<td>Malik [34]</td>
<td>46.6</td>
<td>55.8</td>
<td>59.1</td>
<td>62.0</td>
<td>-</td>
<td>66.20</td>
</tr>
<tr>
<td>Lazebnik [15]</td>
<td>-</td>
<td>-</td>
<td>56.4</td>
<td>-</td>
<td>-</td>
<td>64.6</td>
</tr>
<tr>
<td>Griffin [9]</td>
<td>44.2</td>
<td>54.5</td>
<td>59.0</td>
<td>63.3</td>
<td>65.8</td>
<td>67.60</td>
</tr>
<tr>
<td>Irani [2]</td>
<td>-</td>
<td>-</td>
<td>65.0</td>
<td>-</td>
<td>-</td>
<td>70.40</td>
</tr>
<tr>
<td>Venkatesh [25]</td>
<td>-</td>
<td>-</td>
<td>42.0</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Gemert [7]</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>64.16</td>
</tr>
<tr>
<td>Yang [31]</td>
<td>-</td>
<td>-</td>
<td>67.0</td>
<td>-</td>
<td>-</td>
<td>73.20</td>
</tr>
<tr>
<td>Wang [29]</td>
<td>51.15</td>
<td>59.77</td>
<td>65.43</td>
<td>67.74</td>
<td>70.16</td>
<td>73.44</td>
</tr>
<tr>
<td>SRC [30]</td>
<td>48.8</td>
<td>60.1</td>
<td>64.9</td>
<td>67.7</td>
<td>69.2</td>
<td>70.7</td>
</tr>
<tr>
<td>K-SVD [1]</td>
<td>49.8</td>
<td>59.8</td>
<td>65.2</td>
<td>68.7</td>
<td>71.0</td>
<td>73.2</td>
</tr>
<tr>
<td>D-KSVD [35]</td>
<td>49.6</td>
<td>59.5</td>
<td>65.1</td>
<td>68.6</td>
<td>71.1</td>
<td>73.0</td>
</tr>
<tr>
<td>LC-KSVD [12]</td>
<td>54.0</td>
<td>63.1</td>
<td>67.7</td>
<td>70.5</td>
<td>72.3</td>
<td>73.6</td>
</tr>
<tr>
<td><strong>SDL</strong></td>
<td><strong>55.3</strong></td>
<td><strong>63.4</strong></td>
<td><strong>67.5</strong></td>
<td><strong>70.7</strong></td>
<td><strong>73.1</strong></td>
<td><strong>75.3</strong></td>
</tr>
</tbody>
</table>

  ± 0.5  ± 0.5  ± 0.3  ± 0.3  ± 0.4  ± 0.4

- Computation time (s) for dictionary training

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

- Examples of sparse codes

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

- Examples of sparse codes

![Graphs showing sparse codes for different methods with examples from Class 41 in Caltech101 dataset.](image)

- K-SVD
- LC-KSVD
Key References


3. Q. Zhang and B. Li. Discriminative k-svd for dictionary learning in face recognition, CVPR 2010.
