Recognition, Detection, and Classification
(Part 2)
Computer Vision I
CSE 252A
Lecture 14
Announcements

• Assignment 3 is due Nov 22, 11:59 PM
• Assignment 4 will be released Nov 22
  – Due Dec 6, 11:59 PM
Simplest feature: Image as a Feature Vector

- Consider an n-pixel image to be a point in an n-dimensional space, $x \in \mathbb{R}^n$
- Each pixel value $x_i$ is a coordinate of $x$
Nearest Neighbor Classifier

n classes (different people)
{R_j : j=1,..., n} : set of training images, one per person

\[ \text{label} = \arg\min_j \text{dist}(R_j, I) \]
FIGURE 4.15. The $k$-nearest-neighbor query starts at the test point $x$ and grows a spherical region until it encloses $k$ training samples, and it labels the test point by a majority vote of these samples. In this $k = 5$ case, the test point $x$ would be labeled the category of the black points. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.
Nearest Neighbor Classification

- k Nearest Neighbor with $k=1$
- It can be shown that as number of samples approaches infinity, the error rate of nearest neighbor is at most twice that of the optimal Bayesian classifier. (Why is even close? Training samples are drawn from $P(x \mid \omega_i)P(\omega_i)$)
Comments on Nearest Neighbor

- Sometimes called “Template Matching”
- Variations on distance function (e.g., $L_1$, robust distances)
- Multiple templates per class - perhaps many training images per class
- Expensive to compute $k$ distances, especially when each image is big ($N$ dimensional)
- May not generalize well to unseen examples of class
- Will it apply well to other features (height, weight?)
- Some solutions:
  - Bayesian classification
  - Dimensionality reduction
• If we want to build a minimum-error rate classifier, then we need a very good estimate of $P(\omega_i | x)$

• How do we do this?
• Let’s say our feature space is just 1-dimensional and our feature \( x \in [0,1] \)
• Let’s say we have **10,000 training samples** from which to estimate our *a posteriori* probabilities
• We could estimate these probabilities using a histogram in which we divided the interval into 100 evenly spaced bins
• On average each bin would have 100 samples
• We could estimate \( P(x \mid \omega_i) \) as the number of samples from class \( i \) that fall in the same bin that falls into divided by the total number of samples in that bin
But this plan does not scale as we increase the dimensionality of the feature space!

• Let’s say our feature space is 3 dimensional and our feature \( \mathbf{x} \in [0,1]^3 \)

• Let’s say we still have 10,000 training samples from which to estimate our \( a \) \( posteriori \) probabilities

• If we estimate these probabilities using a histogram in which we divide the volume into the same width bins as before…
On average each bin would only have 0.01 samples! We’re not going be able to estimate probabilities well.
Curse of Dimensionality

Dimensionality Reduction
An idea:
Represent the set of images as a linear subspace

What is a linear subspace?

Let $V$ be a vector space and let $W$ be a subset of $V$. Then $W$ is a subspace if and only if:

1. The null vector $0$ is in $W$
2. If $u$ and $v$ are elements of $W$, then any linear combination of $u$ and $v$ is an element of $W$; $au + bv \in W$
3. If $u$ is an element of $W$ and $c$ is a scalar, then the scalar product $cu \in W$

- A $k$-dimensional subspace is spanned by $k$ linearly independent vectors. It is spanned by a $k$-dimensional orthogonal basis

Example: A 2-D linear subspace of $\mathbb{R}^3$ spanned by $y_1$ and $y_2$
Linear Subspaces & Linear Projection

- A $d$-pixel image $x \in \mathbb{R}^d$ can be projected to a low-dimensional feature space $y \in \mathbb{R}^k$ by

  $$y = Wx$$

  where $W$ is an $k \times d$ matrix

- Each training image is projected to the subspace

- Recognition is performed in $\mathbb{R}^k$ using, for example, nearest neighbor

- How do we choose a good $W$?

Example: A 2-D linear subspace of $\mathbb{R}^3$ spanned by $y_1$ and $y_2$
How do we choose a good $W$?

- Drop dimensions (feature selection)
- Random projections
- Principal Component Analysis
- Linear Discriminant Analysis
- Independent Component Analysis
- Or non-linear dimensionality reduction
Principal component analysis (PCA) of covariance matrix

Assume we have a set of \( n \) feature vectors \( \mathbf{x}_i \) \((i = 1, \ldots, n)\) in \( \mathbb{R}^d \). Write

\[
\mu_x = \frac{1}{n} \sum_i \mathbf{x}_i
\]

\[
\Sigma_x = \frac{1}{n-1} \sum_i (\mathbf{x}_i - \mu)(\mathbf{x}_i - \mu)^T
\]

Eigen decomposition of covariance matrix

\[
\Sigma_x = \mathbf{V} \Lambda \mathbf{V}^T
\]

where

\( \Sigma_x \) is a positive semidefinite \( n \times n \) matrix

\( \mathbf{V} \) is an \( n \times n \) orthogonal matrix

\( \Lambda = \text{diag}(\lambda_1, \ldots, \lambda_n) \), where \( \lambda_i \geq \lambda_{i+1} \geq 0 \)

Columns of \( \mathbf{V} \) are eigenvectors (also called principal component coefficients) corresponding to eigenvalues (also called principal component variances) \( \lambda = (\lambda_1, \ldots, \lambda_n)^T \).
Principal component analysis identifies a collection of linear features that are independent, and capture as much variance as possible from a dataset.
Relationship between singular value decomposition (SVD) and eigen decomposition

\[ A = U \Sigma V^T \]
\[ A^T A = (U \Sigma V^T)^T U \Sigma V^T \]
\[ A^T A = V \Sigma^T U^T U \Sigma V^T \]
\[ A^T A = V \Lambda V^T, \text{ where } \Lambda = \Sigma^T \Sigma \]
\[ A A^T = U \Sigma V^T (U \Sigma V^T)^T \]
\[ A A^T = U \Sigma V^T V \Sigma^T U^T \]
\[ A A^T = U \Sigma \Sigma^T U^T \]
\[ A A^T = U \Lambda U^T, \text{ where } \Lambda = \Sigma \Sigma^T \]

where

\( U \) and \( V \) are orthogonal matrices
\( \Sigma = \text{diag}(\sigma_1, \ldots, \sigma_n) \), where \( \sigma_i \geq \sigma_{i+1} \geq 0 \)
\( \Lambda = \text{diag}(\lambda_1, \ldots, \lambda_n) \), where \( \lambda_i \geq \lambda_{i+1} \geq 0 \)
variances \( \lambda_i = \sigma_i^2 \forall i \)
Data matrix

\[ X \in \mathbb{R}^{m \times n} \quad m \text{ observations, } n \text{ variables} \]

\[ X = \begin{bmatrix}
  x^{1\top} \\
  x^{2\top} \\
  \vdots \\
  x^{m\top}
\end{bmatrix} \]

Data matrix, mean-deviation form

\[ \hat{X} = \begin{bmatrix}
  x^{1\top} - \mu_x^\top \\
  x^{2\top} - \mu_x^\top \\
  \vdots \\
  x^{m\top} - \mu_x^\top
\end{bmatrix} \]

\[ \Sigma_x = \frac{1}{m - 1} \hat{X}^\top \hat{X} \]
Singular value decomposition of (mean-deviation form of) data matrix

\[ \hat{X} = UDV^\top \]

Note: economy SVD can be used

where

\[ D = \text{diag}(\sigma_1, \ldots, \sigma_{\min(m,n)}), \text{ where } \sigma_i \geq \sigma_{i+1} \geq 0 \]

columns of \( V \) (rows of \( V^\top \)) are principal component coefficients of \( \hat{X} \)

Projection of \( \hat{X} \) to principal component axes

\[ \hat{A} = \hat{X}V \] (forward) projection to principal component scores

\[ \hat{A}V^\top = \hat{X} \] back projection

Dimensionality reduction

Columns of \( V \) corresponding to smallest singular values can be removed

\[ \hat{A}' = \hat{X}V' \] (forward) projection to principal component scores

\[ \hat{A}'V'^\top = \hat{X}' \] back projection (with loss of data)
SVD Properties

• $r = \text{rank}(A) = \text{number of non-zero singular values}$
• $U, V$ give an orthonormal bases for the subspaces of $A$:
  – 1st $r$ columns of $U$: Column space of $A$
  – Last $m - r$ columns of $U$: Left nullspace of $A$
  – 1st $r$ columns of $V$: Row space of $A$
  – 1st $n - r$ columns of $V$: (Right) nullspace of $A$

• *For some $d$ where $d \leq r$, the first $d$ column of $U$ provide the best $d$-dimensional basis for columns of $A$ in least squares sense.*
PCA for recognition

Modeling
1. Given a collection of $n$ training images $x_i$, represent each one as a $d$-dimensional column vector.
2. Compute the mean image and covariance matrix.
3. Compute $k$ Eigenvectors of the covariance matrix corresponding to the $k$ largest Eigenvalues and form matrix $W^T = [u_1, u_2, ..., u_k]$ (Or perform using SVD)
   - Note that the Eigenvectors are images.
4. Project the training images to the $k$-dimensional Eigenspace. $y_i = Wx_i$

Recognition
1. Given a test image $x$, project the vectorized image to the Eigenspace by $y = Wx$
2. Perform classification of $y$ to the projected training images.
Example: Eigenfaces

Training images

[ Turk, Pentland 91]
Eigenfaces

Mean Image

Basis Images
# Accuracy of PCA + K-NN

<table>
<thead>
<tr>
<th>Method</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>KNN + 3,072 Features</td>
<td>33.86</td>
</tr>
<tr>
<td>KNN + 200 PCA Comp.</td>
<td>36.54</td>
</tr>
<tr>
<td>KNN + 75 PCA Comp.</td>
<td>39.77</td>
</tr>
<tr>
<td>KNN + 50 PCA Comp.</td>
<td>40.12</td>
</tr>
<tr>
<td>KNN + 40 PCA Comp.</td>
<td>40.93</td>
</tr>
<tr>
<td><strong>KNN + 30 PCA Comp.</strong></td>
<td><strong>41.78</strong></td>
</tr>
<tr>
<td>KNN + 25 PCA Comp.</td>
<td>41.57</td>
</tr>
<tr>
<td>KNN + 15 PCA Comp.</td>
<td>38.75</td>
</tr>
<tr>
<td>KNN + 10 PCA Comp.</td>
<td>34.93</td>
</tr>
</tbody>
</table>
Difficulties with PCA

• Projection may suppress important detail
  – smallest variance directions may not be unimportant

• Method does not take discriminative task into account
  – typically, we wish to compute features that allow good discrimination
  – not the same as largest variance or minimizing reconstruction error
Fisherfaces: Class specific linear projection


• An $n$-pixel image $\mathbf{x} \in \mathbb{R}^d$ can be projected to a low-dimensional feature space $\mathbf{y} \in \mathbb{R}^k$ by

$$\mathbf{y} = W\mathbf{x}$$

where $W$ is an $k$ by $d$ matrix

• Recognition is performed using nearest neighbor in $\mathbb{R}^k$

• How do we choose a good $W$?
PCA & Fisher’s Linear Discriminant

• Between-class scatter
  \[ S_B = \sum_{i=1}^{c} |\chi_i| (\mu_i - \mu)(\mu_i - \mu)^T \]

• Within-class scatter
  \[ S_W = \sum_{i=1}^{c} \sum_{x_k \in \chi_i} (x_k - \mu_i)(x_k - \mu_i)^T \]

• Total scatter
  \[ S_T = \sum_{i=1}^{c} \sum_{x_k \in \chi_i} (x_k - \mu)(x_k - \mu)^T = S_B + S_W \]

• Where
  - \( c \) is the number of classes
  - \( \mu_i \) is the mean of class \( \chi_i \)
  - \( |\chi_i| \) is number of samples of \( \chi_i \).

If the data points \( x_i \) are projected by \( y_i = Wx_i \) and the scatter of \( x_i \) is \( S \), then the scatter of the projected points \( y_i \) is \( W^T S W \).
PCA & Fisher’s Linear Discriminant

- **PCA (Eigenfaces)**
  
  \[ W_{PCA} = \arg\max_W |W^T S_T W| \]
  
  Maximizes projected total scatter

- **Fisher’s Linear Discriminatant**
  
  \[ W_{fld} = \arg\max_W \frac{|W^T S_B W|}{|W^T S_W W|} \]
  
  Maximizes ratio of projected between-class to projected within-class scatter
Computing the Fisher Projection Matrix

\[ W_{opt} = \arg \max_W \frac{|W^T S_B W|}{|W^T S_W W|} \]

\[ = [w_1 \ w_2 \ ... \ w_m] \] 

(4)

where \( \{w_i \mid i = 1, 2, \ldots, m\} \) is the set of generalized eigenvectors of \( S_B \) and \( S_W \) corresponding to the \( m \) largest generalized eigenvalues \( \{\lambda_i \mid i = 1, 2, \ldots, m\} \), i.e.,

\[ S_B w_i = \lambda_i S_W w_i, \quad i = 1, 2, \ldots, m \]

• There are at most \( c-1 \) non-zero generalized Eigenvalues, so \( m \leq c-1 \)
Generalized eigenvalues and generalized eigenvectors for a pair of symmetric matrices

\[ AV = BVA \quad \text{and} \quad V^\top A = \Lambda V^\top B \]

where

- $A$ is symmetric
- $B$ is (symmetric) positive definite
- $V = [v_1 | \cdots | v_n]$ is not orthogonal
- $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_n)$

Eigenvalue $\lambda_i$ corresponds to eigenvector $v_i$.

$S_B$ is $A$
$S_W$ is $B$
Fisherfaces

\[ W = W_{fld} W_{PCA} \]

\[ W_{PCA} = \arg \max_{W} \left| W^{T} S_{T} W \right| \]

\[ W_{fld} = \arg \max_{W} \frac{W^{T} W_{PCA}^{T} S_{B} W_{PCA} W}{W^{T} W_{PCA}^{T} S_{W} W_{PCA} W} \]

- Since \( S_{W} \) is rank \( N-c \), project training set to subspace spanned by first \( N-c \) principal components of the training set.
- Apply FLD to \( N-c \) dimensional subspace yielding \( c-1 \) dimensional feature space.

- Fisher’s Linear Discriminant projects away the within-class variation (lighting, expressions) found in training set.
- Fisher’s Linear Discriminant preserves the separability of the classes.
Harvard Face Database

- 10 individuals
- 66 images per person
- Train on 6 images at 15°
- Test on remaining images
Recognition Results: Lighting Extrapolation

![Graph showing error rate against light direction for different methods: Correlation, Eigenfaces, Eigenfaces (w/o 1st 3), Fisherface.](image-url)
Where have we been so far

- Supervised classification
  - Feature space
  - Nearest Neighbor (kth nearest neighbor)
  - Bayesian (MAP) classifier
- Curse of dimensionality
- Dimensionality reduction
  - Principal component analysis
  - Fisher’s linear discriminant
Example of Feature Extraction

CoralNet
https://coralnet.ucsd.edu/
A Representative Survey

Typical Survey:

- 1,250 images acquired over two weeks
- Goal: Coverage of dominant functional groups
- Annotated with Coral Point Count (CPCe)
- 200 annotations per image
- 250,000 annotations
- 6-9 months to annotate

Pete Edmunds and Vincent Moriarty, Cal State Northridge
Traditional Computer Vision Method

• Color/Texture classification with
  – Preprocess
  – Filtered image + color => feature vector at each pixel.
  – Bag of Visual Words
  – Pool over different sized regions
  – SVM for classification

Automatic Annotation of Survey Images
Step 1. Preprocess

STEP 2A. FILTERING

**STEP2B. FILTERING**

**Step 3. Each Pixel is Mapped to a Visual Word**

\[ w_i \in \mathbb{R}^{24} \quad T_P = \arg \min_i ||w_i - F_P||_2 \]

1. **Image Filter Response**
2. **Texton Map**
3. 135 words in `dictionary`
Step4a. Histograms

The image shows a network of connections with a square highlighted, accompanied by a bar graph indicating normalized histogram count with 135 bins.
STEP4B. HISTOGRAMS AT MULTIPLE SCALES

Normalized histogram count
Next Lectures

• Neural networks