Sparse Learning For Image Classification

Dr. Meng Yang
Smart LV Lab, Computer Vision Institute, Shenzhen University
OUTLINE

- Introduction of sparse learning
- Robust sparse learning
- Discriminative dictionary learning
- Outlook
OUTLINE

- Introduction of sparse learning
- Robust sparse learning
- Discriminative dictionary learning
- Outlook
DATA REPRESENTATION

- Massive High-Dimensional Data

- Low-dimensional structures
Sparse Transformation

Most energy concentrated in a small number of features

\[ y \in \mathbb{R}^m \quad = \quad A \in \mathbb{R}^{m \times n} \quad x \in \mathbb{R}^n \]

\[ \begin{bmatrix} ? & ? & \vdots & ? \\ ? & ? & \vdots & ? \end{bmatrix} \]

\[ x \text{ is a sparse vector.} \]
SPARSE SIGNAL PROCESSING

Signal Processing → Compressive sensing → Sparse Learning

\[ M y = N \Phi \]

Measurement matrix

\[ N \Psi \]

\[ f \]
Visual Cortex → Neural codes → Sparse Learning

- Lifetime sparseness
- Population sparseness

FACE RECOGNITION VIA SPARSE LEARNING

Test image = Training set

Coding coefficient $x$ and residual $e$ are sparse!

$y = Ax - x + e$

Seek the \textit{sparsest} solution:

\[
\min \|x\|_0 + \|e\|_0 \quad \text{subj} \quad y = Ax + e
\]

\[\Rightarrow\]

\[
\min \|x\|_1 + \|e\|_1 \quad \text{subj} \quad y = Ax + e
\]

\[\delta_i(\hat{x}_1) \in \mathbb{R}^N\]

\[r_i = \|y - A\delta_i(\hat{x}_1) - \hat{e}_1\|_2\]

\textbf{Classification criterion:} \texttt{Identity} = \texttt{argmin}_i \{r_i\}. 
OUTLINE

- Introduction of sparse learning
- Robust sparse learning
- Discriminative dictionary learning
- Outlook
Robust Sparse Learning

Sparse coding

Robust coding

Structured coding

\[ \hat{\alpha} = \arg \min_{\alpha} \left\{ \sum_{i=1}^{n} \rho_{\theta} (y_i - r_i \alpha) + \sum_{j=1}^{m} \rho_{\omega} (\alpha_j) \right\} \]

\[ \min_{\alpha} \| [\alpha; \beta] \|_1 \quad \text{s.t.} \quad y = [D; I] \cdot [\alpha; \beta] \]

\[ \min_{\alpha} \| A(\alpha) - Y \|_2^2 + \lambda \| \alpha \|_2^2 \]
Gaussian/Laplacian model can't well fit practical residuals.
\[ \min_\alpha \| y - D\alpha \|_2^2 + \lambda \| \alpha \|_1 \]

works best for Gaussian-distributed coding residuals.

\[ \min_\alpha \| y - D\alpha \|_1 + \lambda \| \alpha \|_1 \]

works best for Laplacian-distributed coding residuals.

\[ \min_\alpha \| y - D\alpha \|_\tau + \lambda \| \alpha \|_1 \]

works best for practical coding residuals.
coding $y$ over $A=[r_1; r_2; \ldots; r_m]$ by maximizing the posterior probability

$$\hat{\alpha} = \arg\max_\alpha \ln P(\alpha | y)$$

$e = y - D\alpha$ or $\alpha$ independent and identically distributed

Regularized Robust Coding

$$\hat{\alpha} = \arg\min_\alpha \left\{ \sum_{i=1}^{n} \rho_\theta (y_i - r_i\alpha) + \sum_{j=1}^{m} \rho_o (\alpha_j) \right\}$$

$\text{RRC}_L1 (\rho_o(x)=|x|), \text{RRC}_L2(\rho_o(x)=x^2)$
ITERATIVE REWEIGHTED ROBUST CODING

Regularized Robust Coding

Taylor expansion of data fidelity

Weighted Robust Coding

\[ \hat{\alpha} = \arg \min_{\alpha} \left\{ \frac{1}{2} \| W^{1/2} (y - A\alpha) \|_2^2 + \sum_{j=1}^{m} \rho_\circ (\alpha_j) \right\} \]
Higher weight value for inliers and lower weight value for outliers.

\[ W_{i,i} = \omega_{\theta}(e_i) = \frac{1}{1 + \exp(\mu e_i^2 - \mu \delta)} \]
ITERATIVE REWEIGHTED ALGORITHM

WHILE not converged DO
  Compute Residual
  Estimate weight
  Weighted Robust Coding
  Reconstruct testing sample
END WHILE
### Face Recognition with Occlusion

**Testing image**

<table>
<thead>
<tr>
<th>Occlusion (EYB)</th>
<th>SRC</th>
<th>GRRC_L1</th>
<th>CESR</th>
<th>RRC_L2</th>
<th>RRC_L1</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>100%</td>
<td>100%</td>
<td>94.7%</td>
<td>100%</td>
<td>100%</td>
</tr>
<tr>
<td>10</td>
<td>100%</td>
<td>100%</td>
<td>94.7%</td>
<td>100%</td>
<td>100%</td>
</tr>
<tr>
<td>20</td>
<td>99.8%</td>
<td>100%</td>
<td>92.7%</td>
<td>99.8%</td>
<td>99.8%</td>
</tr>
<tr>
<td>30</td>
<td>98.5%</td>
<td>99.8%</td>
<td>89.8%</td>
<td>97.6%</td>
<td>96.7%</td>
</tr>
<tr>
<td>40</td>
<td>90.3%</td>
<td>96.5%</td>
<td>83.9%</td>
<td>87.6%</td>
<td>87.4%</td>
</tr>
<tr>
<td>50</td>
<td>65.3%</td>
<td>87.4%</td>
<td>75.5%</td>
<td>87.8%</td>
<td>87.4%</td>
</tr>
</tbody>
</table>
## Face Recognition with Corruption

<table>
<thead>
<tr>
<th>Corruption (EYB)</th>
<th>0~50</th>
<th>60</th>
<th>70</th>
<th>80</th>
<th>90</th>
</tr>
</thead>
<tbody>
<tr>
<td>SRC</td>
<td>100%</td>
<td>99.3%</td>
<td>90.7%</td>
<td>37.5%</td>
<td>7.1%</td>
</tr>
<tr>
<td>CESR</td>
<td>97.4%</td>
<td>96.2%</td>
<td>97.8%</td>
<td>93.8%</td>
<td>41.5%</td>
</tr>
<tr>
<td><strong>RRC_L₂</strong></td>
<td>100%</td>
<td>100%</td>
<td>99.8%</td>
<td>97.8%</td>
<td>43.3%</td>
</tr>
<tr>
<td><strong>RRC_L₁</strong></td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
<td>99.6%</td>
<td>67.1%</td>
</tr>
</tbody>
</table>
OUTLINE

- Introduction of sparse learning
- Robust sparse learning
- Discriminative dictionary learning
- Outlook
The choice of the dictionary that sparsifies the signals is crucial for the success of this model.”

CLASS-SPECIFIC DL

- Metaface[18], DLSI[19], CS-DL[21], FDDL[20]
Predefined bases (e.g., wavelet, DCT) too general

Training data matrix may have a big size (e.g., SRC)

Discriminative Dictionary Learning

*Discriminative sparse coding coefficients*

*Discriminative class-specific sub-dictionary*
MAIN IDEA

**Sparse Coefficient**

**Discrimination of representation residual and coding coefficient.**

\[ D = [D_1, D_2, \ldots, D_c] \]

**GOOD FOR**

\[ D_i \]

**BAD FOR**

\[ X_i \]

**Fisher criterion**

\[ X_j \]

**SMALL** within-class scatter

**BIG** between-class scatter

**Discrimination of representation residual and coding coefficient.**
FDDL MODEL

\[
\min_{D,X} \left\{ \sum_{i=1}^{K} \left( r(A_i, D, X_i) + \lambda_1 \|X\|_1 + \lambda_2 f(X) \right) \right\}
\]

Discriminative data fidelity term

\[
r(A_i, D, X_i) = \|A_i - DX_i\|_F^2 + \|A_i - D_i X_i\|_F^2 + \sum_{j=1}^{K} \|D_j X_i\|_F^2
\]

Discriminative coefficient term

\[
f(X) = tr(S_{W}(X) - S_{R}(X)) + \eta \|X\|_F^2
\]

\[
A = [A_1, A_2, \ldots, A_K], \quad A_i : \text{training samples from class } i.
\]

\[
D = [D_1, D_2, \ldots, D_K], \quad D_i : \text{sub-dictionary of the } i\text{th class.}
\]

\[
X = [X_1, X_2, \ldots, X_K], \quad X_i : \text{coding coefficients of } A_i \text{ over } D.
\]
### DIGIT RECOGNITION

#### Learned dictionary atoms

<table>
<thead>
<tr>
<th>Algorithms</th>
<th>FDDL</th>
<th>SRSC</th>
<th>REC-L</th>
<th>REC-BL</th>
<th>SDL-G</th>
</tr>
</thead>
<tbody>
<tr>
<td>Error rate (%)</td>
<td>2.89</td>
<td>6.05</td>
<td>6.83</td>
<td>4.38</td>
<td>6.67</td>
</tr>
<tr>
<td>SDL-D</td>
<td>3.54</td>
<td>3.98</td>
<td>5.2</td>
<td>4.2</td>
<td>3.61</td>
</tr>
<tr>
<td>DLSI</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>KNN</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SVM</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>COPAR</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
HYBRID DL

- COPAR[28], JDL[27], SVDL[22],
SVDL

- **SPARSE REPRESENTATION**
  promising performance & *multiple* training images for each class

- **GENERIC TRAINING SET**
  Various variation of *generic* facial images

- **ADAPTIVE VARIATION DICTOINARY**
  Sparse variation dictionary *adaptive* for the gallery set
INTRODUCTION TO SVDL

\[ \hat{\alpha} = \arg \min_{\alpha} \| y - [G, D] \alpha \|^2_2 + \lambda \| \alpha \|_1 \quad \text{id} = \arg \min_{i} \left\{ \| y - g_i \hat{\alpha}_i - D \hat{\alpha}_D \|^2_2 \right\} \]
SPARSE VARIATION DICTIONARY LEARNING

For each $g_i$

**Generic Training Set**

{Reference subset $R_i$, Variation subset $X_i$}

**SVDL Model with $c$ classes**

$$\min \sum_{i=1}^{c} \left\{ p(g_i, R_i, \gamma_i) + q(D, X_i, \gamma_i) \right\}$$

**Adaptive projection learning**

Projection coefficient $\gamma_i$ is learned based on $g_i$ and $R_i$.

The projected variation set is generated via $Y_i = X_i \times \gamma_i$

**Sparse variation dictionary learning**

$$q(D, X_i, \gamma_i) = \|Y_i - DB_i\|_F^2 + \lambda_2 \|B_i\|_1 + \lambda_3 \|d_j\|_1$$
**SINGLE-SAMPLE FACE RECOGNITION**

![Face Recognition Examples](Image)

<table>
<thead>
<tr>
<th>Methods</th>
<th>CMU Multiple PIE with 100 subjects</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Pose1-S2</td>
<td>Pose2-S3</td>
<td>Corrution 20%</td>
<td>40%</td>
<td>20% Block</td>
</tr>
<tr>
<td>NN</td>
<td>26.0</td>
<td>8.7</td>
<td>44.4</td>
<td>31.1</td>
<td>35.0</td>
</tr>
<tr>
<td>SVM</td>
<td>26.0</td>
<td>8.7</td>
<td>44.4</td>
<td>31.1</td>
<td>35.0</td>
</tr>
<tr>
<td>SRC</td>
<td>25.0</td>
<td>7.3</td>
<td>44.8</td>
<td>29.8</td>
<td>38.4</td>
</tr>
<tr>
<td>DMMA</td>
<td>27.1</td>
<td>5.3</td>
<td>49.3</td>
<td>15.6</td>
<td>51.4</td>
</tr>
<tr>
<td>AGL</td>
<td>66.7</td>
<td>24.9</td>
<td>5.1</td>
<td>2.4</td>
<td>53.6</td>
</tr>
<tr>
<td>ESRC</td>
<td>63.9</td>
<td>31.8</td>
<td>68.8</td>
<td>25.3</td>
<td>68.4</td>
</tr>
<tr>
<td>ESRC-KSVD</td>
<td>67.1</td>
<td>29.9</td>
<td>72.6</td>
<td>29.1</td>
<td>68.9</td>
</tr>
<tr>
<td>SVDL</td>
<td><strong>77.8</strong></td>
<td><strong>38.3</strong></td>
<td><strong>100</strong></td>
<td><strong>97.2</strong></td>
<td><strong>87.7</strong></td>
</tr>
</tbody>
</table>
Latent Dictionary Learning (LDL)

Latent vector for $d_m$

$w_{j,m}$ indicates the relationship between atom $d_m$ and $j^{th}$ class label.
Latent sparse representation

\[
\min_{D,X,W} \sum_{j=1}^{C} \left\| A_j - D \text{diag}(w_j) X_j \right\|_F^2 + \lambda_1 \left\| X_j \right\|_1 + \lambda_2 \left\| X_j - M_j \right\|_F^2
\]

\[
+ \lambda_3 \sum_{j=1}^{C} \sum_{l \neq j} \sum_{n=1}^{N} \sum_{m \neq n} w_{j,m} \left( d_m^T d_n \right)^2 w_{l,n}
\]

s.t. \( w_{j,m} \geq 0 \ \forall \ j, m; \)

\[
\sum_{m} w_{j,m} = \delta, \ \forall \ m;
\]

Latent dictionary incoherence

Discriminative coefficient
## Experimental Results

<table>
<thead>
<tr>
<th>Methods</th>
<th>Accuracy (%)</th>
<th>Methods</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Qiu 2011</td>
<td>83.6</td>
<td>LCKSVD</td>
<td>91.2</td>
</tr>
<tr>
<td>Sadanand 2012</td>
<td>90.7</td>
<td>COPAR</td>
<td>90.7</td>
</tr>
<tr>
<td>SRC</td>
<td>92.9</td>
<td>JDL</td>
<td>90.0</td>
</tr>
<tr>
<td>DLSI</td>
<td>92.1</td>
<td>FDDL</td>
<td>93.6</td>
</tr>
<tr>
<td>DKSVD</td>
<td>88.1</td>
<td>LDL</td>
<td><strong>95.0</strong></td>
</tr>
</tbody>
</table>
OUTLINE

- Introduction of sparse learning
- Robust sparse learning
- Discriminative dictionary learning
- Outlook
ROBUST SPARSE LEARNING

- Beyond occlusion/corruption
  - Pose
  - Expression
  - Aging

- Beyond 1-D robust representation
Beyond small data
DICTIONARY LEARNING

- Beyond shallow dictionary learning
非常感谢各位！

Question?