Basic Concepts of Pattern Recognition and Feature Selection

Xiaojun Qi

-- REU Site Program in CVMA (2010 Summer)

Outline

- Pattern Recognition
  - Pattern vs. Features
  - Pattern Classes
  - Classification
- Feature Selection Techniques
  - PCA

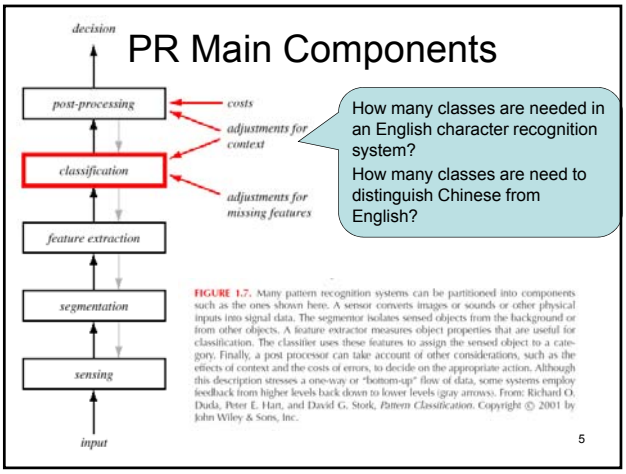
Introduction to Pattern Recognition

- Pattern Recognition (PR): Classify what inside of the image
- Applications:
  - Speech Recognition/Speaker Identification
  - Fingerprint/Face Identification
  - Signature Verification
  - Character Recognition
  - Biomedical: DNA Sequence Identification
  - Remote Sensing
  - Meteorology
  - Industrial Inspection
  - Robot Vision

Introduction to PR (Cont.)

- Pattern recognition deals with classification, description, and analysis of measurements taken from physical or mental processes.
- Pattern recognition
  - Take in raw data
  - Determine the category of the pattern
  - Take an action based on the category of the pattern

PR Main Components



How many classes are needed in an English character recognition system?

How many classes are need to distinguish Chinese from English?

FIGURE 1.7. Many pattern recognition systems can be partitioned into components such as the ones shown here. A sensor converts images or sounds or other physical inputs into signal data. The segmentor isolates sensed objects from the background or from other objects. A feature extractor measures object properties that are useful for classification. The classifier uses these features to assign the sensed object to a category. Finally, a post processor can take account of other considerations, such as the effects of context and the costs of errors, to decide on the appropriate action. Although this description stresses a one-way or "bottom-up" flow of data, some systems employ feedback from higher levels back down to lower levels (gray arrows). From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

PR Main Components (Cont.)

- Sensing: Design of transducers
- Segmentation and grouping (into a composite object)
- Feature extraction: A set of characteristic measurements (numerical or non-numerical), and their relations are extracted to represent patterns for further process.
- Classification: The process or events with same similar properties are grouped into a class. The number of classes is task-dependent.
- Post-processing: Considering the effects of context and the cost of errors

PR Example

- Fish-packing plant -- species determination
  - Separate sea bass from salmon using optical sensing
  - Image features
    - Length
    - Lightness
    - Width
    - Number and shape of fins, etc.
  - Establish models for objects to be classified
    - Descriptions in mathematics form
    - Considering noise or variations in population itself and sensing

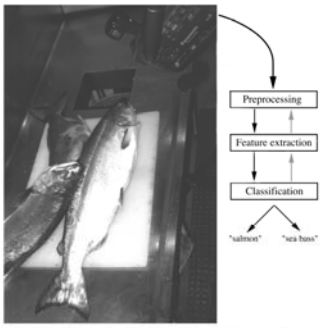


FIGURE 1.1. The objects to be classified are first sensed by a transducer (camera), whose signals are preprocessed. Next the features are extracted and finally the classification is emitted, here either "salmon" or "sea bass." Although the information flow is often chosen to be from the source to the classifier, some systems employ information flow in which earlier levels of processing can be altered based on the tentative or preliminary response in later levels (gray arrows). Yet others combine two or more stages into a unified step, such as simultaneous segmentation and feature extraction. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

First Feature Extraction

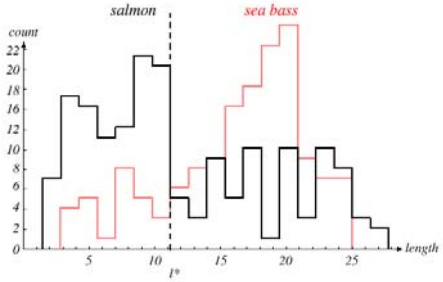


FIGURE 1.2. Histograms for the length feature for the two categories. No single threshold value of the length will serve to unambiguously discriminate between the two categories; using length alone, we will have some errors. The value marked l^* will lead to the smallest number of errors, on average. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Second Feature Extraction

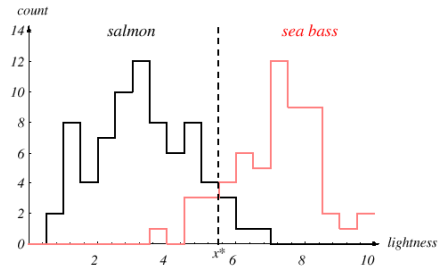


FIGURE 1.3. Histograms for the lightness feature for the two categories. No single threshold value x^* (decision boundary) will serve to unambiguously discriminate between the two categories; using lightness alone, we will have some errors. The value x^* marked will lead to the smallest number of errors, on average. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Classification -- Classifier Design

- Feature space
  - Feature vector

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

- Scattering plot for training samples
- Classifier : design of decision boundary on scattering plot
  - Partition the feature space into several regions.

Linear Classifier

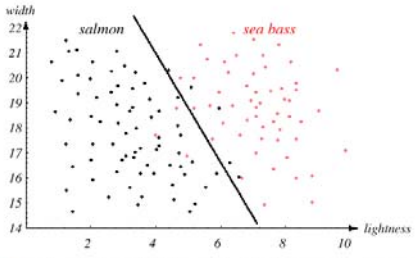


FIGURE 1.4. The two features of lightness and width for sea bass and salmon. The dark line could serve as a decision boundary of our classifier. Overall classification error on the data shown is lower than if we use only one feature as in Fig. 1.3, but there will still be some errors. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

The Best Classifier

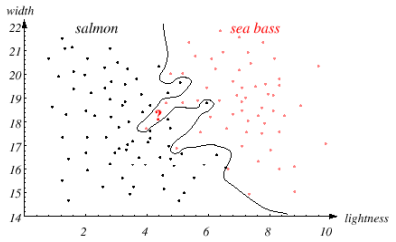


FIGURE 1.5. Overly complex models for the fish will lead to decision boundaries that are complicated. While such a decision may lead to perfect classification of our training samples, it would lead to poor performance on future patterns. The novel test point marked **?** is evidently most likely a salmon, whereas the complex decision boundary shown leads it to be classified as a sea bass. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

The Optimal Classifier

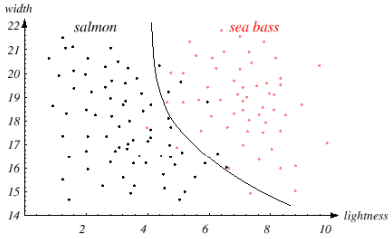


FIGURE 1.6. The decision boundary shown might represent the optimal tradeoff between performance on the training set and simplicity of classifier, thereby giving the highest accuracy on new patterns. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

14

PR Feature Extraction

- Seek distinguishing features that are invariant to irrelevant transformations.
  - Distinguishing features
    - Feature values are similar in the same category and very different in different categories.
  - Irrelevant transformations
    - Rotation, Scale, and Translation, (RST invariance, major concern)
    - Occlusion
    - Projective distortion
    - Non-rigid deformations
- Feature selection (those are most effective)

15

PR Classification

- Assign an object to a category by using the feature vector.
- Difficulty of classification depends on the variability of the feature values in the same category relative to the difference between feature values in different categories.
- The variability of feature values in the same category may come from noise

16

PR Post-Processing

- Consider the cost of action
  - Minimize classification error rate
  - Minimize risk (total expected cost)
- Exploit context (input-dependent information) to improve system performance
  - E.g., use the context information for OCR or speech recognition
- Multiple classifier (different from multiple features)
  - Each classifier operates on different aspects of the input (e.g., speech recognition = acoustic recognition + lip reading)
  - Decision fusion

17

Areas Related to PR

- Image Processing
- Speech Processing
- Artificial Intelligence
- Associate Memory
- Neural and Fuzzy
- Probability and Statistics (Statistical)
  - Regression (find functional description of data for new input prediction)
  - Interpolation (infer the function for intermediate ranges of input)
  - Density estimation (for Maximum Likelihood and Maximum A Prior classification)
- Formal language (Syntactic)
- Neural network architecture design and training (Neural)

18

Concepts of Pattern Recognition

- Pattern: A pattern is the description of an object.
- Pattern Class: It is a category determined by some given common attributes.
- Pattern recognition can be defined as the categorization of input data into identifiable classes via the extraction of significant features or attributes of the data from a background of irrelevant detail.
- The problem of pattern recognition may be regarded as one of discriminating of the input data, not between individual patterns but between populations, via the search for features or invariant attributes among members of a population.

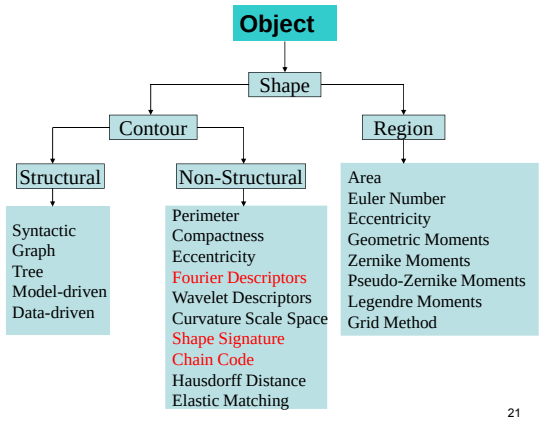
19

Object Feature -- Shape

- **What features can we get from an Object?**

- Perimeter
- Area
- Eccentricity: The ratio of the major to the minor axis
- Curvature: The rate of change of slope. That is: Use the difference between the slopes of adjacent boundary segments as a descriptor of curvature at the point of intersection of the segments.
- Chain Code…

20



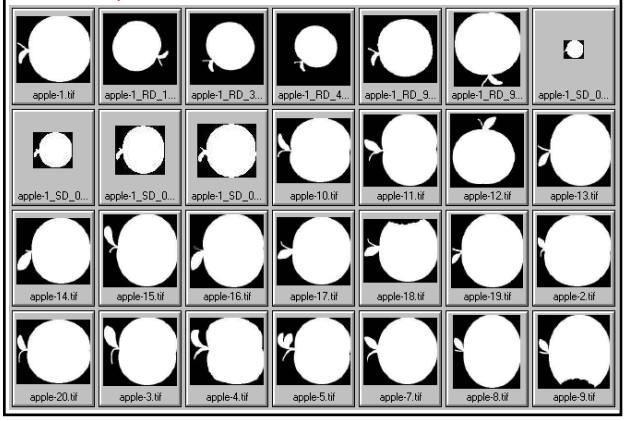
21

Object Representation and Recognition

- Representing a region involves two choices:
  - We can represent the region in terms of its external characteristics (its boundary)
    - Focus on Shape Characteristics.
  - We can represent the region in terms of its internal characteristics (the pixels comprising the region)
    - Focus on regional properties such as color and texture.
- Describe the region based on the chosen representation.
  - A region may be represented by its boundary, and the boundary can be described by features such as:
    - Length,
    - The orientation of the straight line joining its extreme points,
    - The number of concavities in the boundary.

22

The features selected as descriptors should be as insensitive as much as possible to variations in size, translation, and rotation.



Shape Signatures

- A signature is a 1-D functional representation of a boundary and may be represented in various ways.
- Regardless of how a signature is generated, the basic idea is to reduce the boundary representation to a 1-D function, which presumably is easier to describe than the original 2-D boundary. Some sample shape signatures are:
  1. Complex Coordinates
  2. Central Distance
  3. Central Complex Coordinates
  4. Chordlength
  5. Cumulative Angular Function
  6. Curvature Function
  7. Area Function

24

Shape Signatures --Complex Coordinates

- The boundary can be represented as the sequence of coordinates s(t) = [x(t), y(t)] for t = 0, 1, 2, …, N-1, where x(t) = x_t and y(t) = y_t; (x_t, y_t)'s are encountered in traversing the boundary in the counterclockwise direction and N is the total number of points on the boundary.

Z(t) = x(t) + iy(t)

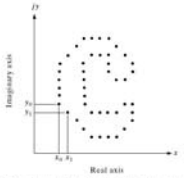
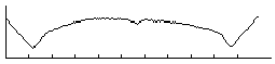


FIGURE 11.13 A digital boundary and its representation as a complex sequence. The points (x_0, y_0) and (x_1, y_1) shown are (arbitrarily) the first two points in the sequence.

Shape Signatures -- Central Distance

$$r(t) = ([x(t) - x_c]^2 + [y(t) - y_c]^2)^{1/2}$$

$$x_c = \frac{1}{N}\sum_{t=0}^{N-1} x(t), \quad y_c = \frac{1}{N}\sum_{t=0}^{N-1} y(t)$$



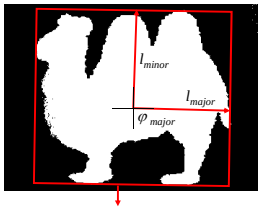
Shape Signatures --Central Complex Coordinates

$$z(t) = [x(t) - x_c] + i[y(t) - y_c]$$

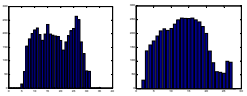
$$x_c = \frac{1}{N}\sum_{t=0}^{N-1} x(t), \quad y_c = \frac{1}{N}\sum_{t=0}^{N-1} y(t)$$

Object Feature: Possible Boundary Features (1)

- **Simple features**
  - Area : *A*
  - Circumference: *r*
  - Euler's number: #parts - #holes
  - Direction: φ_{major}
  - Eccentricity: $||l_{major}|| / ||l_{minor}||$
  - Elongatedness: w_{BB} / h_{BB}
  - Rectangularity: A / A_{BB}
  - Compactness: r^2 / A
  - Gray value/color/texture statistics
  - Projections
- Not all invariant



$w_{BB} \times h_{BB}$ bounding box



Object Feature: Possible Boundary Features (2)

- **Moments**

$$m_{pq} = \int_{x=-\infty}^{\infty}\int_{y=-\infty}^{\infty} x^p y^q I(x,y)\,dxdy$$

- 0^{th} order (i.e., p + q = 0) : size
- 1^{st} order (i.e., p + q = 1) : center-of-mass
- 2^{nd} order (i.e., p + q = 2): orientation
- higher order : shape

Object Feature: Possible Boundary Features (2) Cont.

- Central moments: translation invariant

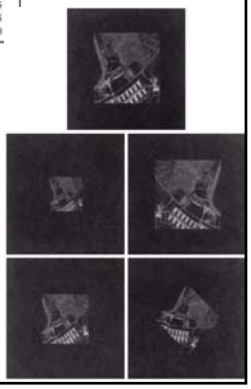
$$\mu_{pq} = \int_{x=-\infty}^{\infty}\int_{y=-\infty}^{\infty} (x-\bar{x})^p (y-\bar{y})^q I(x,y)\,dxdy$$

$$= \int_{x=-\infty}^{\infty}\int_{y=-\infty}^{\infty} (x-\frac{m_{10}}{m_{00}})^p (y-\frac{m_{01}}{m_{00}})^q I(x,y)\,dxdy$$

- 7 Hu moments: translation, rotation, scale invariant
- 47 Zernike moments: translation, rotation, scale invariant

Invariant (Log)	Original	Half Size	Mirrored	Rotated 2°	Rotated 45°
ϕ_1	6.249	6.226	6.919	6.253	6.318
ϕ_2	17.180	16.954	19.955	17.270	16.803
ϕ_3	22.655	23.531	26.689	22.836	19.724
ϕ_4	22.919	24.236	26.901	23.130	20.437
ϕ_5	45.749	48.349	53.724	46.136	40.525
ϕ_6	31.830	32.916	37.134	32.068	29.315
ϕ_7	45.589	48.343	53.590	46.017	40.470

TABLE 11.3 Moment invariants for the images in Figs. 11.25(a)-(e).



1) Seven moments invariants are calculated for each of these images, and the logarithm of the results are taken to reduce the dynamic range.
2) As the Table shows, the results are in reasonable agreement with the invariants computed for the original image.
3) The major cause of error can be attributed to the digital nature of the data, especially for the rotated images.

Object Feature: Possible Boundary Features (3) -- Fourier Descriptors

- Fourier Transform of the Signature s(t):

$$u_n = \frac{1}{N}\sum_{t=0}^{N-1} s(t)e^{-j2\pi nt/N}$$

$$\text{for } n = 0,1,\ldots,N-1$$

The complex coefficients u_n are called the Fourier descriptors of the boundary, and are denoted as FDn.

Object Feature: -- Fourier Descriptors

- The inverse Fourier transform of these coefficients restores s(t). That is:

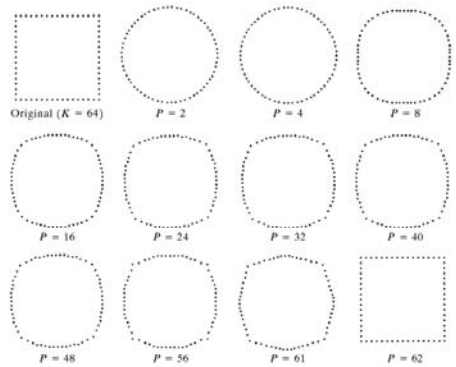
$$s(t) = \sum_{n=0}^{N-1} u_n e^{j2\pi nt/N} \quad \text{for t} = 0,1,\ldots,\text{N-1}$$

- Suppose that only the first P coefficients are used (that is, setting $u_n = 0$ for n > P-1). The result is the following approximation to s(k):

$$\hat{s}(t) = \sum_{n=0}^{P-1} u_n e^{j2\pi nt/N}$$

for t = 0, 1, 2, …, N-1.

FIGURE 11.14 Examples of reconstruction from Fourier descriptors. *P* is the number of Fourier coefficients used in the reconstruction of the boundary.



The goal is to use a few Fourier descriptors to capture the gross essence of a boundary. These coefficients carry shape information and can be used as the basis for differentiating between distinct boundary shapes.

Transformation	Boundary	Fourier Descriptor
Identity	$s(k)$	$a(u)$
Rotation	$s_r(k) = s(k)e^{j\theta}$	$a_r(u) = a(u)e^{j\theta}$
Translation	$s_t(k) = s(k) + \Delta_{xy}$	$a_t(u) = a(u) + \Delta_{xy}\delta(u)$
Scaling	$s_s(k) = \alpha s(k)$	$a_s(u) = \alpha a(u)$
Starting point	$s_p(k) = s(k - k_0)$	$a_p(u) = a(u)e^{-j2\pi k_0 u/K}$

$$s_t(k) = [x(k) + \Delta x] + j[y(k) + \Delta y)$$

$$s_p(k) = x(k - k_0) + jy(k - k_0)$$

1) Magnitude |FDn| is translation and rotation invariant
2) |FD0| carries scale-information
3) "Low-frequency" terms (*t* small): smooth behavior
4) "High-frequency" terms (*t* large): jaggy, bumpy behavior

Object Feature: -- Normalized Fourier Descriptor

$$f = \left[\frac{|FD_1|}{|FD_0|}, \frac{|FD_2|}{|FD_0|}, \ldots, \frac{|FD_m|}{|FD_0|}\right]$$

Why?

When two shapes are compared, m=N/2 coefficients are used for central distance, curvature and angular function.

m=N coefficients are used for complex coordinates.

$$d = \sqrt{\sum_{i=1}^{m}(f_i^q - f_i^t)^2}$$

$$\text{where } f_q = (f_q^1, f_q^2, \ldots, f_q^m) \text{ and } f_t = (f_t^1, f_t^2, \ldots, f_t^m)$$

are the feature vectors of the two shapes respectively.

- Complex FFT example:

A = [2 3 4 4] ;
B = [1 3 4 7] ;
C = A + B * i ;
fft(A) = [13 -2 + i -1 -2 – i] ;
fft(B) = [15 -3 + 4i -5 -3 - 4i] ;
fft(C) = [13 +15i -6 - 2i -1 - 5i 2-4i] ;

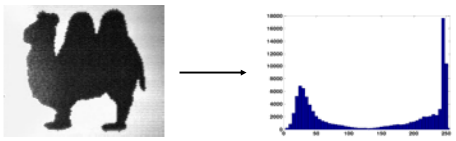
37

- Criteria for shape representation
  - Rotation, scale and translation Invariant
  - Compact & easy to derive
  - Perceptual similarity
  - Robust to shape variations
  - Application Independent
- FD satisfies all these criteria
- Problem
  - Different shape signatures can be used to derive FD, which is the best?

38

Object Feature: Possible Regional Features (1) -- Color Histogram

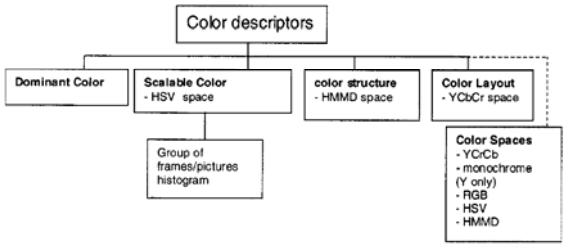
- **Histogram** (gray-scale images)



  - Invariant to translation, rotation, and small variations
  - Normalized histogram is invariant to scale
  - Not very sensitive to noise
  - But: removes a lot of information!

39

Object Feature: Possible Regional Features (2) -- MPEG-7 Color Descriptors



40

Object Feature: -- MPEG-7 Scalable Color Descriptor

- A color histogram in HSV color space
- Encoded by Haar Wavelet transform

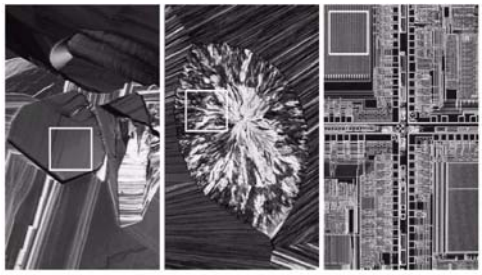
41

Object Feature: Possible Regional Features (3) -- Texture Features

- The three principal approaches used in image processing to describe the texture of a region are statistical, structural, and spectral.
  - Statistical approaches yield characterizations of textures as smooth, coarse, grainy, and so on.
  - Structural techniques deal with the arrangement of image primitives, such as the description of texture based on regularly spaced parallel lines.
  - Spectral techniques are based on properties of the Fourier spectrum and are used primarily to detect global periodicity in an image by identifying high-energy, narrow peaks in the spectrum.

42

Object Feature: -- Statistical Approaches



a b c

FIGURE 11.22 The white squares mark, from left to right, smooth, coarse, and regular textures. These are optical microscope images of a superconductor, human cholesterol, and a microprocessor. (Courtesy of Dr. Michael W. Davidson, Florida State University.)

TABLE 11.2
Texture measures for the subimages shown in Fig. 11.22.

Texture	Mean	Standard deviation	R (normalized)	Third moment	Uniformity	Entropy
Smooth	82.64	11.79	0.002	−0.105	0.026	5.434
Coarse	143.56	74.63	0.079	−0.151	0.005	7.783
Regular	99.72	33.73	0.017	0.750	0.013	6.674

$$\text{Mean}: m = \sum_{i=0}^{L-1} z_i\, p(z_i)$$

$$\text{Standard Deviation}: \sigma = \sqrt{\sum_{i=0}^{L-1} (z_i - m)^2\, p(z_i)}$$

$$\text{R}: R = 1 - \frac{1}{1+\sigma^2(z)}$$

$$\text{Third Moment}: \mu_3(z) = \sum_{i=0}^{L-1} (z_i - m)^3\, p(z_i)$$

$$\text{Uniformity}: U = \sum_{i=0}^{L-1} p^2(z_i)$$

$$\text{Entropy}: e = -\sum_{i=0}^{L-1} p(z_i) \log_2 p(z_i)$$

Standard Deviation is a measure of gray-level contrast that can be used to establish descriptors of relative smoothness.

R: Normalized variance in the range of [0, 1]

Third moment is a measure of the skewness of the histogram.

Uniformity: Histogram based measure.

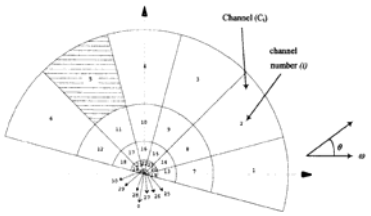
Average Entropy is a measure of variability and is 0 for a constant image.

Object Feature: -- MPEG-7 Homogenous Texture Descriptor

- Partition the frequency domain into 30 channels (modeled by a 2D-Gabor function)
- Compute the energy and energy deviation for each channel
- Compute mean and standard variation of frequency coefficients
- F = {f_{DC}, f_{SD}, e_1,…, e_{30}, d_1,…, d_{30}}
- An efficient implementation:
  - Radon transform followed by Fourier transform

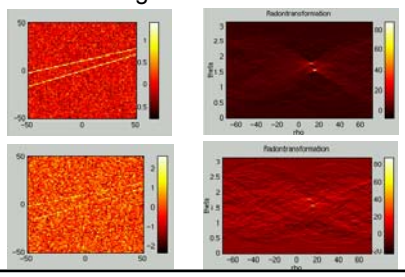
2-D Gabor Function

- It is a Gaussian weighted sinusoid
- It is used to model individual channels
- Each channel filters a specific type of texture



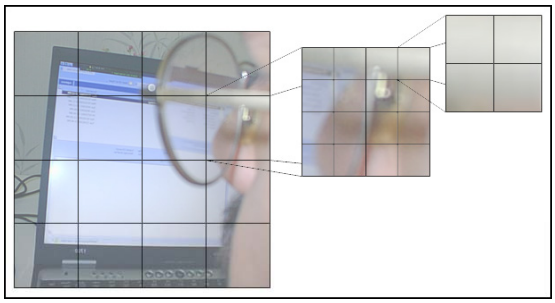
Radon Transform

- Transform images with lines into a domain of possible line parameters
- Each line will be transformed to a peak point in the resulted image



Object Feature: MPEG-7 Non-Homogenous Texture Descriptor

- Represent the spatial distribution of five types of edges
  - vertical, horizontal, 45°, 135°, and non-directional
- Dividing the image into 16 (4x4) blocks
- Generating a 5-bin histogram for each block
- It is scale invariant

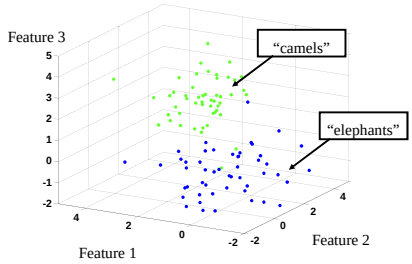


1	-1	1	1	$\sqrt{2}$	0	0	$\sqrt{2}$	2	-2
1	-1	-1	-1	0	$-\sqrt{2}$	$-\sqrt{2}$	0	-2	2

49

Feature Space (1)

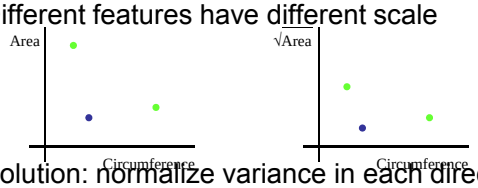
- End result: a *k*-dimensional space,
  - –in which each dimension is a **feature**
  - –containing *N* (labeled) **samples** (objects)



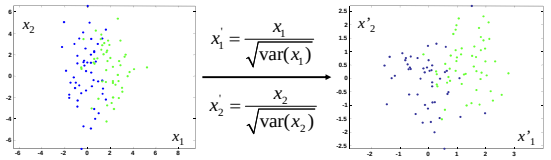
50

Feature Space (2)

- Different features have different scale



- Solution: normalize variance in each direction



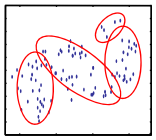
51

Feature Space (3)

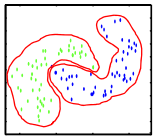
- Pattern recognition

What is our basic problem?

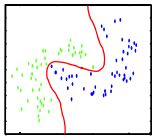
Clustering: find natural groups of samples in unlabelled data



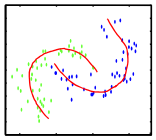
Density estimation: make a statistical model of the data



Classification: find functions separating the classes



Regression: fit lines or other functions to data (not in this course)



52

Summary

- Features are derived from measurements
- Application-dependent knowledge tells what features are important
- Invariance is important to make discrimination easier
- Recognition:
  - –Noise removal
  - –Shading removal
  - –Segmentation and labeling
  - –Features: Simple, Skeletons, Moments, Polygons, Fourier descriptors, ….

53

Examples of Automatic Pattern Recognition Systems

Object Recognition – Matching Technique

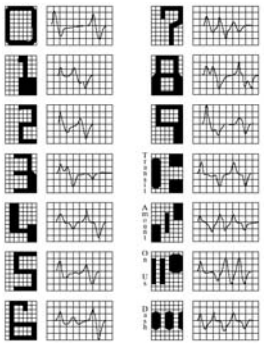


FIGURE 12.7 American Bankers Association E-13B font character set and corresponding waveforms.

54

Feature Selection: Principal Components Analysis (PCA)

- Consider a data set D={x_1, x_2, …, x_M} of N-dimensional vectors. This data set can be a set of M face images.
- The mean and the covariance matrix is given by

$$\mu = \frac{1}{M}\sum_{m=1}^{M} x_m$$

$$\sum = \frac{1}{M}\sum_{m=1}^{M}[x_m - \mu][x_m - \mu]^T$$

- Where the covariance matrix is an NxN symmetric matrix. This matrix characterizes the scatter of the data set.

Example

A = [0 0 0 ;
1 0 0 ;
1 1 0 ;
1 0 1 ;]

Here: The dimension N = 3 and M = 4. That is, A (i.e., a data set) contains a set of 4 vectors, each of which has 3 elements.

X1 = [0 0 0]' ;

X2 = [1 0 0]' ;

X3 = [1 1 0]' ;

X4 = [1 0 1]' ;

μ = [0.75 0.25 0.25] ;

Mx1 = X1 - μ = [-0.75 -0.25 -0.25]' ;

Mx2 = X2 - μ = [0.25 -0.25 -0.25]' ;

Mx3 = X3 - μ = [0.25 0.75 -0.25]' ;

Mx4 = X4 - μ = [0.25 -0.25 0.75]' ;

Example (Cont.)

(X1 – μ) (X1 - μ)^T = [0.5625 0.1875 0.1875 ;
0.1875 0.0625 0.0625 ;
0.1875 0.0625 0.0625];

(X2 – μ) (X2 - μ)^T = [0.0625 -0.0625 -0.0625;
-0.0625 0.0625 0.0625 ;
-0.0625 0.0625 0.0625 ;]

(X3 – μ) (X3 - μ)^T = [0.0625 0.1875 -0.0625;
0.1875 0.5625 -0.1875 ;
-0.0625 -0.1875 0.0625 ;]

(X4 – μ) (X4 - μ)^T = [0.0625 -0.0625 0.1875 ;
-0.0625 0.0625 -0.1875 ;
0.1875 -0.1875 0.5625 ;]

Example (Cont.)

$$\sum = \begin{bmatrix} 0.1875 & 0.0625 & 0.0625 \\ 0.0625 & 0.1875 & -0.0625 \\ 0.0625 & -0.0625 & 0.1875 \end{bmatrix}$$

This covariance matrix is a symmetric matrix.

Each diagonal value $\sum_{i,i}$ indicates the variance of the ith element of the data set.

Each off-diagonal element $\sum_{i,j}$ indicates the covariance between the ith and jth element of the data set.

PCA Algorithm (Cont.)

- A non-zero vector U_k is the eigenvector of the covariance matrix if $\sum u_k = \lambda_k u_k$
- It has the corresponding eigenvalue λ_k
- If $\lambda_1, \lambda_2, ..., \lambda_K$ are K largest and distinct eigenvalues, the matrix U=[u1 u2 … uk] represent the K dominant eigenvectors.

PCA Algorithm (Cont.)

- The eigenvectors are mutually orthogonal and span a K-dimensional subspace called the principal subspace.
- When the data are face images, these eigenvectors are often referred to as eigenfaces.

PCA Algorithm (Cont.)

- If U is the matrix of dominant eigenvectors, an N-dimensional input x can be linearly transformed into a K-dimensional vector α by:

$$\alpha = U^T (x - \mu)$$

- After applying the linear transform U^T, the set of transformed vectors $\{\alpha_1, \alpha_2, \ldots \alpha_M\}$ has scatter

$$U^T \Sigma U = \begin{bmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ldots\ldots & \\ & & & \lambda_k \end{bmatrix}$$

- PCA chooses U so as to maximize the determinant of this scatter matrix.

61

PCA Algorithm

- An original vector x can be approximately constructed from its transformed α as:

$$\tilde{x} = \sum_{k=1}^{K} \alpha_k u_k + \mu$$

- In fact, PCA enables the training data to be reconstructed in a way that minimizes the squared reconstruction error over the data set. This error is:

$$\varepsilon = \frac{1}{2} \sum_{m=1}^{M} \left\| x_m - \tilde{x}_m \right\|^2$$

62

PCA Algorithm (Cont.)

- Geometrically, PCA consists of projection onto K orthonormal axes.
- These principal axes maximize the retained variance of the data after projection.
- In practice, the covariance matrix is often singular, particularly, if M<N.
- However, the K<M principal eigenvectors can still be estimated using Singular Value Decomposition (SVD) or Simultaneous Diagonalization.

63

```
[pc, newdata, variance, t2] = princomp(A)
pc =
   0.8165         0    0.5774
   0.4082   -0.7071   -0.5774
   0.4082    0.7071   -0.5774
newdata =
  -0.8165   -0.0000   -0.1443
   0.0000   -0.0000    0.4330
   0.4082   -0.7071   -0.1443
   0.4082    0.7071   -0.1443
variance =
   0.3333
   0.3333
   0.0833
t2 =
   2.2500
   2.2500
   2.2500
   2.2500
```

64

PCA Summary

- The PCA method generates a new set of variables, called principal components.
- Each principal component is a linear combination of the original variables.
- All the principal components are orthogonal to each other so there is no redundant information.
- The principal components as a whole form an orthogonal basis for the space of the data.

65

The First Principal Component

- The first principal component is a single axis in space. When you project each observation on that axis, the resulting values form a new variable. And the variance of this variable is the maximum among all possible choices of the first axis.

66

The Second Principal Component

- The second principal component is another axis in space, perpendicular to the first. Projecting the observations on this axis generates another new variable. The variance of this variable is the maximum among all possible choices of this second axis.

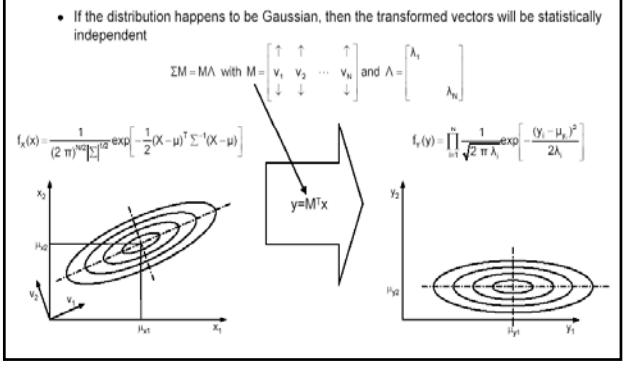
67

Principal Components

- The full set of principal components is as large as the original set of variables.
- But it is commonplace for the sum of the variances of the first few principal components to exceed 80% of the total variance of the original data. By examining plots of these few new variables, researchers often develop a deeper understanding of the driving forces that generated the original data.

68

PCA Illustration



PCA Illustration Explanation

- Given an nxn matrix that does have eigenvectors, there are n of them.
- Scale the vector by some amount before multiplying it, the same multiple of it will be obtained.
- All the eigenvectors of a matrix are perpendicular, i.e., at right angles to each other, no matter how many dimensions you have.➔ You can express the data in terms of these perpendicular eigenvectors, instead of expressing them in terms of the x and y axes.

70

Possible Use of PCA

- Dimensionality reduction
- The determination of linear combinations of variables
- Feature selection: the choice of the most useful variables.
- Visualization of multidimensional data
- Identification of underlying variables.
- Identification of groups of objects or of outliers

71