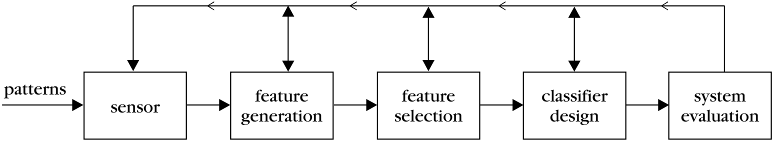
Pattern Recognition Lecture
“Feature Extraction”

Prof. Dr. Marcin Grzegorzek

Research Group for Pattern Recognition
www.pr.informatik.uni-siegen.de

Institute for Vision and Graphics
University of Siegen, Germany

Pattern Recognition Chain



We are here today!

Overview

1. Introduction
2. MPEG-7 Image Descriptors
3. Regional Features
4. Features for Shape and Size Characterisation
5. Typical Features for Speech and Audio Classification

Overview

1 Introduction

2 MPEG-7 Image Descriptors

3 Regional Features

4 Features for Shape and Size Characterisation

5 Typical Features for Speech and Audio Classification

Introduction to Image Feature Generation (1)

- The major goal can be summarised as follows:
  Given an image, or a region within an image, generate the features that will subsequently be fed to a classifier in order to classify the image in one of the possible classes.

- A digital image results from sampling of a continuous image function $I(x, y)$ to a two-dimensional array $I(m, n)$ with $m = 0, \ldots, N_x - 1$ and $n = 0, \ldots, N_y - 1$.

- The intensity of grey level image pixels $I(m, n)$ is quantised in N_g levels and N_g is known as the depth of the image. Then, a pixel $I(m, n)$ can take one of the values $0, 1, \ldots, N_g - 1$.

Introduction to Image Feature Generation (2)

- Features are generated from images, because using row image data is highly inefficient. Already for a small 64×64 image the number of pixels 4096 is too large for many classification techniques.

- The goal is to generate features that exhibit high information packing properties. Features should encode efficiently the relevant information residing in the original data.

Types of Image Features

- Colour Features
- Texture Features
- Shape Features

Introduction to Audio Feature Generation

- In contrast to images, audio signals are one dimensional.
- Statistical and spectral analysis can be applied.
- The same goal: High information packing properties

Overview

1. Introduction
2. MPEG-7 Image Descriptors
3. Regional Features
4. Features for Shape and Size Characterisation
5. Typical Features for Speech and Audio Classification

MPEG-7 Colour Descriptors

- Dominant Colour
- Scalable Colour
- Colour Layout
- Colour Structure

MPEG-7 Texture Descriptors

- Edge Histogram
- Homogeneous Texture
- Texture Browsing

MPEG-7 Shape Descriptors

- Region Shape
- Contour Shape

http://www.chiariglione.org/mpeg/

Overview

1. Introduction
2. MPEG-7 Image Descriptors
3. Regional Features
4. Features for Shape and Size Characterisation
5. Typical Features for Speech and Audio Classification

Texture Features

- Although there is no clear definition of “texture”, we describe an image by the look at it as fine or coarse, smooth or irregular, homogeneous or inhomogeneous...

- Our goal here is to generate features that somehow quantify this kind of properties of an image region.

- These features will emerge by exploiting space relations underlying the grey level distribution.

Texture - First Order Statistics Features (1)

- Let I be the random variable representing the grey levels in the region of interest. The first order histogram $P(I)$ is defined as

$$P(I) = \frac{\text{number with pixels with grey level } I}{\text{total number of pixels in the region}}$$

- The following quantities can be now defined:

$$\text{Moments:} \quad m_i = E[I^i] = \sum_{I=0}^{N_g-1} I^i P(I), \quad i = 1, 2, \ldots$$

$$\text{Central moments:} \quad \mu_i = E[(I - E[I])^i] = \sum_{I=0}^{N_g-1} (I - m_1)^i P(I)$$

Texture - First Order Statistics Features (2)

- The most frequently used central moments are μ_2, μ_3, and μ_4. $\mu_2 = \sigma^2$ is the variance, μ_3 is known as the skewness, and μ_4 as the kurtosis of the histogram.

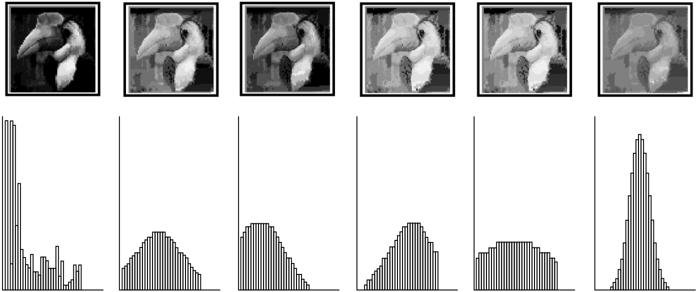
- Other quantities that result from the first order histogram are:

$$\text{Absolute Moments:} \quad \widehat{\mu}_i = E[\|I - E[I]\|^i] = \sum_{I=0}^{N_g - 1} \|I - E[I]\|^i P(I)$$

$$\text{Entropy:} \quad H = -E[\log_2 P(I)] = -\sum_{I=0}^{N_g - 1} P(I) \log_2 P(I)$$

- Entropy is a measure of histogram uniformity. The closer to the uniform distribution $P(I) = \mathrm{constant}$, the higher the H.

Texture - First Order Statistics Features (3)



Original Normal −Skewed +Skewed Platykurtic Leptokurtic

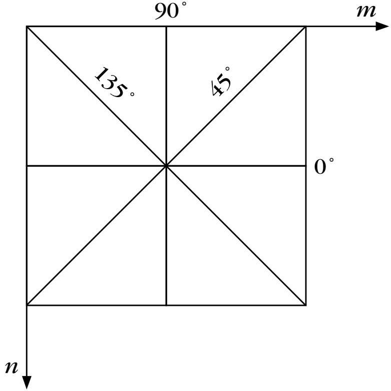
$\mu_3 \rightarrow$	587	0	-169	169	0	0
$\mu_4 \rightarrow$	16609	7365	7450	7450	9774	1007
$H \rightarrow$	4.61	4.89	4.81	4.81	4.96	4.12

Texture - Second Order Statistics Features (1)

- The first order statistics features provide information related to the distribution of grey levels, however, they don't provide any information about the relative positions of the various grey levels in the image.
- This type of information can be extracted from the second order histograms, where the pixels are considered in pairs.
- Two more parameters are used in this case, namely the relative distance among the pixels and their relative orientation.

Texture - Second Order Statistics Features (2)

- Let d be the relative distance measured in pixel numbers. The orientation ϕ is quantised in four directions: horizontal, diagonal, vertical, and anti-diagonal (0°, 45°, 90°, 135°):



Texture - Second Order Statistics Features (3)

- For each combination of d and ϕ a two-dimensional histogram is defined:

$$0^\circ : \; P(I(m,n) = I_1, I(m \pm d, n) = I_2)$$

$$= \frac{\text{no. of pixel pairs at distance } d \text{ with values } I_1, I_2}{\text{total number of possible pairs}}$$

- In a similar way

$$45^\circ : \; P(I(m,n) = I_1, I(m \pm d, n \mp d) = I_2)$$

$$90^\circ : \; P(I(m,n) = I_1, I(m, n \mp d) = I_2)$$

$$135^\circ : \; P(I(m,n) = I_1, I(m \pm d, n \pm d) = I_2)$$

Texture - Second Order Statistics Features (4)

- For each of these histograms an array is defined, known as the co-occurrence or spatial dependence matrix.

$$\mathbf{A} = \frac{1}{R}\left[\begin{array}{cccc} \eta(0,0) & \eta(0,1) & \eta(0,2) & \eta(0,3) \\ \eta(1,0) & \eta(1,1) & \eta(1,2) & \eta(1,3) \\ \eta(2,0) & \eta(2,1) & \eta(2,2) & \eta(2,3) \\ \eta(3,0) & \eta(3,1) & \eta(3,2) & \eta(3,3) \end{array}\right]$$

- $\eta(I_1, I_2)$ is the number of pixel pairs, at a relative position (d, ϕ), which have grey level values I_1 and I_2 respectively. R is the total number of possible pixel pairs. Thus,

$$\frac{1}{R}\eta(I_1, I_2) = P(I_1, I_2)$$

Local Linear Transforms (1)

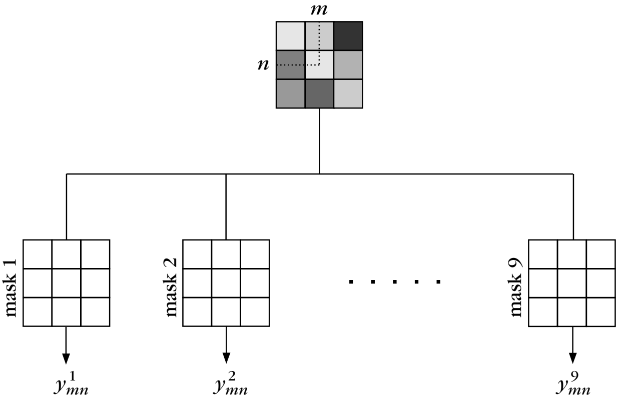
- Let us consider a neighbourhood of size $N \times N$ cantered at pixel location (m, n). Let $\mathbf{x}_{mn}$ be the vector with elements being the N^2 points within the area, arranged in a row-by-row mode.

- A local linear transform or local feature extractor is defined as

$$\mathbf{y}_{mn} = \mathbf{A}^{\mathrm{T}}\mathbf{x}_{mn} \equiv \begin{bmatrix} \mathbf{a}_1{}^{\mathrm{T}} \\ \mathbf{a}_2{}^{\mathrm{T}} \\ \vdots \\ \mathbf{a}_{N^2}{}^{\mathrm{T}} \end{bmatrix} \mathbf{x}_{mn}$$

Local Linear Transforms (2)

- The problem of local linear transform can be interpreted as a series of N^2 filtering operations



Overview

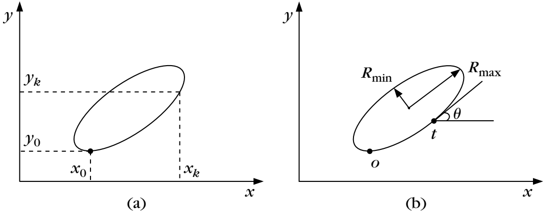
1. Introduction
2. MPEG-7 Image Descriptors
3. Regional Features
4. Features for Shape and Size Characterisation
5. Typical Features for Speech and Audio Classification

Introduction

- While texture features describe whole images or image regions, shape and size features are related to objects.
- Some objects have exactly the same shape and can be distinguished by the texture, other have exactly the same texture, but different shapes.
- Extraction methods for shape features depend on segmentation algorithms and, therefore, their performance is limited for images with heterogeneous backgrounds.

Fourier Features (1)

- Let (x_k, y_k) with $k = 0, \dots, N-1$ be the coordinates on the boundary of an object.



- For each pair (x_k, y_k) we define the complex variable $u_k = x_k + jy_k$ and obtain the DFT f_l

$$f_l = \sum_{k=0}^{N-1} u_k \exp\left(-j\frac{2\pi}{N} lk\right), \quad l = 0, 1, \dots, N-1$$

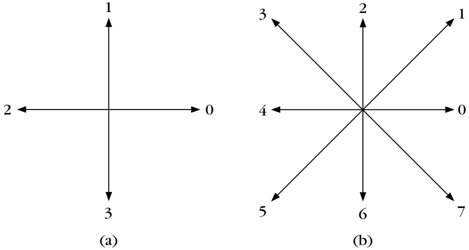
Fourier Features (2)

- The coefficients f_l are known as Fourier descriptors of the boundary.
- Once f_l are available, the u_k can be recovered and the boundary can be reconstructed.
- However, the goal of pattern recognition is not to reconstruct the boundary. Thus, a smaller number of coefficients is usually used.

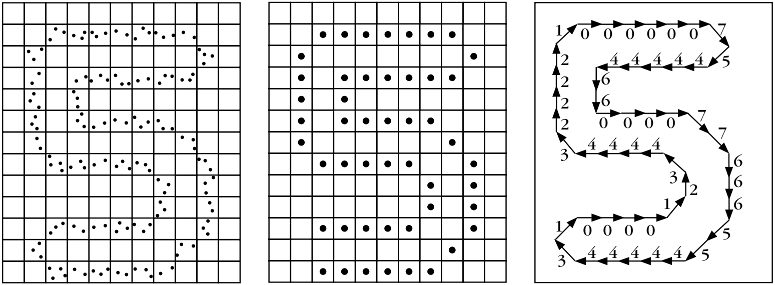
Chain Codes - Introduction

- Chain coding is the most widely used technique for shape description.

- Directions for a four-directional (a) and an eight-directional (b) chain code are defined as follows:



Chain Codes - Example



(a) (b) (c)

(a) - original sample image

(b) - its resampled version

(c) - the resulting chain code.

CBIR Example using Shape Features

Content-Based Image Retrieval using Shape Features

Online YouTube Video

Overview

1 Introduction

2 MPEG-7 Image Descriptors

3 Regional Features

4 Features for Shape and Size Characterisation

5 Typical Features for Speech and Audio Classification

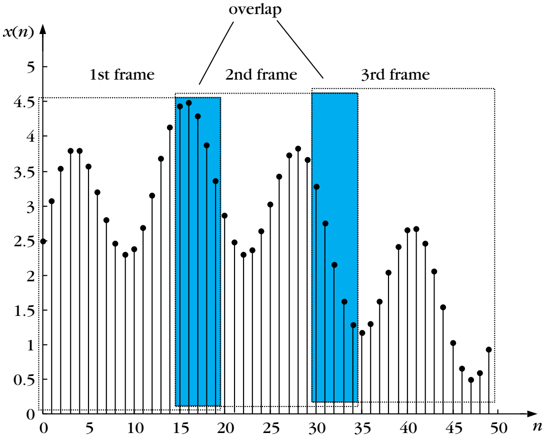
Applications

- Speech Recognition Systems (e. g. BERTI 09131 610017)
- Audiovisual Data Segmentation and Indexing
- Content-Based Retrieval from Music Databases (e. g., Querying by Humming)
- Automatic Music Genre Classification
- ...

Short Time Processing of Signals (1)

- The statistical properties of the speech and audio signals vary with time (nonstationary signals).
- In order to use tools for stationary signals (e. g., Fourier transform), the signal is divided to a series of successive frames.
- Each frame consists of a finite number, N, of samples.
- During the time interval of a frame, the signal is assumed to be “reasonably stationary” (quasistationary, see Figure on the next slide).

Short Time Processing of Signals (2)



Three successive frames, each of length $N = 20$ samples. The overlap between successive frames is 5 samples.

Short Time Processing of Signals (3)

- Choosing the length, N, is a problem-dependent task.

- On the one hand, N has to be high enough to include useful part of information. On the other hand, it has to be small for the stationary assumption.

- For speech signals sampled at a frequency of $f_s = 100$ KHz, reasonable frame sizes range from 100 to 200 samples, corresponding to 10-20 msecs time duration.

- For music signals sampled at 44.1 KHz, reasonable frame sizes range from 2048 to 4096 samples, corresponding to 45-95 msecs.

Short Time Processing of Signals (3)

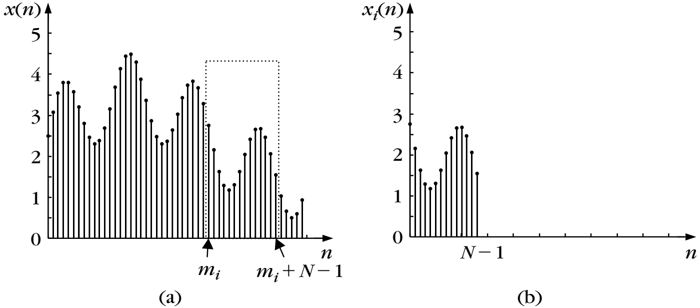
- Dividing the signal in a sequence of successive frames is equivalent to multiplying the signal segment by a window sequence, $w(n)$, of a finite duration N

$$w(n) = \begin{cases} 1 & 0 \leq n \leq N-1 \\ 0 & \text{elsewhere} \end{cases}$$

- For different frames, the window is shifted to different points m_i in the time axis. Hence, if $x(n)$ denotes the signal sequence, the samples of the frame no. i can be written as

$$x_i(n) = x(n + m_i)w(n)$$

Short Time Processing of Signals (4)



A signal segment (a) and the resulting frame (b) after the application of a rectangular window sequence of duration equal to 14 samples and shifted at m_i.

Short Time Processing of Signals (5)

- Multiplying a sequence by a window in the time domain smooths out its Fourier transform by convolving it with the Fourier transform of the window sequence.

- Some of the effects of this smoothing action can be minimised by using the so called Hamming window

$$w(n) = \begin{cases} 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right) & 0 \leq n \leq N-1 \\ 0 & \text{elsewhere} \end{cases}$$

Short Time Processing of Signals (6)

- We divide a speech signal into a sequence of F frames, each of length N.

- Then, for each frame we compute the DFT as

$$X_i(m) = \sum_{n=0}^{N-1} x_i(n) \exp\left(-j\frac{2\pi}{N}mn\right), \quad m = 0, \ldots, N-1$$

Short Time Processing of Signals (7)

- Selecting $I \leq N$ DFT coefficients from each frame, we construct a sequence of feature vectors

$$\mathbf{x}_i = \begin{bmatrix} X_i(0) \\ \vdots \\ X_i(I) \end{bmatrix}, \quad i = 1, 2, \ldots, F$$

- Thus, the pattern of interest (i. e., the speech segment) is not represented by a single feature vector but by a sequence of feature vectors

$$\mathbf{x} \to (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_i, \ldots, \mathbf{x}_F)$$

Short Time Processing of Signals (8)

- Another very important quantity defined for quasistationary processes is the short-time autocorrelation

$$r_i(k) = \frac{1}{N} \sum_{n=0}^{N-1-|k|} x_i(n)x(n+|k|)$$

- The limits in the sum indicate that outside the interval $[0, N-1-|k|]$ the product $x_i(n)x_i(n+|k|)$ is zero.

Cepstrum (1)

- Let $x(0), x(1), \ldots, x(N-1)$ be the samples from the current frame. The FT of this sequence is defined as the periodic complex function

$$X(\omega) = \sum_{n=0}^{N-1} x(n) \exp(-j\omega Tn)$$

with period in the frequency domain $\frac{2\pi}{T}$, where T is the sampling period.

- The coefficients of the DFT are the samples of FT taken at the frequency points $0, \frac{2\pi}{NT}, \ldots, \frac{2\pi}{NT}(N-1)$

$$X(m) = \sum_{n=0}^{N-1} x(n) \exp\left(-j\frac{2\pi}{N}mn\right), \quad m = 0, \ldots, N-1$$

Cepstrum (2)

- Assuming[1] $T = 1$ the inverse FT is defined as

$$x(n) = \frac{1}{2\pi} \int_{-\pi}^{\pi} X(\omega) \exp(j\omega n) d\omega, \quad n = 0, \ldots, N-1$$

- That is, the resulting samples are equal to the samples of the original sequence and identical to what is obtained by the inverse DFT

$$x(n) = \frac{1}{N} \sum_{m=0}^{N-1} X(m) \exp\left(j \frac{2\pi}{N} mn\right), \quad n = 0, \ldots, N-1$$

[1] Without loss of generality.

Cepstrum (3)

- The cepstrum, $c(n)$, of a sequence, $x(n)$, is the sequence resulting from the inverse FT of the logarithm of the magnitude of its FT. That is

$$c(n) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \log_{10} |X(\omega)| \exp(j\omega n) d\omega$$

Cepstrum - Summary

The computational steps to obtain the cepstral coefficients of a frame $x_i(n), n = 0, \ldots, N-1$ are the following:

- Extend the length of the frame by appending $M - N$ zeros at the end of the frame.
- Obtain the DFT of length M of the extended frame.
- Compute the logarithm of the magnitude of the DFT coefficients.
- Compute the inverse DFT of length M.

Spectral Features - Introduction

- Let $x_i(n), n = 0, \ldots, N-1$ be the samples of the frame no. i and $X_i(m), m = 0, \ldots, N-1$ the corresponding DFT coefficients.

- The following features are common in speech/audio recognition:
  - Spectral Centroid
  - Spectral Roll-Off
  - Spectral Flux
  - Fundamental Frequency

Spectral Features - Spectral Centroids

- Definition

$$C(i) = \frac{\sum_{m=0}^{N-1} m|X_i(m)|}{\sum_{m=0}^{N-1} |X_i(m)|}$$

- The centroid is a measure of the spectral shape. High values of the centroid correspond to “brighter” acoustic structures with more energy in the high frequencies.

Spectral Features - Spectral Roll-Off

- The spectral roll-off is the frequency sample $m_c^R(i)$ below which the $c\%$ of the magnitude distribution of the DFT coefficients is concentrated.

$$\sum_{m=0}^{m_c^R(i)} |X_i(m)| = \frac{c}{100} \sum_{m=0}^{N-1} |X_i(m)|$$

- This measure indicates where the most of the spectral energy is concentrated.

Spectral Features - Spectral Flux

- Definition

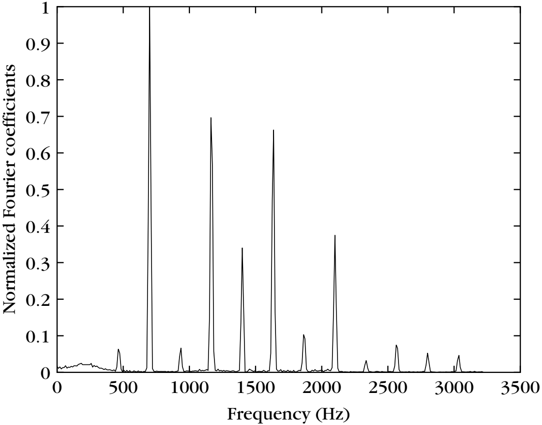
$$F(i) = \sum_{m=0}^{N-1} (N_i(m) - N_{i-1}(m))^2$$

- Here, $N_i(m)$ is the normalised (by its maximum value) magnitude of the respective DFT coefficient of the frame no. i and is measure of the local spectral change between successive frames.

Spectral Features - Fundamental Frequency (1)

- Audio signals produced by musical instruments and voiced speech segments are harmonic and can be characterised by its fundamental frequency.
- For men's voiced speech signals it lies in the range of 80 to 200 Hz, for women's in the range of 150 to 350 Hz.
- For musical instruments the fundamental frequency may vary a lot and, in some cases, may not be present in the frequency spectrum (see Figure on the next slide).

Spectral Features - Fundamental Frequency (2)



Normalised DFT coefficients of a clarinet sound, whose fundamental frequency is absent.

Time Domain Features

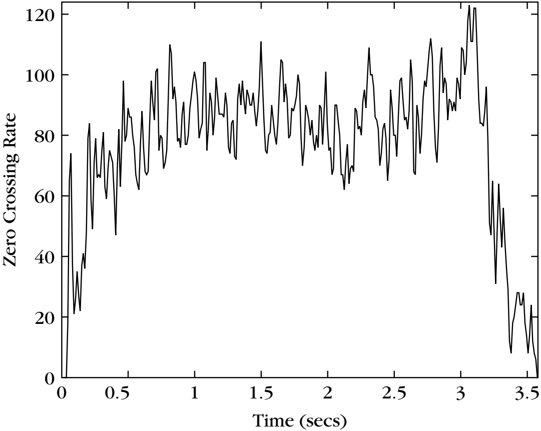
- The zero-crossing rate measures the noisiness of the signal and is defined as

$$Z(i) = \frac{1}{2N} \sum_{n=0}^{N-1} |\mathrm{sgn}[x_i(n)] - \mathrm{sgn}[x_i(n-1)]|$$

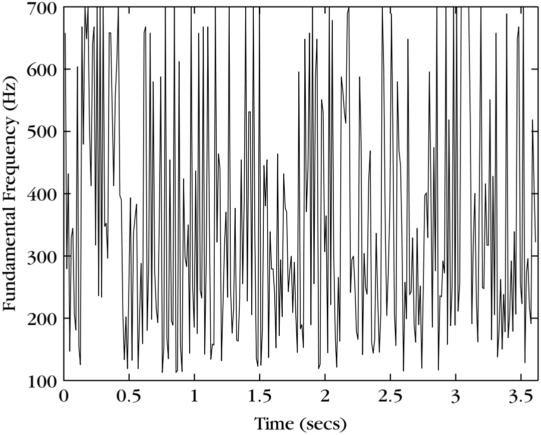
- Energy is used to discriminate voiced from unvoiced speech signals. Its definition is

$$E(i) = \frac{1}{N} \sum_{n=0}^{N-1} |x_i(n)|^2$$

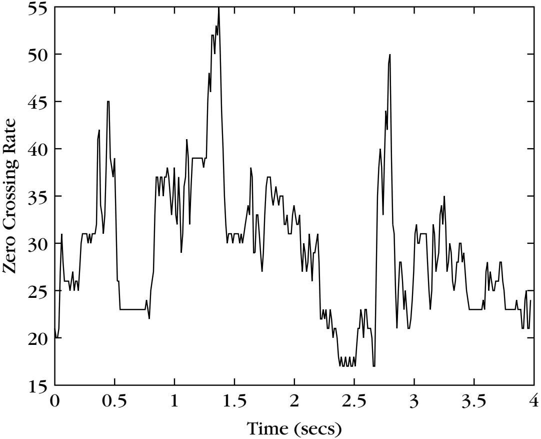
Noisy Example - Zero-Crossing Rate



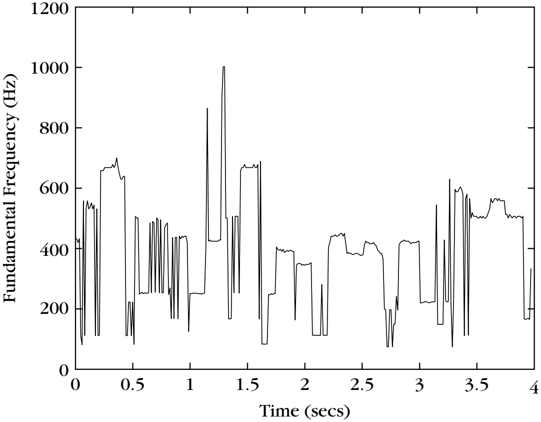
Noisy Example - Fundamental Frequency



Piano Example - Zero-Crossing Rate



Piano Example - Fundamental Frequency


Fundamental Frequency (Hz)
Time (secs)