Fast Classification of Handwritten Digits Using 2D-DCT Based Sparse PCA

Darya Ismailova and Wu-Sheng Lu

Department of Electrical and Computer Engineering
University of Victoria, Victoria, BC, Canada
Introduction

PCA for Multi-Category Classification

2-D DCT

2-D DCT-Based Sparse PCA

Application to Handwritten Digit Recognition

Conclusions
The problem of handwritten digit recognition (HWDR) has broad applications, where both accuracy and speed of digit recognition are critical indicators of system performance.

Problem itself: given a training data set \( \{D_j, j = 0, 1, \ldots, 9\} \) develop an approach to train a multi-class classifier to recognize the digit outside the training data.

The primary challenge of the HWDR problem: variation in the handwriting styles.
PCA for Multi-Category Classification

Given a data set \( \{x_i, i = 1, 2, \ldots, n\}, x_i \in \mathbb{R}^{m \times 1} \), its average vector and covariance matrix are defined as

\[
\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i, \quad C = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x})^T
\]

Since \( C \succeq 0 \) its singular value decomposition (SVD) is identical to its eigen-decomposition

\[
C = U S U^T
\]

where \( U = [u_1 \ u_2 \ldots u_m] \) is orthogonal, \( S = \text{diag}\{ \sigma_1, \sigma_2 \ldots \sigma_m \} \) with \( \sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_m \).
PCA for Multi-Category Classification

- An $L_2$-optimal rank-$K$ approximation of covariance matrix $C$ can be obtained as

$$C \approx U_K S_K U_K^T$$

- The usefulness of approximation may be understood from two perspectives:
  1. Dimension reduction from $R^m$ to $R^K$
  2. Supervised multi-category classification
Example of the supervised multi-category classification

(a) Plot in the Original Space $\mathbb{R}^3$

(b) Projection of Training Dataset to the 2-dimensional Subspace

**Figure:** Example “double semi-circle” data set for supervised classification of 1000 random samples.
PCA for Multi-Category Classification

Example of the supervised multi-category classification

(a) Plot in the Original Space $R^3$

(b) Projection of Training Dataset to the 1-dimentional Subspace

Figure: Example Data Set for Supervised Classification of 1000 Random Samples in Double Semi-Circle
Given a digital image of size $N$ by $N$ represented by its light intensity \( \{x(i,j), i,j = 1, 2, \ldots, N\} \), the 2-D DCT of the image is a 2-D array of the same size, denoted by \( \{D(k, l), k, l = 1, 2, \ldots, N\} \) where

\[
D(k, l) = \frac{2\alpha(k)\alpha(l)}{N} \sum_{i=0}^{N-1} \sum_{j=0}^{N-1} x(i,j) \cdot \cos \left( \frac{(2i + 1)k\pi}{2N} \right) \cos \left( \frac{(2j + 1)l\pi}{2N} \right)
\]

with

\[
\alpha(k) = \begin{cases} 
  \frac{1}{\sqrt{2}}, & \text{for } k = 0 \\
  1, & \text{for } k \neq 0
\end{cases}
\]
Important property of DCT is energy compaction.

Figure: (a) An example digit from MNIST database, (b) the 2-D DCT of the image of size 28 by 28 in (a), and (c) the 784 DCT coefficients as a 1-D sequence.
The main point is to use 2-D DCT at the pre-processing stage is to reduce the dimension $m$ of the input data space.

Given training data $\{x_i, i = 1, 2, \ldots, n\}$ with $x_i \in \mathbb{R}^{mx1}$, first re-shape vector $x_i$ to its original image size and apply 2-D DCT.

Convert the 2-D DCT coefficients to a 1-D sequence by zig-zag scanning the coefficients.

Retain the first $r$ DCT coefficients to construct $d_i \in \mathbb{R}^r$ thus constructing a reduced data set $\{d_i, i = 1, 2, \ldots, n\}$. 
The equation for PCA holds for reduced data set as:

$$C \approx U_K S_K U_K^T$$

$$\bar{d} = \frac{1}{n} \sum_{i=1}^{n} d_i, \quad C = \frac{1}{n-1} \sum_{i=1}^{n} (d_i - \bar{d})(d_i - \bar{d})^T$$

$$S_K = \text{diag}\{\sigma_1, \sigma_2, \ldots, \sigma_K\}, \quad U_K = [u_1 \ u_2 \ldots u_K]$$

The $L$ data classes $\{x^{(j)}_i, i = 1, 2, \ldots, n_j\}$ for $j = 0, 1, \ldots, L - 1$ are well represented by $L$ reduced “data” sets $\{\bar{d}_j \in \mathbb{R}^r, U_K^{(j)}\}$. 
To classify a test point $\mathbf{x} \in \mathbb{R}^m$:

(i) Apply 2-D DCT to the image constructed from $\mathbf{x}$ and keep the first $r$ DCT coefficients to construct vector $\mathbf{d} \in \mathbb{R}^r$;

(ii) Project point $\mathbf{d} - \overline{\mathbf{d}}_j$ into the $j$th data class: $\mathbf{z}_j = \mathbf{U}_K^{(j)} \mathbf{T} (\mathbf{d} - \overline{\mathbf{d}}_j)$;

(iii) Approximate point $\mathbf{d}$ in the $j$th class as $\hat{\mathbf{d}}_j = \mathbf{U}_K^{(j)} \mathbf{z}_j + \overline{\mathbf{d}}_j$;

(iv) Compute $e_j = \| \mathbf{d} - \hat{\mathbf{d}}_j \|$ for $j = 0, 1, \ldots, L - 1$;

(v) Classify point $\mathbf{x}$ to class $j^*$ if $e_{j^*}$ reaches the minimum among $\{e_j, j = 0, 1, \ldots, L - 1\}$. 
The MNIST Database

(a) Training set  
(b) Testing set

Figure: Typical images from the MNIST database
Generation of Input Data for HWDR

- The MNIST database: 60,000 labeled handwritten digits in the training set, 10,000 handwritten digits in the test set. Each data sample is a vector of length 784 representing a 28 by 28 gray-scale image of the digit.

- Input data for the algorithm: ten sets of training data $\mathcal{D}_j = \{(x^{(j)}_i, y_j), i = 1, 2, \ldots, n_j\}$ for $j = 0, 1, \ldots, 9$, where each $\mathcal{D}_j$ contains a total of $n_j$ digits representing the same numeral $j$. In our experiments, $n_j$ was set to 1200 for all sets and, for a fixed $j$, $\{x^{(j)}_i, i = 1, 2, \ldots, 1200\}$ were selected at random from those in the training data that collects all the digits representing numeral $j$. 

Darya Ismailova and Wu-Sheng Lu (UVic)
Appropriate ranges for $r$ and $K$ were found to be $180 \leq r \leq 400$ and $22 \leq K \leq 31$, respectively.

A smaller $r$ yields a faster classifier, but using an $r$ too small degrades recognition accuracy.
### Performance Evaluation of 2-D DCT Based Sparse PCA

<table>
<thead>
<tr>
<th></th>
<th>Sparse PCA</th>
<th>Conventional PCA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dim. of the classifier input</td>
<td>196</td>
<td>784</td>
</tr>
<tr>
<td>K</td>
<td>26</td>
<td>25</td>
</tr>
<tr>
<td>Accuracy</td>
<td>96.21%</td>
<td>96.26%</td>
</tr>
<tr>
<td>Normalized time</td>
<td>0.643</td>
<td>1</td>
</tr>
</tbody>
</table>
Conclusions

- A 2-D DCT-based sparse PCA classifier for handwritten digit recognition has been proposed.
- The ability of 2-D DCT to compress image-related signals allows a significant dimensionality reduction of the input space.
- The sparse PCA classifier is shown to perform HWDR considerably faster than the conventional PCA classifier without sacrificing recognition accuracy.
Q & A
Appendix
2-D DCT-Based Sparse PCA. The Algorithm

**Input:** Training data $D_j = \{(x_i^{(j)}, y_j), i = 1, 2, \ldots, n_j\}$ for $j = 0, 1, \ldots, L - 1$; target dimension of reduced input space $r$; number $K$ of principal components to be retained; and testing data $T$.

**Step 1:** Apply 2-D DCT to $D_j, j = 0, 1, \ldots, L - 1$ to obtain reduced data set $R_i = \{(d_i^{(j)}, y_j), i = 1, 2, \ldots, n_j\}$ of dimension $r$.

**Step 2:** For $j = 0, 1, \ldots, L - 1$ compute

$$
\overline{d}_j = \frac{1}{n} \sum_{i=1}^{n_j} d_i^{(j)}, \quad C_j = \frac{1}{n_j - 1} \sum_{i=1}^{n_j} (d_i^{(j)} - \overline{d}_j)(d_i^{(j)} - \overline{d}_j)^T
$$

and the $K$ eigenvectors $U_K^{(j)} = [u_1^{(j)} \ u_2^{(j)} \ldots u_K^{(j)}]$ associated with the $K$ largest eigenvalues of $C_j$. 
2-D DCT-Based Sparse PCA. The Algorithm

**Step 3:** For each vector $\mathbf{x}$ from test data $\mathcal{T}$:

(i) Compute vector $\mathbf{d} \in \mathbb{R}^r$ by applying 2-D DCT to the image constructed from $\mathbf{x}$ and retaining the $r$ most significant DCT coefficients (refer to Sec. 3.A);

(ii) Perform projections for $j = 0, 1, \ldots, L - 1$;

(iii) Compute approximating points $\hat{d}_j = U_K^{(j)} \mathbf{z}_j + \bar{d}_j$ for $j = 0, 1, \ldots, L - 1$;

(iv) Compute $e_j = \|d - \hat{d}_j\|$ for $j = 0, 1, \ldots, L - 1$;

(v) Classify point $\mathbf{x}$ to class $j^*$ if $e_{j^*}$ reaches the minimum among \{e$_j$, $j = 0, 1, \ldots, L - 1$\}. 


