Feature Extraction in Computational Intelligence

Evangelia Micheli-Tzanakou, PhD

Department of Biomedical Engineering,
RUTGERS UNIV.
cil.rutgers.edu

IJCNN 2004 Budapest
International Joint Conference on Neural Networks

Learning from Data

- Design of experiments, data recording and analysis
- Now what do you do?

Size of the Data Set Matters

- If you do not know what to do:
  - Try simple tools first
  - Then more complex ones
  - Validate them properly on separate test sets

Statistics or DATA MINING?

- Statistics deals with small data sets- data mining deals with large data sets
- Statistics addresses focused questions- Data Mining unfocused
- Statistics-uses probabilistic inference based on population models
- Data Mining-????

Challenges

- Huge data sets-memory problems
- How much data we really need?
- Different types of data-how do we handle them?
- What if the data are correlated
- What if we have complex data structures?

Then…

- Learn more about Computational Intelligence
- Learn more about Feature Extraction
- Best of all: Know your data!

Pattern recognition

A **Pattern** is a description of an object

The object belongs to a **Class** or a **Set** where each element shares common properties. For example:

1. The alphabet is a set of objects (letters) with the property that all appear in a text.
2. Humans form a set of objects (men, women) with common properties (2 feet, 2 arms, well developed cerebral cortex)

Pattern Recognition (cont.)

- In Pattern Recognition we extract "relevant" information about an object via experiments

  and

  Use these measurements (=features) to classify an object.

Pattern Recognition (cont.)

- Arrange the measurements of the object in a pattern vector

$$\mathbf{x}=[x_1\ x_2\ \ldots.\ x_n]$$

- "Extract" characteristic features or attributes from the input data
- Operate on the pattern vector to obtain a feature vector

$$\mathbf{F}=[y_1\ y_2 \ldots.\ y_m],\quad m<n$$

y_i is a feature.

Feature - Any local attribute or property of a specific configuration of some object or image that is critical in distinguishing that object or image from others.

Feature detector - A perceptual mechanism that detects single distinctive features in complex displays. Generally thought to be the receptive fields of neurons, such as simple and complex cells, that respond to orientations, size, spatial frequency, etc.

Visual Processes/Mechanisms

Preattentive - A parallel, effortless process which signals where texture gradients (feature differences) are located, and directs focal attention.

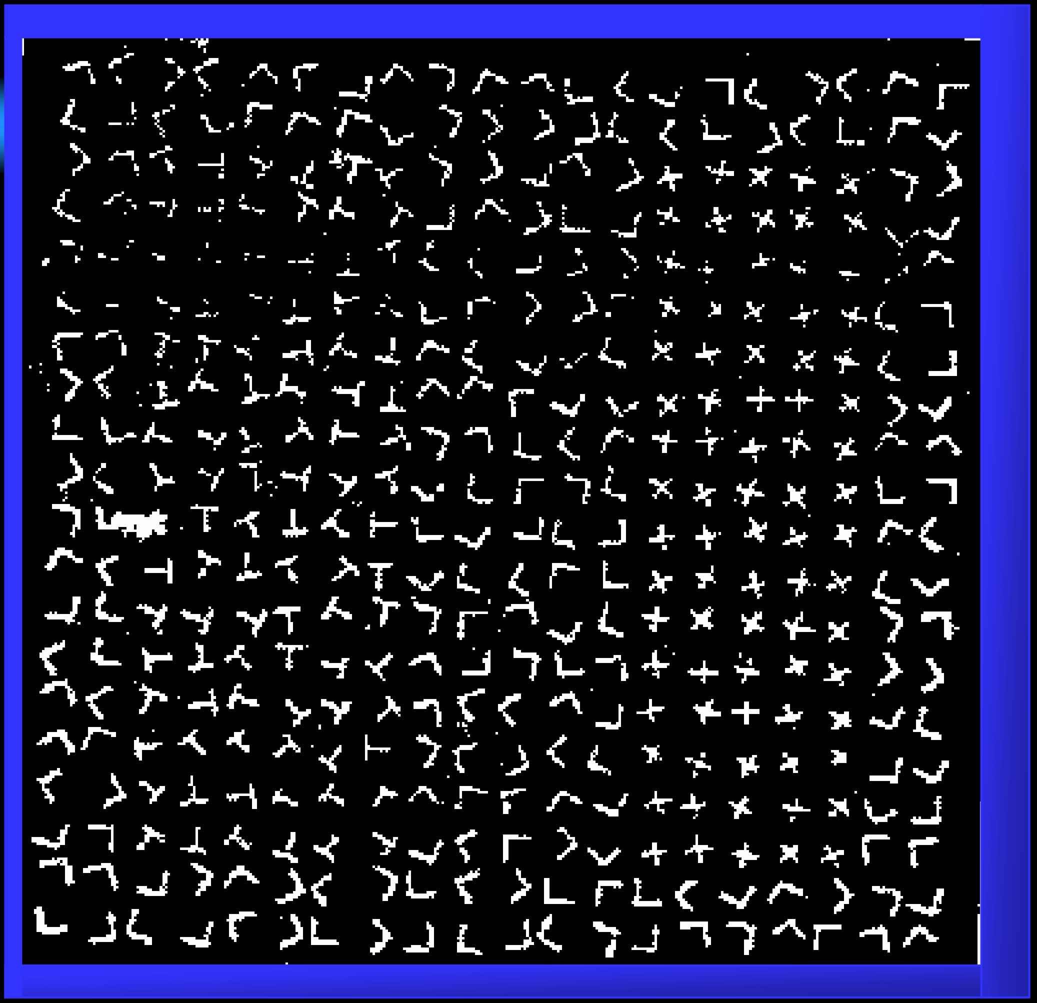
Focal Attention - A searchlight which scrutinizes each element of the texture in a serial fashion, and signals what is in the texture by synthesizing features in the same spatial location into whole objects.

e.g. Texton theory

Proposes that the visual system applies some local spatial filtering which is followed by some non-linearity such as threshold taking, and then a second spatial filtering such as averaging which separates the areas of different luminance distributions obtained by the threshold taking.

Texton theory

- Textons: specific texture spatial properties to which the pre-attentive processes are highly sensitive. They include elongated blobs of specific orientations, terminators, color, motion, spatial frequency, and line crossings
- Pre-attentive vision selects the areas where texton activity is highest due to greater number and density



Models

- Several models exist within the Texton theory

- Each model starts with some local conspicuous feature (whether they are textons, size, orientation, or spatial frequency) being extracted from the texture.

- Next the outputs of these feature detectors are pooled in some way over the different texture regions

Models…..

- These stages are somewhat analogous to the operations of simple cells, complex cells, and hyper-complex cells, respectively, as found in the visual system.
- Finally, these pooled outputs are compared in order to find differences among them which in turn will segregate the textures.

Feature selection

- A feature vector can be thought of as a vector in an n-dim vector space, where “components” are the projections on the feature axes and correspond to the magnitude of the features
- Features and feature vectors are samples from a probability distribution whose statistical properties can be estimated from a random sample of the population

Feature selection…..

- Select from the initial set of features, that subset which best discriminates between two or more previously defined groups of objects

- The last step is called *Feature Selection*

Feature selection

- Intraset features
  - Those which characterize properties common to all members of a given class
  - Intraset features that contain no information that permits discrimination may be ignored

Feature selection

- Interset features have values that permit differentiation between the classes under study

- Features that discriminate best between groups are selected with statistical tests

→

- This results in a small subset of "information rich" features that are then used to design a decision (=classification) rule

Feature selection

- Feature selection reduces the dimensionality of the feature space
- Feature selection discards information poor features

Classification

- View the recognition problem as that of generating "decision boundaries" separating *m classes* on the basis of the observed vector

Important Characteristics of Features

- Discrimination
  - How good are the features
- Reliability
  - How reliable is the decision rule
- Independence
  - Features should be uncorrelated with each other
- Small numbers
  - Complexity in recognition increases with the number of features used

Important Points

- Normalization
  - The usual concept of distance may not be useful
  - One method of "norming" the space is
    - Calculate the variances of the features:

      If σ_k = variance of the k^{th} feature of all sample points (from all classes) then

      $$ {}^{N}x_k / \sigma_k $$

      are the normalized values

Dimensionality of the Feature Space

- ***Questions***

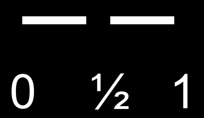
  - Why not use a large number of features in designing a decision function?

  - Doesn't the accuracy increase as we add more and more features?

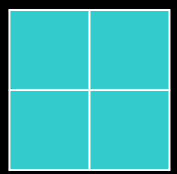
Dimensionality of the Feature Space

- ***Answer***
  - NO, because the dimensionality of the vector space increases and the number of sample points necessary to give a meaningful estimate of the decision rule parameters increases dramatically

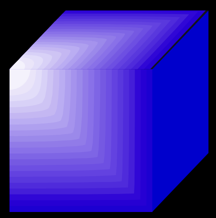
Dimensionality of the Feature Space



2 samples (1-D)



4 samples (2-D)



8 samples (3-D)

Dimensionality of the Feature Space

- ……….For an n-dimensional cube, we would need 2^n evenly distributed points for the same density, and even then, the feature space would be sparsely populated :

Dimensionality Curse

Dimensionality of the Feature Space

Rule of thumb:

If M=number of sample feature vector per class and

If n=number of features

then

$$M/n > 5$$

Analysis Methods

- average power
- Fourier analysis
- wavelets
- fractal dimension
- entropy
- moments
- Hjorth parameters
- modular neural network

Analysis Methods

- average power
- Fourier analysis
- wavelets
- fractal dimension
- entropy
- moments
- Hjorth parameters
- modular neural network

Wavelets

- If a signal contains frequency components emerging and vanishing in certain time intervals, then a time and a frequency localization is required
- Historically, this is done with the *Short Time Fourier Transform (STFT) or Gabor Transform*

Wavelets

- There exists a Heisenberg's Uncertainty Principle between time and frequency

$$\Delta t \Delta f \geq \frac{1}{4\pi}$$

In order to overcome the resolution limitation of the STFT a decomposition of square integrable signals has been developed

$$h_{a,b}(t) = |a|^{\frac{1}{2}} h(t-b) a, b \in \_, a \neq 0$$

Wavelets

These families of functions $h_{a,b}$ are generated from a single function *h(t)* by the operation of dilations and translations

$$CWT_x(b,a) = \langle x(t), |a|^{-1/2} h^*((t-b)/a) \rangle = |a|^{-1/2} \int x(t) h^*((t-b)/a)\,dt$$

Where *x(t)* is a continuous function, * represents the complex conjugation and < > represents the inner product.

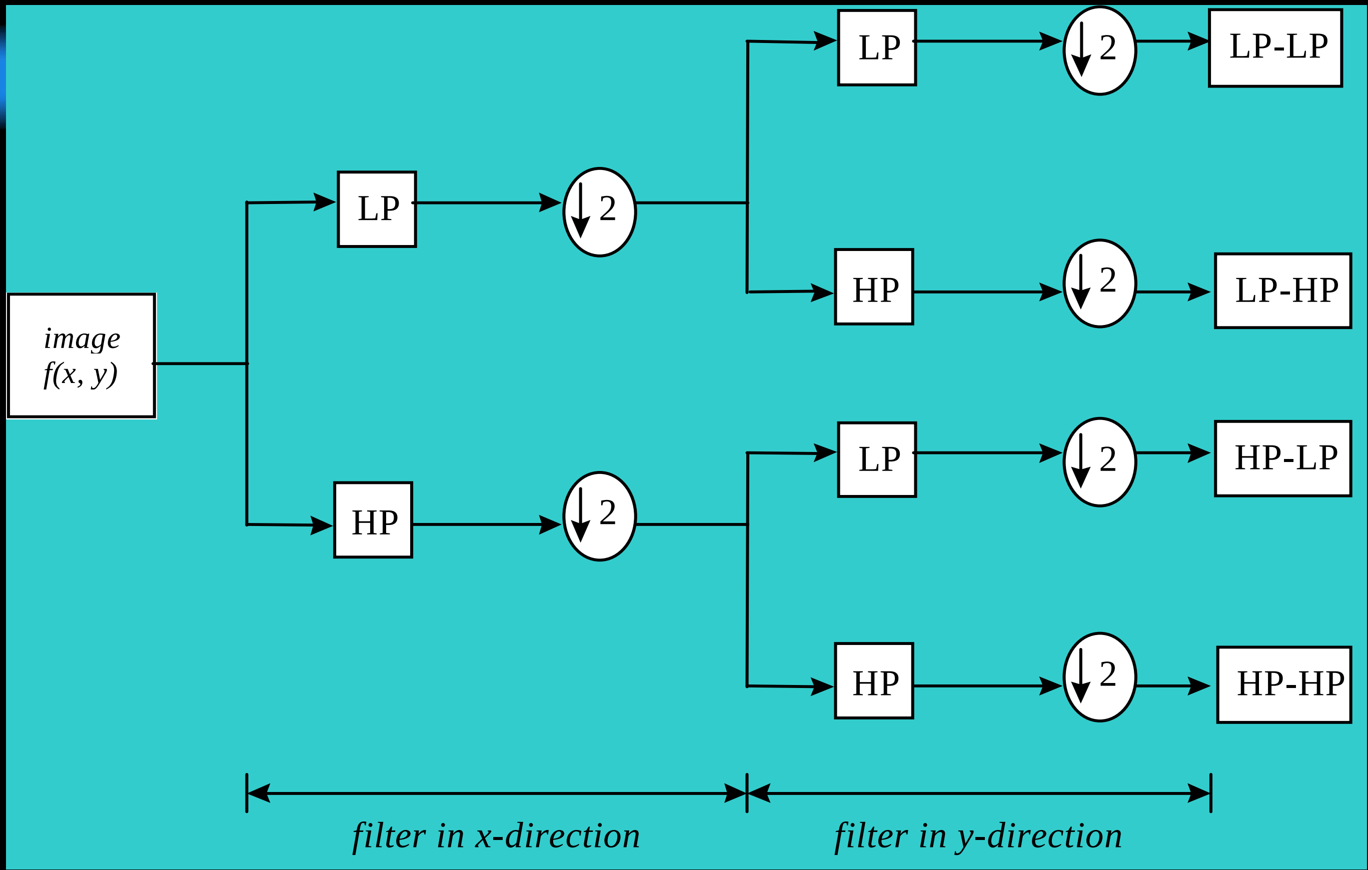
Wavelets

- The last equation is interpreted as a multi-resolution decomposition of the signal into a set of channels having the same bandwidth in a logarithmic scale
- For the STFT the phase space is uniformly sampled
- In the wavelet transform the sampling in frequency is logarithmic
- The latter enables one to analyze higher frequencies in shorter windows and lower frequencies in longer windows in time

Wavelets

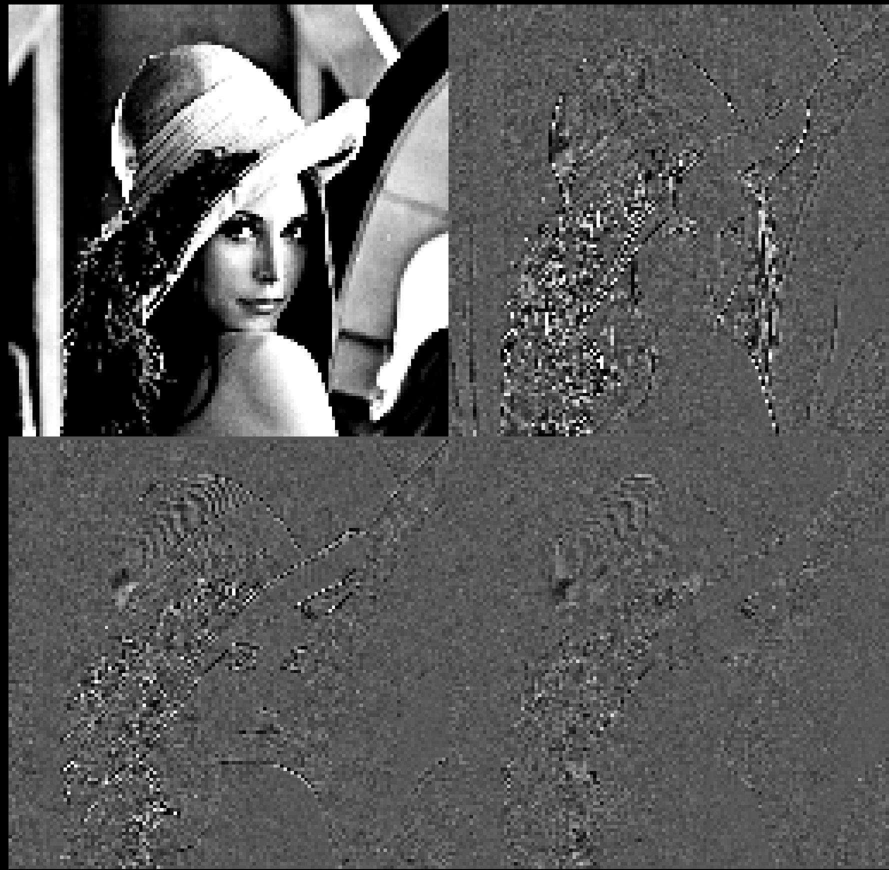
- Taking the wavelet transform of an image involves convolving a pair of filters, one *high pass* and one *low pass,* with the image

Wavelets



Wavelet transform algorithm - sub-band decomposition of one octave.
HP = high-pass, LP = low-pass, ↓2 represents decimation by 2

Wavelets



(a) Lena (b) Octave 1

The wavelet transform of Lena.bmp. Note that (b) has been enhanced to accentuate the detail coefficients (high pass components).

Wavelets

Discrete Wavelet Series

$$h_{ik}(t) = 2^{\frac{i}{2}} h\left(2^{-i}t - k\right).$$

$$x(t) = \sum_{i \in Z} \sum_{k \in Z} C_{i,k} \alpha_{i,k}(t)$$

$$C_{i,k} = \int x(t) h_{i,k}^{*}(t) dt$$

Wavelets

Discrete Wavelet Transform (DWT)

$$x[n] = \sum_{i=I}^{\infty} \sum_{k \in Z} a_{i,k} g_i\left[n - 2^i k\right] + \sum_{k \in Z} b_{I,k} h_I\left[n - 2^I k\right]$$

$$DWT\left\{x[n]; 2^I, k2^I\right\} = a_{i,k} = \sum x[n] g_i^*\left[n - 2^i k\right]$$

$$b_{I,k} = \sum_n x[n] h_I^*\left[n - 2^I k\right]$$

where the $g_i[n - 2^i k]$'s are the discrete wavelets and the $h_I[n - 2^I k]$ are the scaling sequences.

Wavelets, an example….

Wavelet Transform

What it is:

- A signal processing technique
- Projects a signal into multiple frequency bands
- Overcomes the *uncertainty principle* associated with the Short Term Fourier Transform (STFT), which uses "fixed" analysis windows.

The Uncertainty Principle:

- High frequency resolution → poor time localization
- High time localization → poor frequency resolution

How Does the Wavelet Transform Overcome This ?

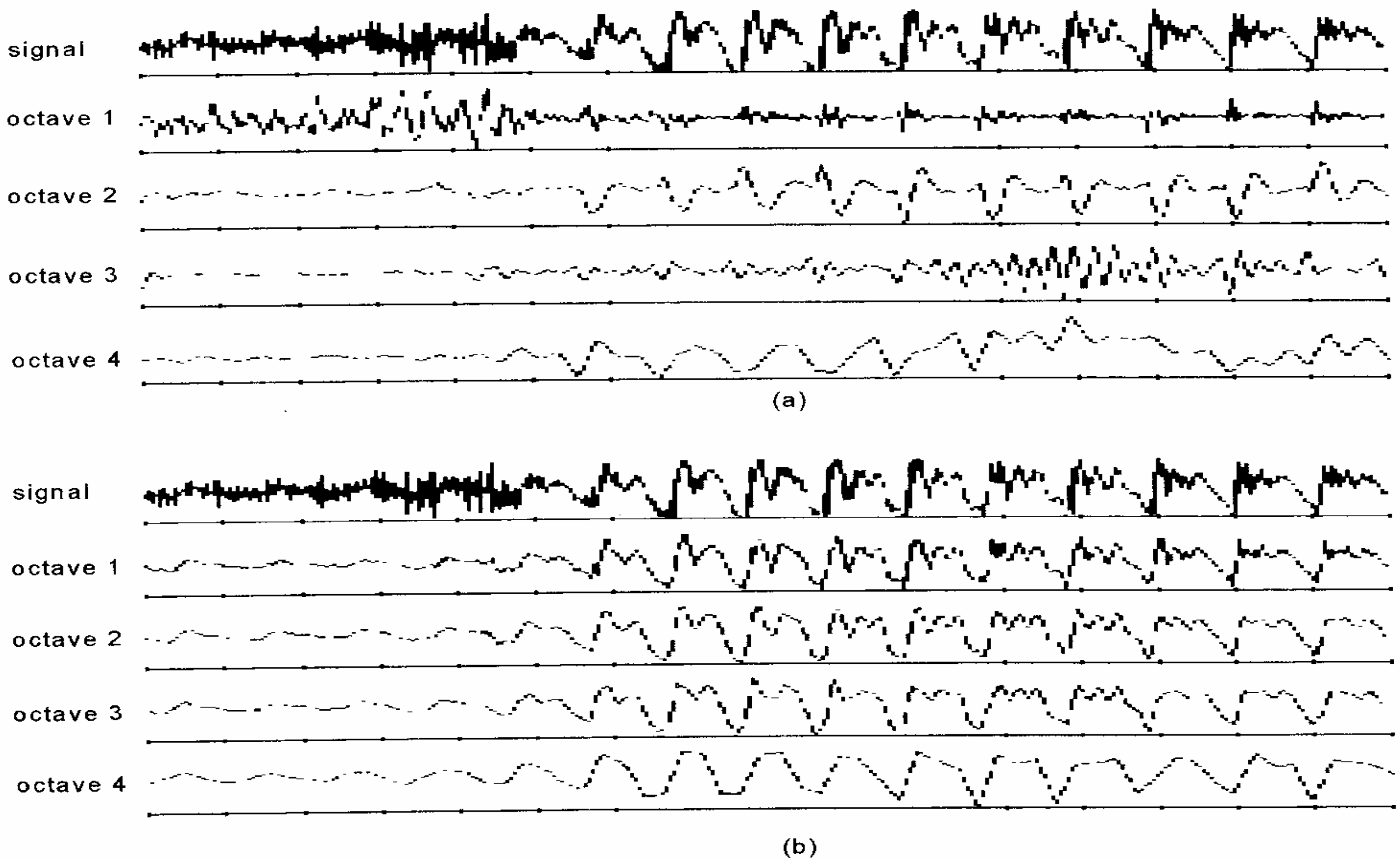
- The time analysis window compresses and dilates to produce a fine to coarse resolution of the signal

Motivation:

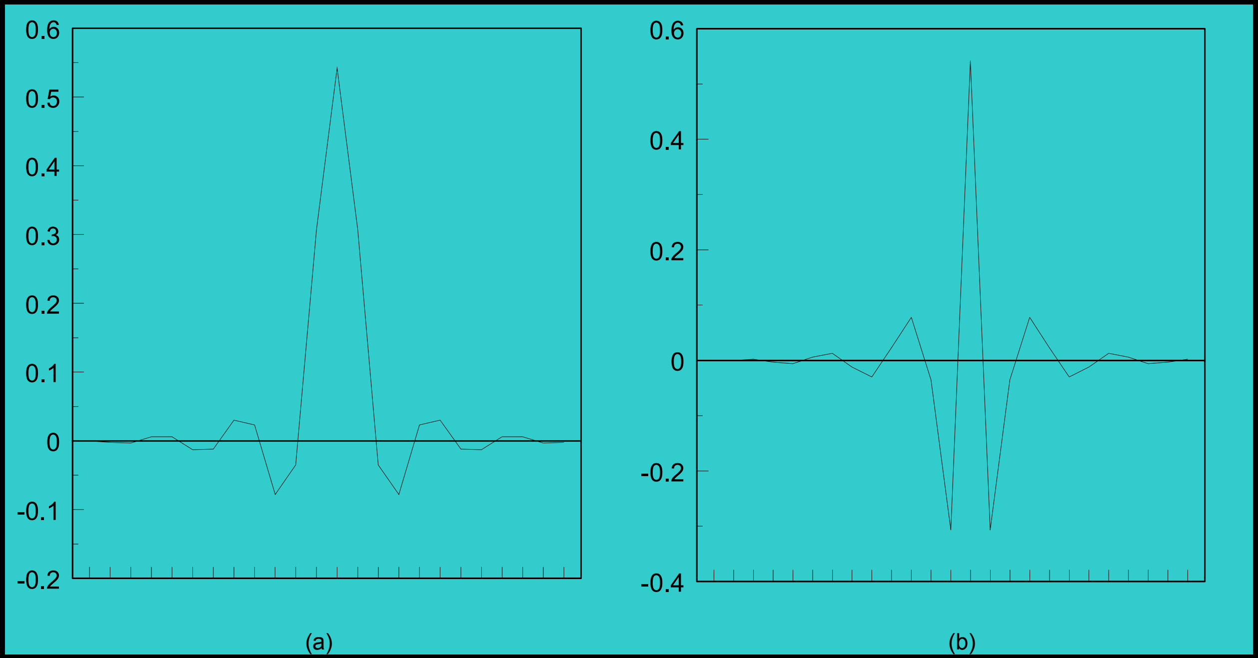
- Wavelet Transform better suited for extracting features in nonstationary signals such as speech
- **Vowels**: Long time analysis window to extract low frequency features
- **Consonants**: Short time analysis window to extract high frequency features.

Speaker Identification

Generating Wavelets from a Waveform: (a)signal detail (b) discrete approx.

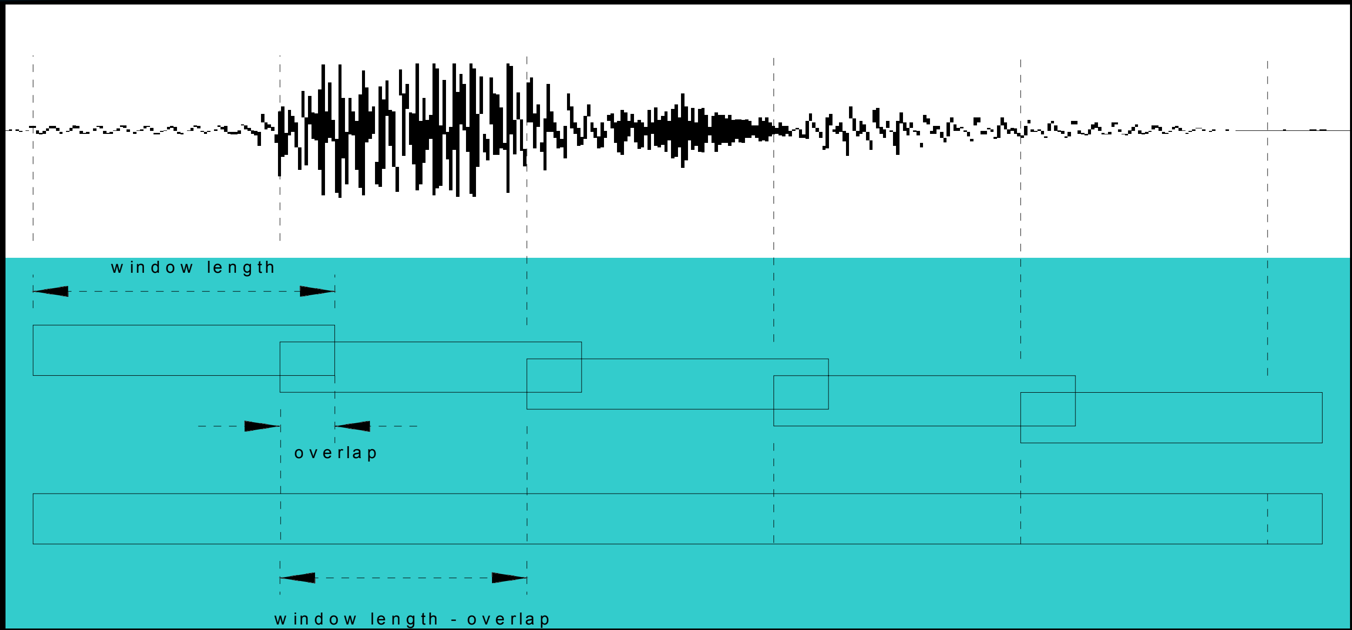


Wavelets



High Pass and Low Pass filter coefficients

Speaker Identification



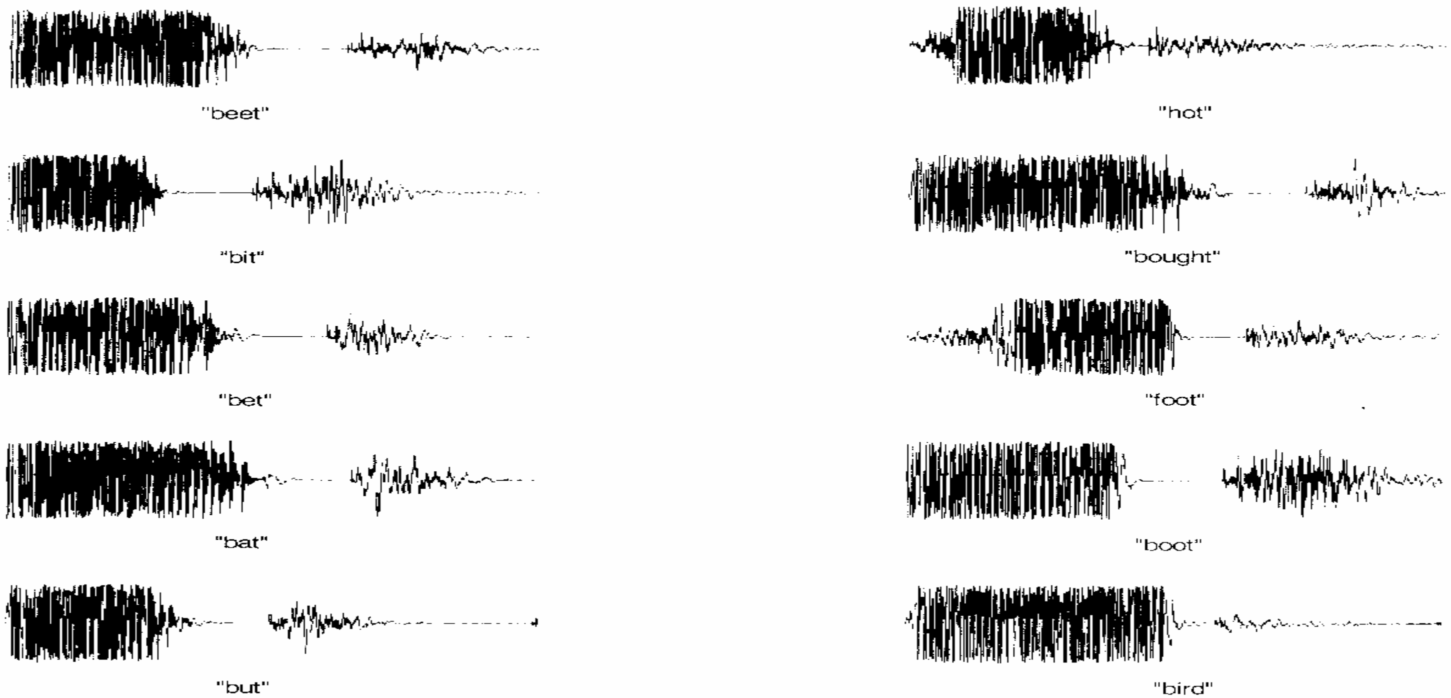
Processing with overlapping windows

Evangelia Micheli-Tzanakou, PhD

Speaker Identification

Test utterances articulated three times each by six subjects and phonetic representation.

utterance	vowel	utterance	vowel
beet	IY	hot	AA
bit	IH	bought	AO
bet	EH	foot	UH
bat	AE	boot	UW
but	AH	bird	ER



Example of sampled utterances.

Summary of Results

Table a. Speaker recognition without noise.

	FP	MZ	KL	BL	YA	DZ
FP	👍	–	–	–	–	–
MZ	☒	👎	–	–	–	–
KL	–	–	👍	–	–	–
BL	–	–	–	👍	–	–
YA	–	–	–	–	👍	–
DZ	–	–	–	–	–	👍

Table b. Speaker recognition against -20 dB noise.

	FP	MZ	KL	BL	YA	DZ
FP	👍	–	–	–	–	–
MZ	–	👎	–	–	–	☒
KL	–	☒	👎	–	–	–
BL	–	–	–	👍	–	–
YA	–	–	–	–	👎	☒
DZ	–	–	–	–	–	👍

Table c. Speaker recognition against competing speaker (cocktail party effect)

	FP	MZ	KL	BL	YA	DZ
FP	👍	–	–	–	–	–
MZ	☒	👎	–	–	–	–
KL	–	–	👍	–	–	–
BL	–	–	–	👍	–	–
YA	–	–	–	–	👍	–
DZ	–	–	–	–	–	👍

Speaker Identification

Table c. Interspeaker speech recognition for three words of two utterances each using **Alopex Template Matching**.

	beet1	boot1	bought1
beet2	👍	–	–
beet3	👍	–	–
boot2	–	👍	–
boot3	–	👎	☒
bought2	–	–	👍
bought3	–	–	👍

Table d. Interspeaker speech recognition for three words of two utterances each using an **Artificial Neural Network with ALOPEX**.

	beet1	boot1	bought1
beet2	👎**.5**	.5	.5
beet3	👍**1.0**	0.0	0.0
boot2	0.0	👍**1.0**	0.0
boot3	.19	👍**1.0**	0.0
bought2	0.0	.01	👍**1.0**
bought3	0.0	0.0	👍**1.0**

Table e. Interspeaker speech recognition for three words of two utterances each using an **Artificial Neural Network with Backpropogation**.

	beet1	boot1	bought1
beet2	👎**.13**	☒.89	.01
beet3	👍**.89**	.09	.04
boot2	.63	👎**.23**	.30
boot3	☒.89	👎**.09**	.04
bought2	.04	.01	👍**.98**
bought3	.04	.01	👍**.98**

Speaker Identification

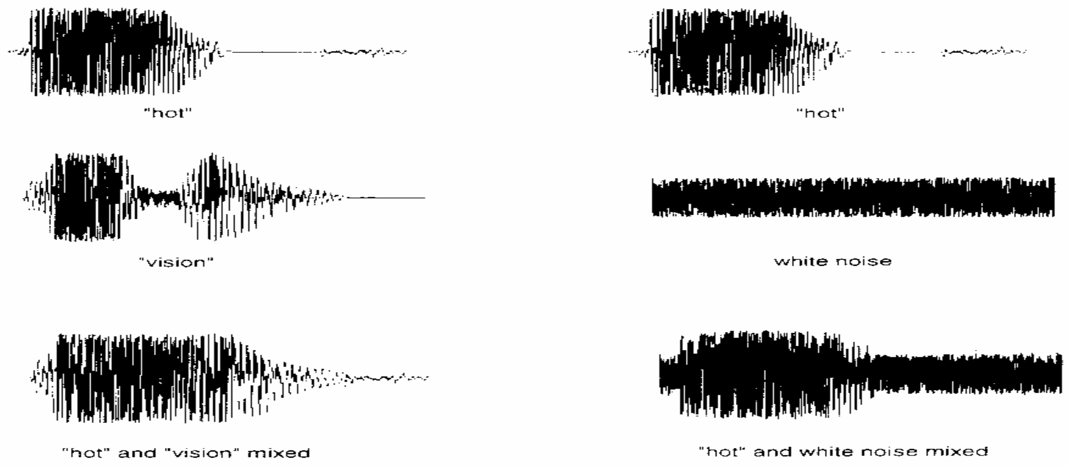
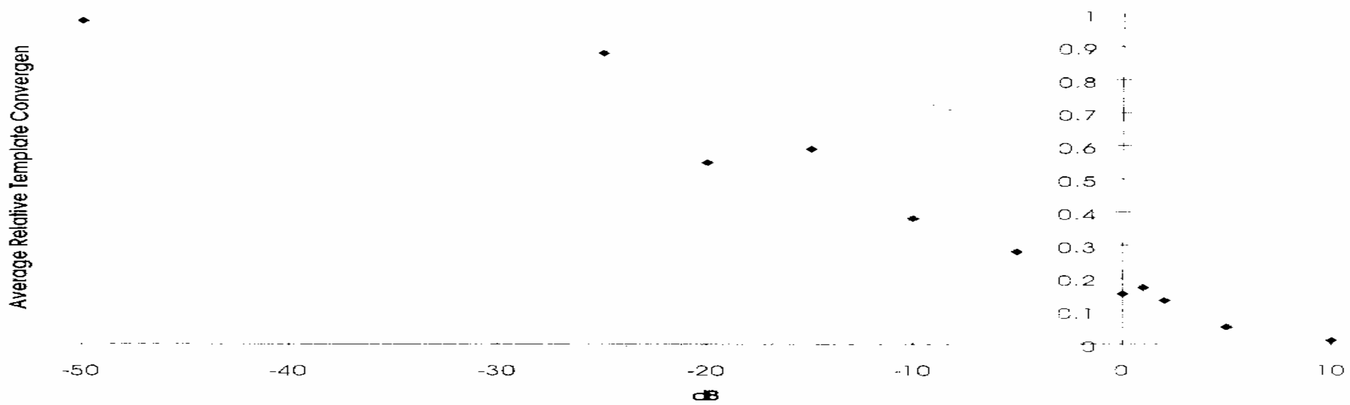
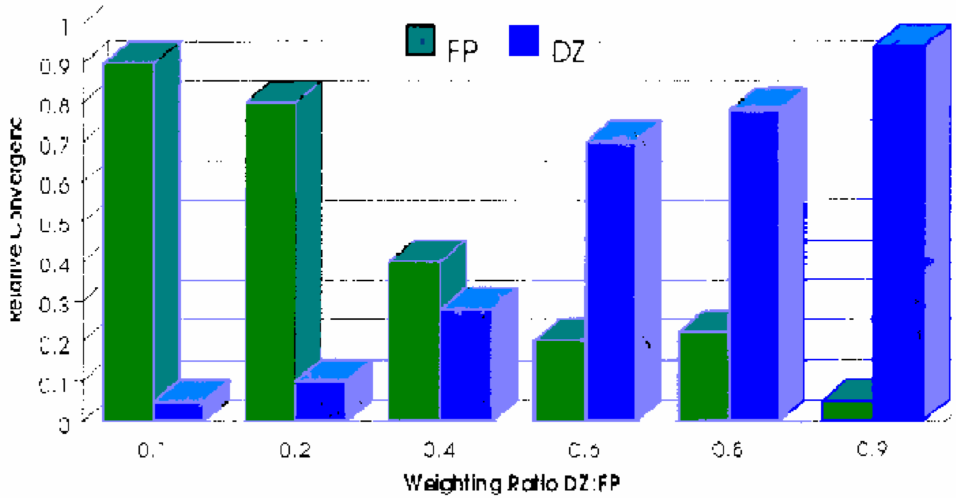


Figure 7.1. Input speech patterns before (top) and after (bottom) waveform altering for the utterance *hot*.



Average speaker recogntion performance against white noise corruption.

Speaker Identification



Relative speaker recogntion performance against competing speaker.

Speaker Identification

Conclusions

Achievements:

- Wavelets used as tokens for speech/speaker recognition
- Speaker classification against competing noise
- Reduction of feature space to 10% signal size

Limitations:

- Small speaker database
- Small vocabulary database
- "Offline" processing
- Reasonable speaker cooperation still assumed

Analysis Methods

- average power
- Fourier analysis
- wavelets
- fractal dimension
- entropy
- moments
- Hjorth parameters
- modular neural network

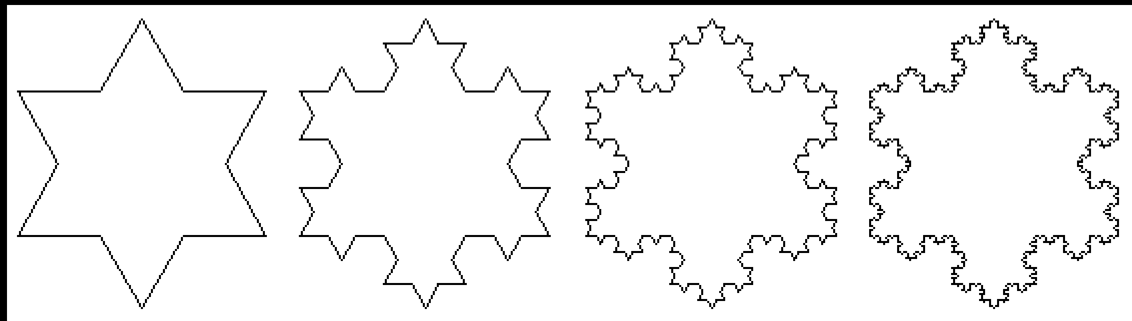
Fractal dimension

What is a fractal?

- Self-Similarity - small part should resemble the whole
- “an object whose Hausdorff- Besicovich (H-B) dimension strictly exceeds its topological dimension”
- Results from an recursive iterative equations
- Wiggly Lines or Surfaces

Fractal dimension

von Koch's Curve (1904) - "On a continuous curve without any tangent, obtained through an elementary geometrical construction



Each side L is replaced by 4/3L – length tends towards infinite – yet curve never goes outside circumcircle of original triangle or inside inner circle inside triangle

Fractal dimension

Mathematical Development

- Defining dimensions of objects

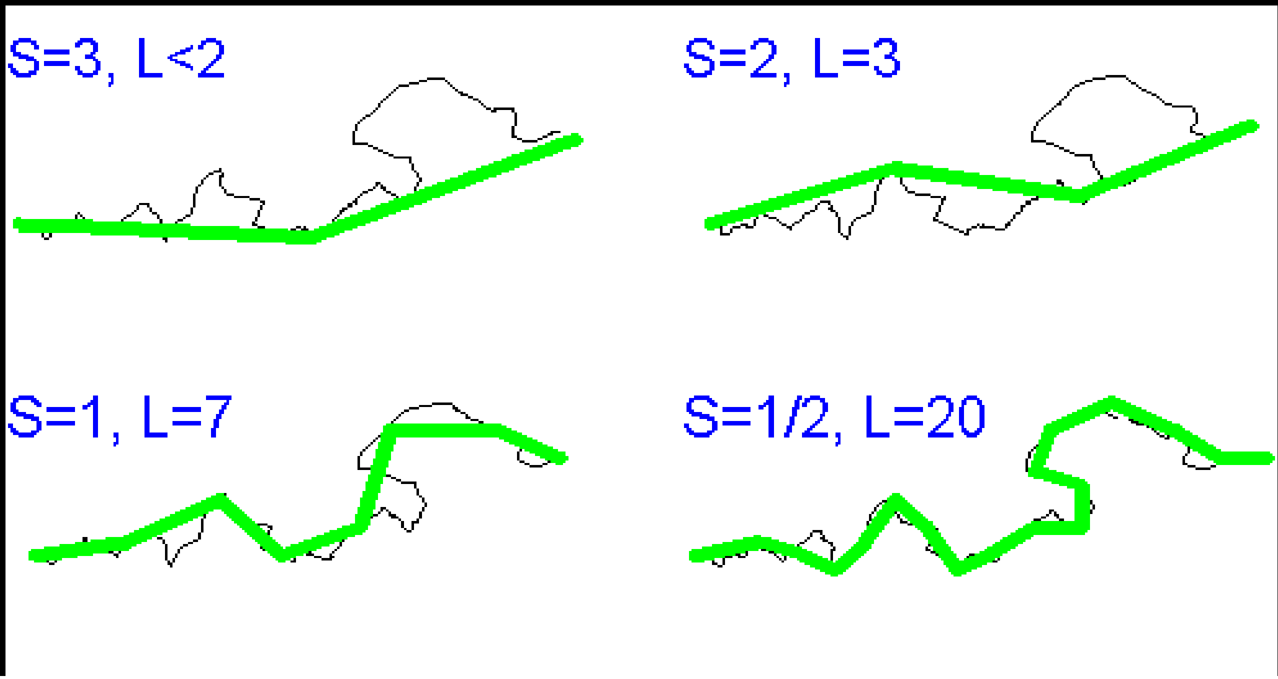
Euclidean Geometry

Point – 0D Line – 1D Plane – 2D Space – 3D

- Hausdorff(1919) & Besicovich(1935) – calculation of dimensions
- Von Koch's Curve H-B dimension log 4/log 3 = 1.2618...
- Cantor's dust H-B dimension log 2/log 3 = 0.6309...

Fractal dimension

- Proposed Fractal Dimension ("Fractional Dimension")
- Measuring the Coastline of England



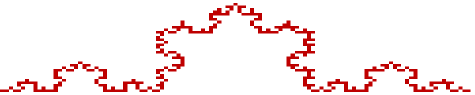
Fractal dimension

$$D = \frac{\log(L2/L1)}{\log(S1/S2)} = \frac{\log(20/7)}{\log(½/1)} = 1.51$$

L2, L1 are the measured lengths of the curves (in units)
S2, S1 are the sizes of the units (ie. the scales)

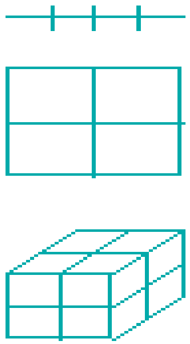
Fractal dimension

Self-similarity and Dimension



A **self-similar** shape can be constructed from **N** copies each scaled by a factor **s** from the original.

Simple Euclidean shapes are also self-similar.



1-D N parts scaled by $s = 1/N$ $\quad N = 1/s^1$

2-D N parts scaled by $s = 1/N^{1/2}$ $\quad N = 1/s^2$

3-D N parts scaled by $s = 1/N^{1/3}$ $\quad N = 1/s^3$

- **GENERALIZE:** for an object of N parts, each scaled by ratio s from the whole

$$N = 1/s^D = (1/s)^D$$

- defines the fractal (or similarity) dimension

$$D = \frac{\log N}{\log 1/s}$$

Fractal dimension

- **Fractals and Images**
  - Measurement of the texture or roughness of an image
  - The higher the FD the rougher the surface
- Methods of calculating Fractal Dimension
  - **Statistical differences in pixel intensity**
  - Box counting method
  - Gabor filters
  - Wavelets

Fractal dimension Sarkar and Chaudhuri's algorithm

- Start with a M x M image, G levels of gray scale and D = log(N)/log(1/r)
- D = dimension, N = number of parts comprising the set, scaling of 1/r from whole
- For a square: N parts scaled by 1/N1/2, thus Nr2 = 1 or D = 2
- Divide up the image into size s x s where M/2 > s > 1 such that r = s/M
- Imagine the two dimensional image is a topological map in three dimensions. On each size grid s x s can be built a column of boxes sized s x s x s' where $\lfloor G/s' \rfloor = \lfloor M/s \rfloor$ with indices starting with 1 for the bottom box.
- Find the lowest and highest boxes intersected by the image in the current column of boxes and name them k and l respectively.
- Add up the differences (1 – k + l) for all areas s x s for the current scale r and call it N(r)
- Do this for all scales and the result will be a vector N(r) where 1/r = 2, 4, 8, ….M/2
- Plot log(N[r]) vs. log(1/r) and calculate the slope using a least square linear fit..

This is the fractal dimension

Evangelia Micheli-Tzanakou, PhD

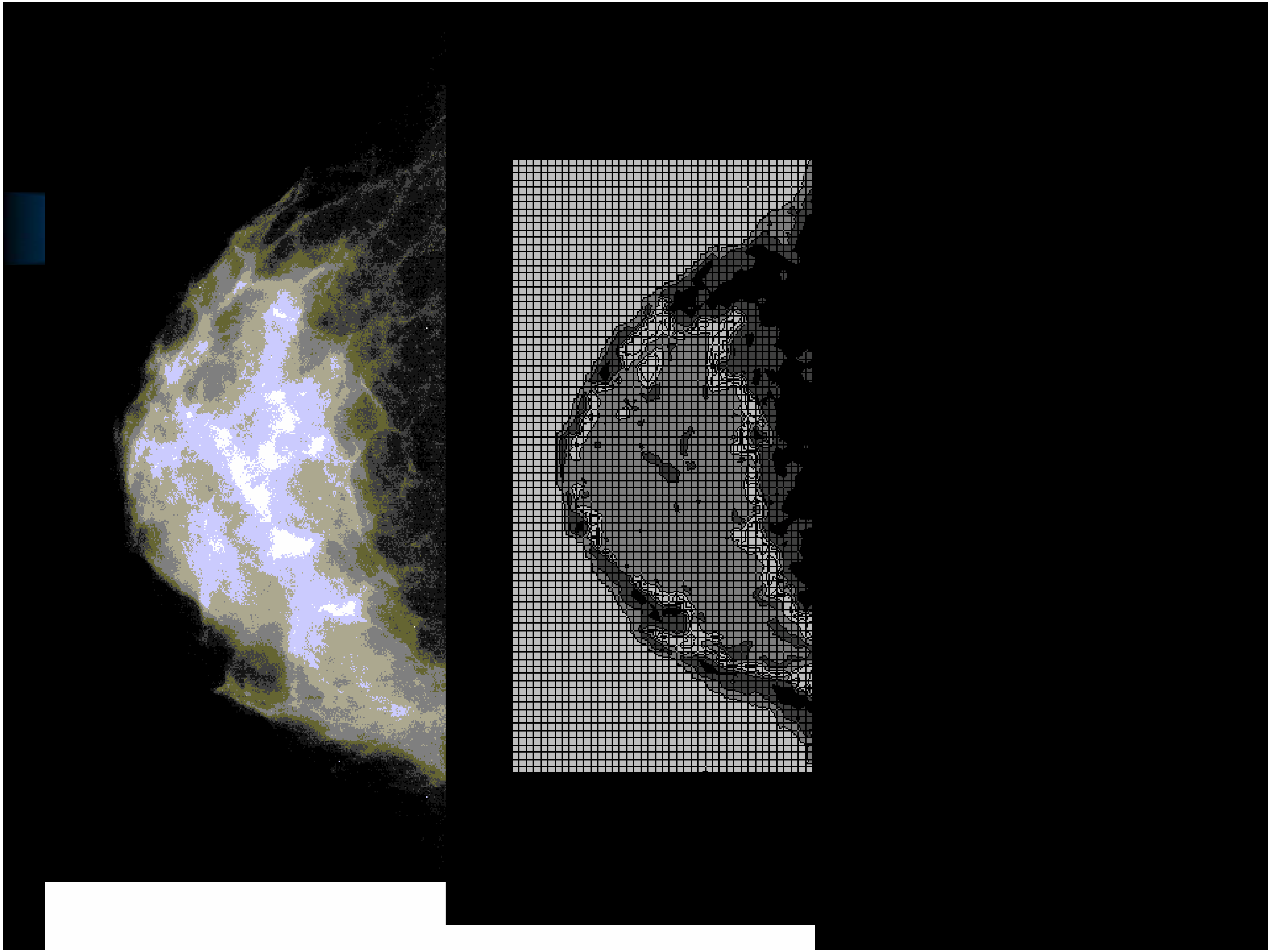
An example…..

MAMMOGRAPHY

The leading cause of death of women affected by breast cancer

Classification is performed in two basic steps:

- feature extraction
- neural network classification



Network Topology

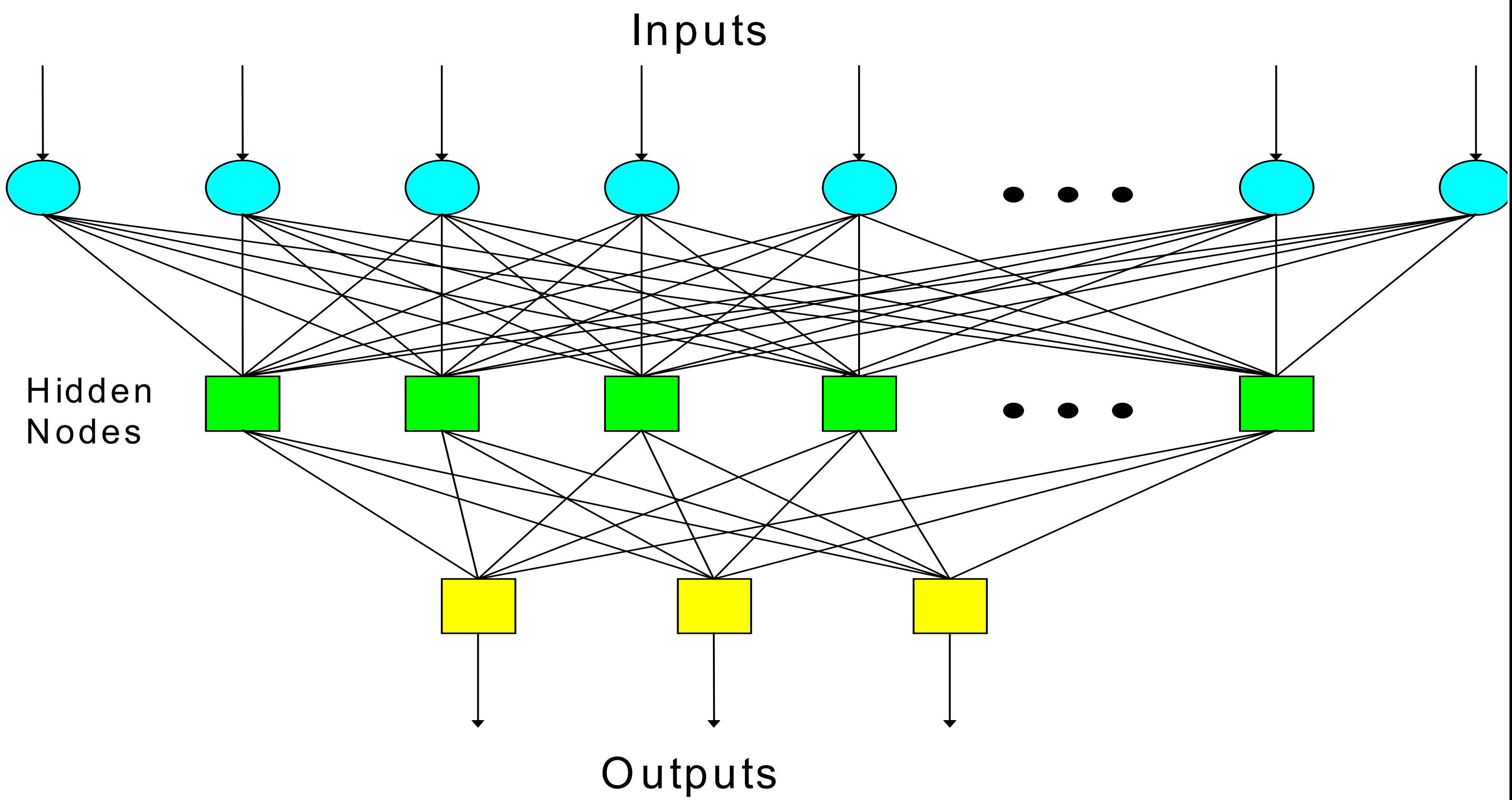
The two basic types of network topologies used in our experiments were:

A Three Layer Network

- one input, one hidden, and one output layer, classified between the three types of images by using three output nodes
    - Normal
    - Mass
    - Microcalcifications

This type of architecture did not identify exactly the three types of patterns

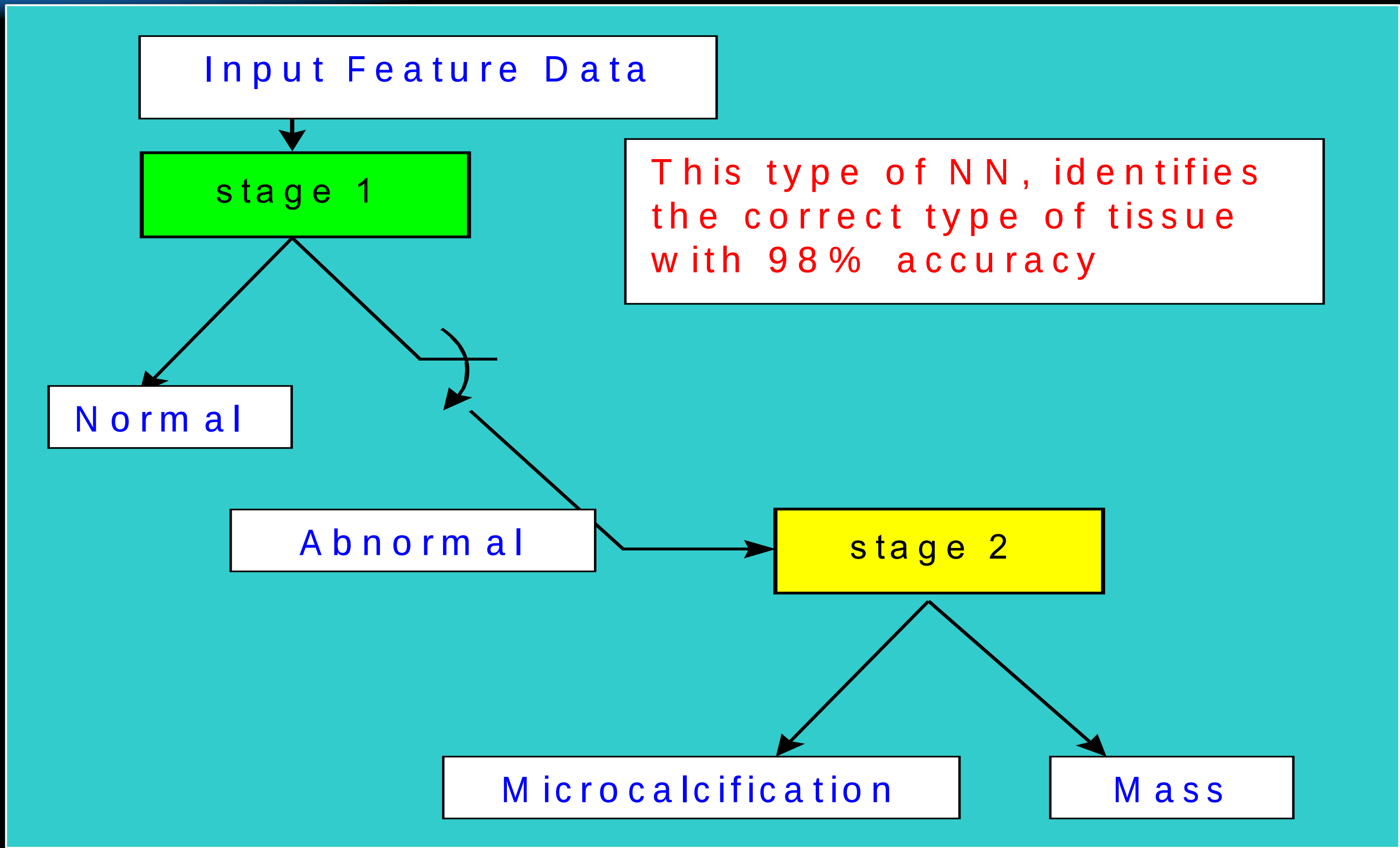
NN ARCHITECTURE



A Binary Tree Network

- images were classified into two categories at a time
- each stage contained a single three layer network as in the three layer NN, however each three layer network contained only two output units

Binary Tree NN

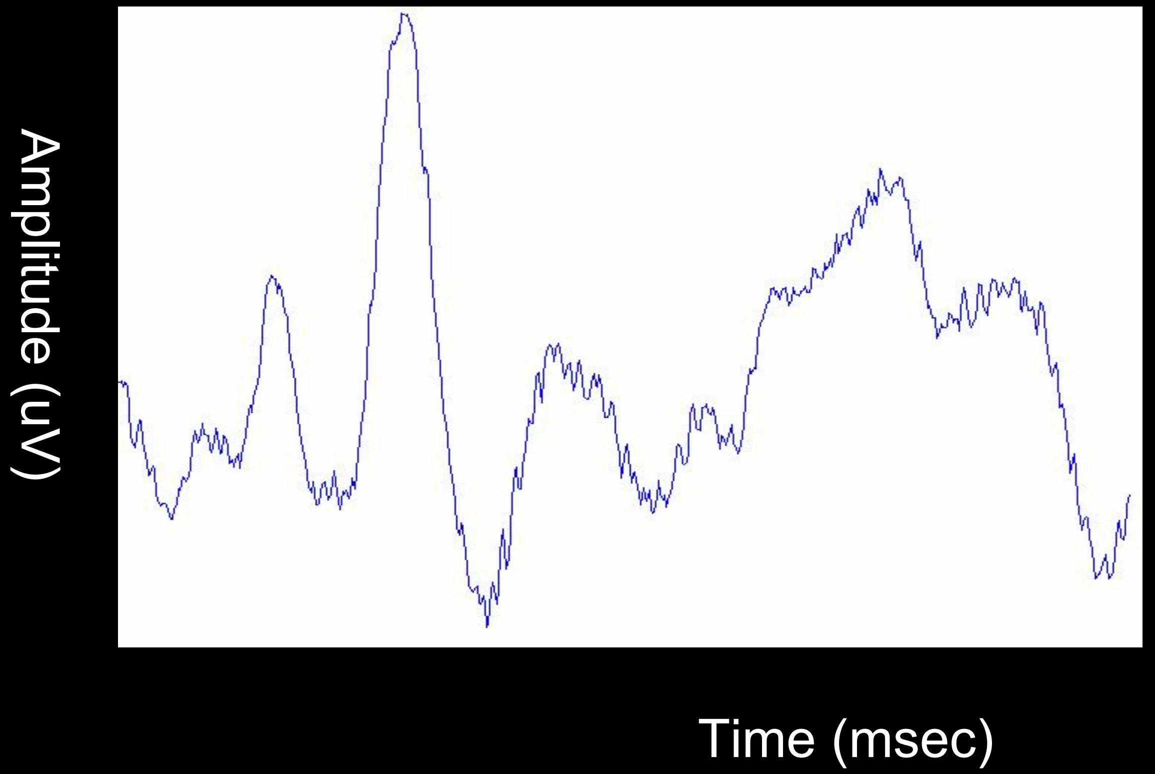


Another example…..

In signal processing

Fractal Analysis of EMG & Evoked Potential Signals

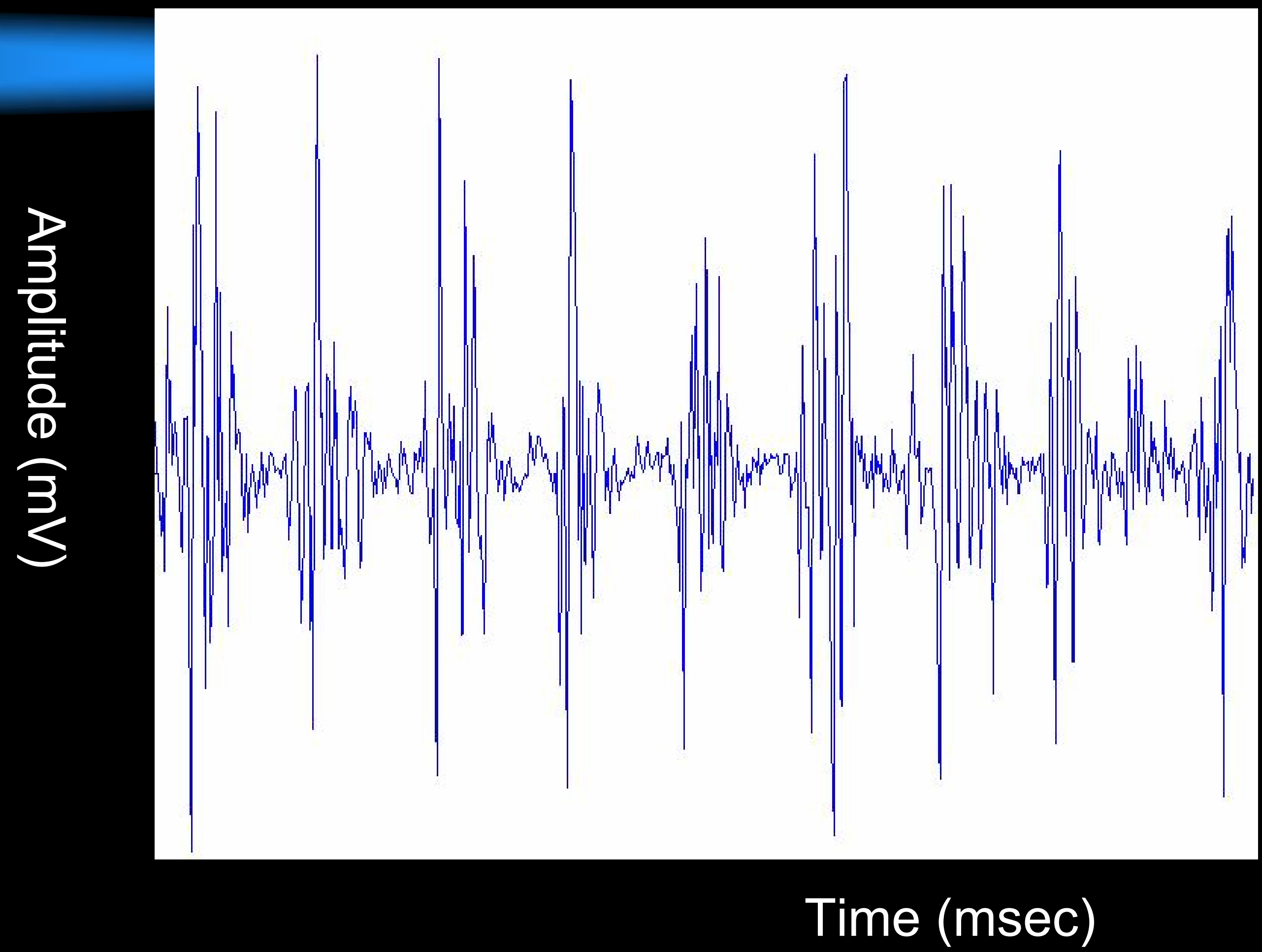
Evoked Potential Signal



EMG (Electromyography) Signals

- EMG is a test that measures muscle response to nervous stimulation (electrical activity within muscle fibers).
- The electromyography (EMG) measures the response of muscle fibers to electrical activity. It's used to help determine the kind of muscle condition that might be causing muscle weakness, including muscular dystrophy and nerve dysfunctions.

EMG Signal



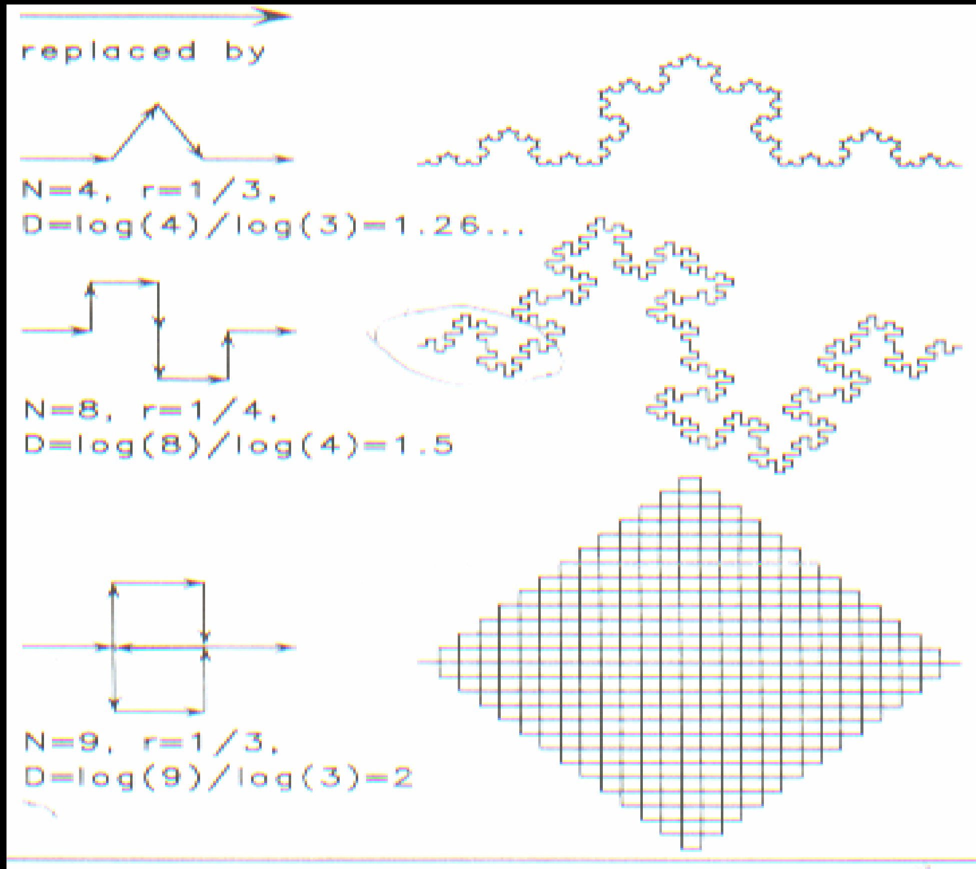
Evangelia Micheli-Tzanakou, PhD

Fractal Dimension

- In medicine, waveforms showing repetitive patterns (ECG, EEG, EMG) are often analyzed in the terms of Fractal Dimension.

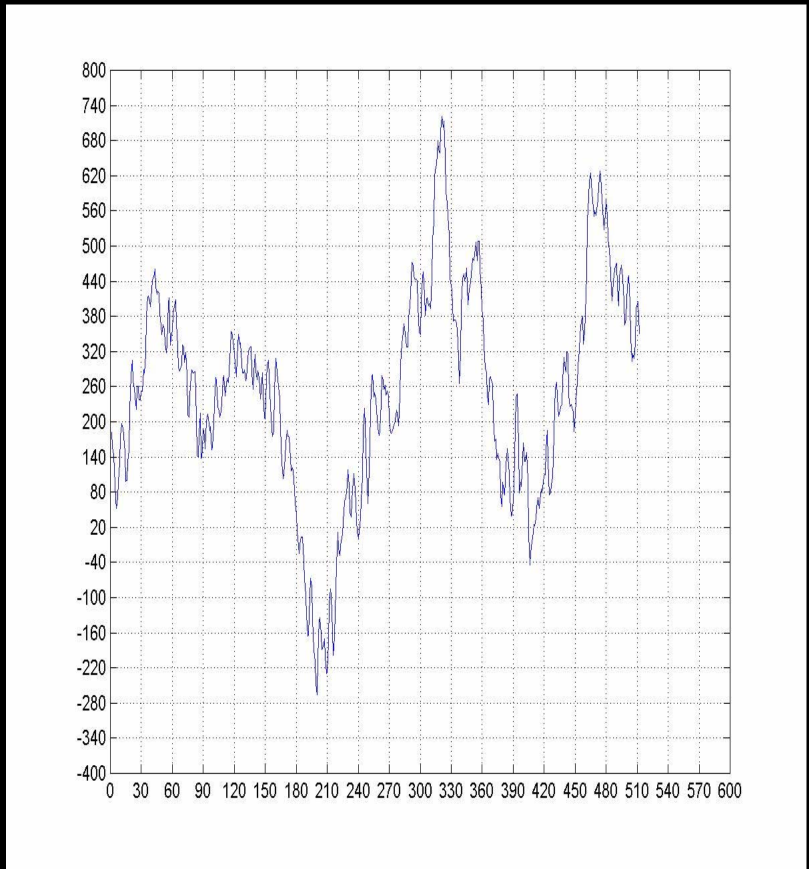
Fractal Dimension

- Fractals are of rough or fragmented geometric shape that can be subdivided in parts, each of which is approximately a reduced copy of the whole.
- Fractal Dimension measures the degree of fractal boundary fragmentation or irregularity over multiple scales
- D=log(N)/log(1/r)



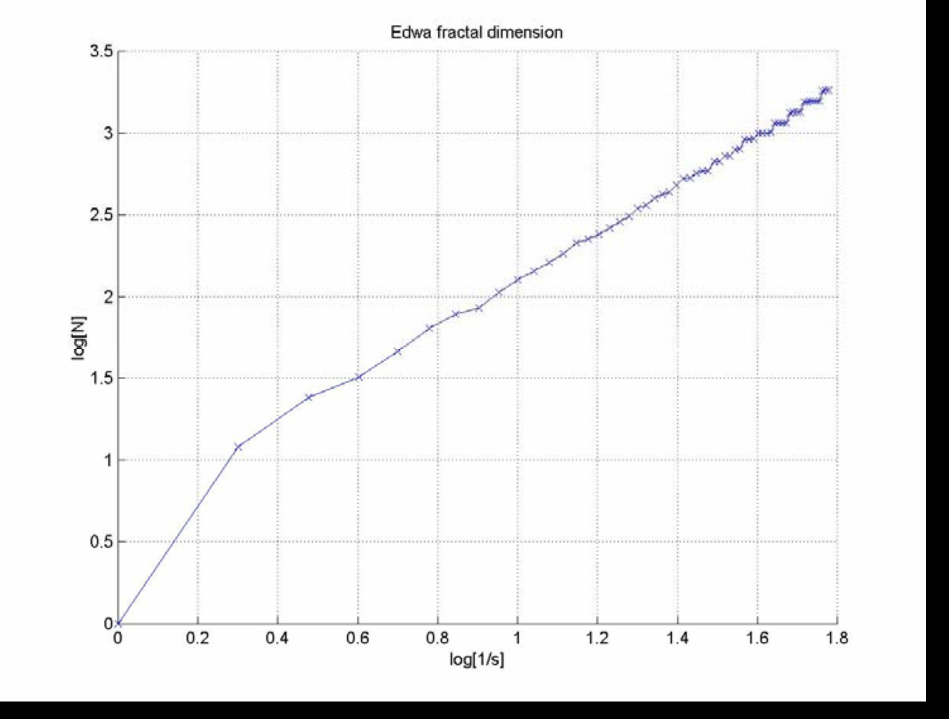
Fractal Dimension

- Box-Counting Method (Barnsley, 1988): It works by covering fractal (its image) with boxes (squares) and then counting how many boxes are needed to cover the fractal completely. Repeating this measurement with different sizes of boxes will result into logarithmical function of box size *(*x-axis) and number of boxes needed to cover the fractal *(*y-axis). The slope of this function is referred as box dimension. Box dimension is taken as an appropriate approximation of fractal dimension.



Fractal Dimension

- $D_{\text{box-counting}} = \Delta \log N(s) / \Delta \log(1/s)$



Results

- D of Evoked Potential Signals

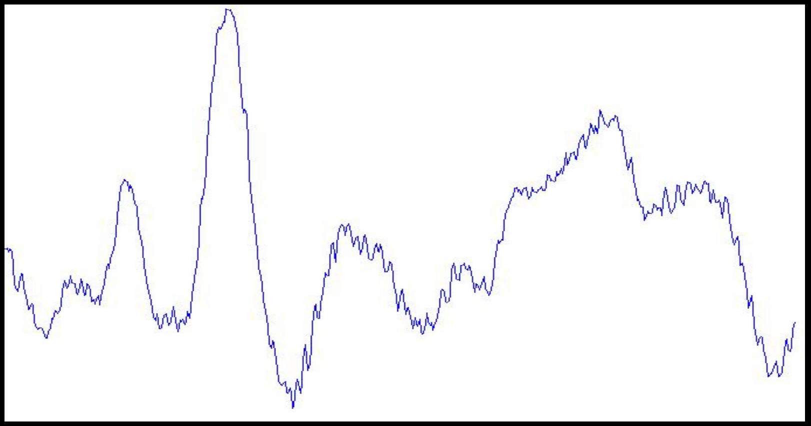
1.3996	1.3532
1.4193	1.3270
1.2070	1.4587
1.2584	1.3730

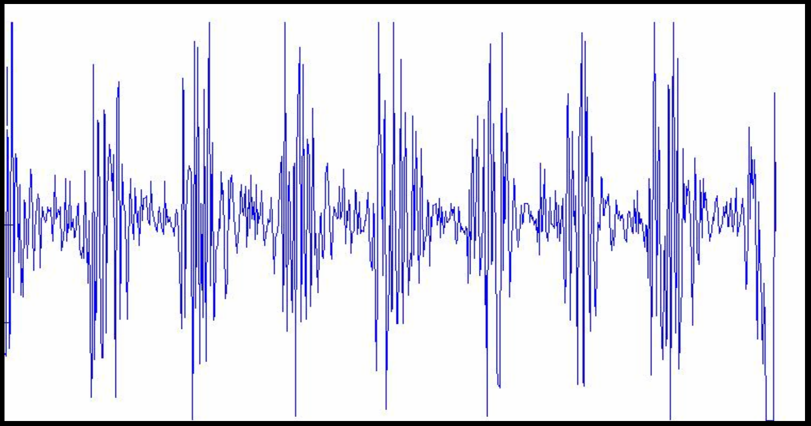
- D of EMG Signals

1.6016	1.4914
1.4674	1.4058
1.4450	1.4104
1.5500	

Results

- Fractal Dimension indicates the fragmentation or irregularity of the signal curve over multiple scales.
- D1=1.2070 (upper)
- D2=1.6016 (lower)





Discussion

- The results above suggest that fractal dimension may be useful as alternative means to evaluate the EMG and Evoked Potential signals. High D value may mean muscle’s irregular state of trembling, which is one symptom of Parkinson's disease.

Analysis Methods

- average power
- Fourier analysis
- wavelets
- fractal dimension
- entropy
- moments
- Hjorth parameters
- modular neural network

Entropy

- Information content in a source is denoted by entropy: $H = -\Sigma p_i \log_2 p_i$ (bits)
- Shannon Coding Theorem states that a source with entropy H can be encoded with an arbitrarily small error probability at rate R bits/source output as long as

$$R > H$$

Entropy

- What does this have to do with the wavelet transform?
  - the wavelet transform changes the <u>statistics</u> of the image
  - has the potential to decrease entropy depending on the image being transformed

Analysis Methods

- average power
- Fourier analysis
- wavelets
- fractal dimension
- entropy
- moments
- Hjorth parameters
- modular neural network

Moments

$$m_{p,q} = \iint_R f(x,y)\, x^p y^q \, dx\, dy$$

Hu, 1962: **Central:**

$$\mu_{p,q} = \iint (x-\bar{x})^p (y-\bar{y})^q f(x,y)\, dx\, dy$$

$$\bar{x} = \frac{m_{1,0}}{m_{0,0}}, \quad \bar{y} = \frac{m_{0,1}}{m_{0,0}}$$

Moments

Invariant Moments

$$\phi_1 = \mu_{2,0} + \mu_{0,2}$$

$$\phi_2 = (\mu_{2,0} - \mu_{0,2})^2 + 4\mu_{1,1}^2$$

$$\phi_3 = (\mu_{3,0} - 3\mu_{1,2})^2 + (\mu_{0,3} - 3\mu_{2,1})^2$$

$$\phi_4 = (\mu_{3,0} + \mu_{1,2})^2 + (\mu_{0,3} + \mu_{2,1})^2$$

$$\phi_5 = (\mu_{3,0} - 3\mu_{1,2})(\mu_{0,3} + \mu_{1,2})[(\mu_{3,0} + \mu_{1,2})^2 - (\mu_{2,1} + \mu_{0,3})^2] + (\mu_{0,3} - 3\mu_{2,1})(\mu_{0,3} + \mu_{2,1})[(\mu_{0,3} + \mu_{2,1})^2 - (\mu_{1,2} + \mu_{3,0})^2]$$

$$\phi_6 = (\mu_{2,0} - \mu_{0,2})[(\mu_{3,0} + \mu_{1,2})^2 - (\mu_{2,1} + \mu_{0,3})^2] + 4\mu_{1,1}(\mu_{0,3} + \mu_{1,2})(\mu_{0,3} + \mu_{2,1})$$

$$\phi_7 = (3\mu_{2,1} - \mu_{0,3})(\mu_{3,0} + \mu_{1,2})[(\mu_{3,0} + \mu_{1,2})^2 - 3(\mu_{2,1} + \mu_{0,3})^2] + (\mu_{3,0} - 3\mu_{2,1})(\mu_{0,3} + \mu_{2,1})[(\mu_{0,3} + \mu_{2,1})^2 - 3(\mu_{1,2} + \mu_{3,0})^2]$$

Moments values

$|10^{-40}|$ to $|10^{41}|$ (for 256x256 images)

$$x' = \ln(|\ln(|x|)|)$$

Moments

- Have been used successfully both in one and two dimensional data.

Analysis Methods

- average power
- Fourier analysis
- wavelets
- fractal dimension
- entropy
- moments
- Hjorth parameters
- modular neural network

Hjorth parameters (coefficients)

- Another way of looking at features using moments and their higher order combinations
- Mostly used with one dimensional data
- These are:
  - Activity
  - Mobility
  - Complexity

Hjorth Coefficients

TABLE I

Derivation of parameters

Frequency		Time
Spectral moment	Parameter definition	Time operation
m_0	$m_0 = \text{activity} = \sigma_a{}^2$	$f(t)$
$m_1 = 0$	—	—
m_2	$(m_2/m_0)^{\frac{1}{2}} = \text{mobility} = \sigma_d/\sigma_a$	$\mathrm{d}(f)/\mathrm{d}t$
$m_3 = 0$	—	—
m_4	$\dfrac{(m_4/m_2)^{\frac{1}{2}}}{(m_2/m_0)^{\frac{1}{2}}} = \text{complexity} = \dfrac{\sigma_{dd}/\sigma_d}{\sigma_d/\sigma_a}$	$\dfrac{\mathrm{d}^2(f)}{\mathrm{d}t^2}$
$m_5 = 0$	—	—
$m_6, m_8, \ldots$	....	....

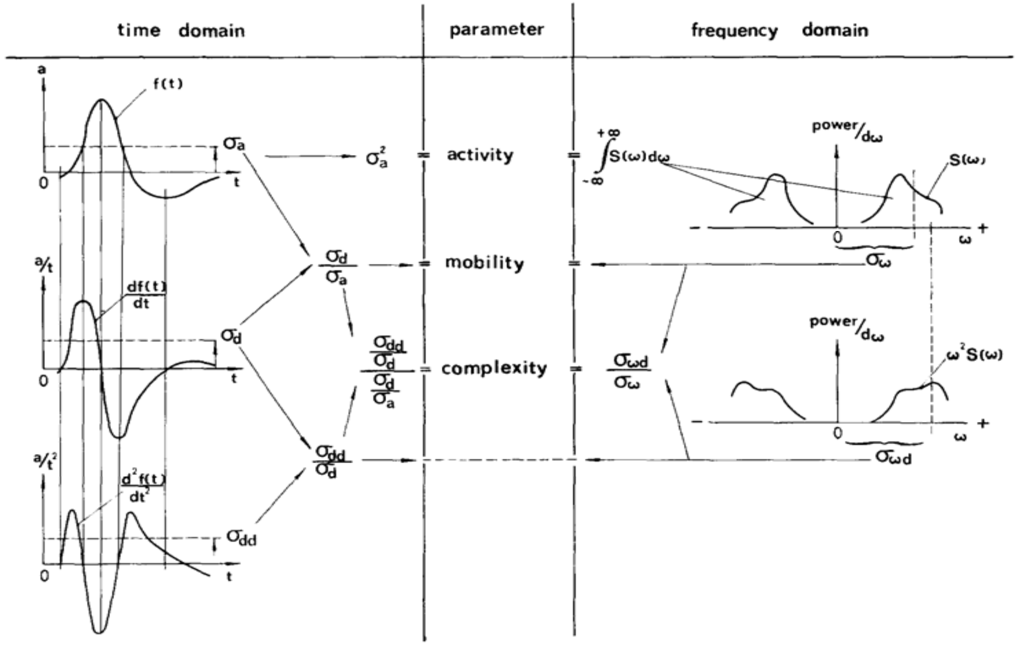


Fig. 1
Parametric connections between time domain properties and spectral characteristics of an arbitrary signal.

[Hjorth 1970]

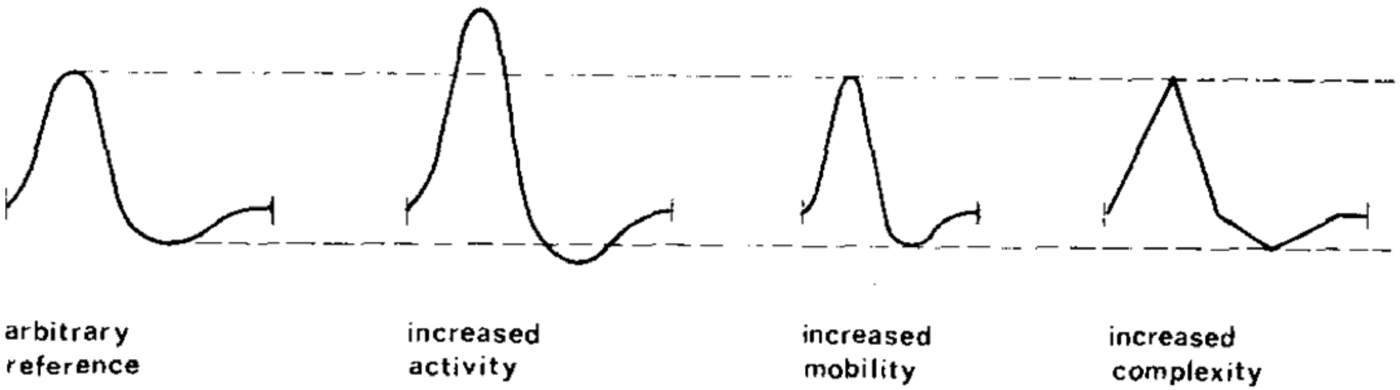


Fig. 3
Characteristic changes of a curve shape, illustrating the dependence of the individual parameters.

Time-Frequency Analysis

TOP:

VEP waveform from data file

X axis: Time (ms)

Y axis: Amplitude (mv)

MIDDLE:

Time-Frequency Analysis

X axis: Time (ms)

(matches Top timescale)

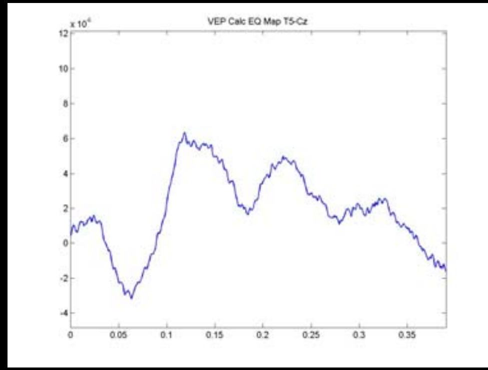
Y axis: Frequency (Hz)

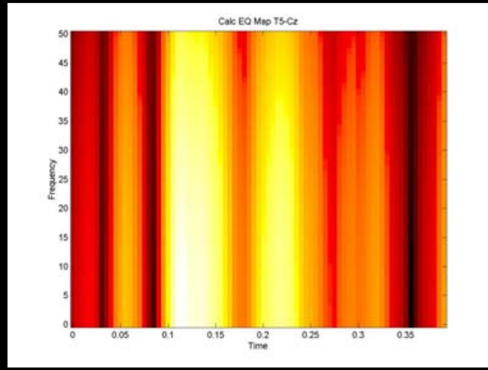
BOTTOM:

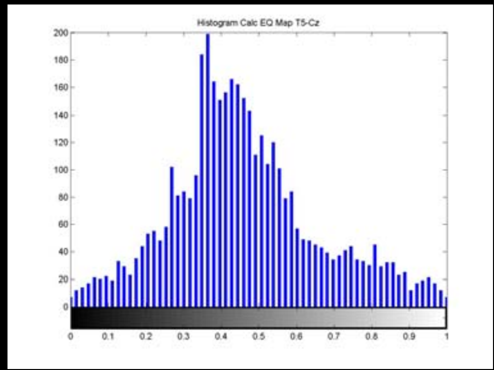
Histogram of Time-Frequency Amplitudes

X axis: Normalized Amplitudes (0–1)

Y axis: Count of Amplitudes in Time-Frequency Space

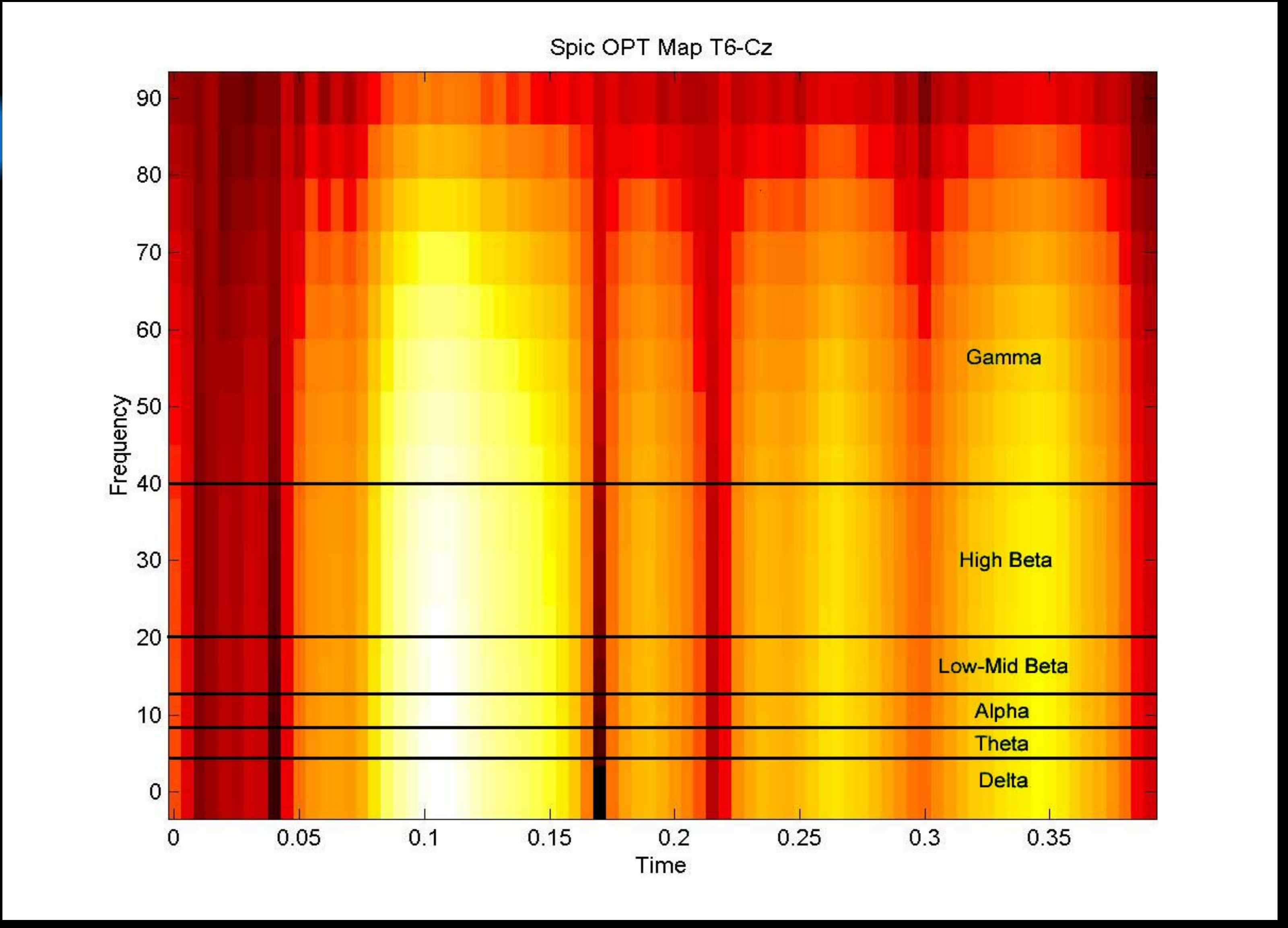






Evangelia Micheli-Tzanakou, PhD

Brain Frequencies



Evangelia Micheli-Tzanakou, PhD

Analysis Methods

- average power
- Fourier analysis
- wavelets
- fractal dimension
- entropy
- moments
- Hjorth parameters
- modular neural network

Modular Neural Networks

- Once you have all these features, what do you do with them?
- Use a Modular Neural Network with each module processing a different set of features
  - Integrate all “input networks” into one for the final output

An example…..

An example…..

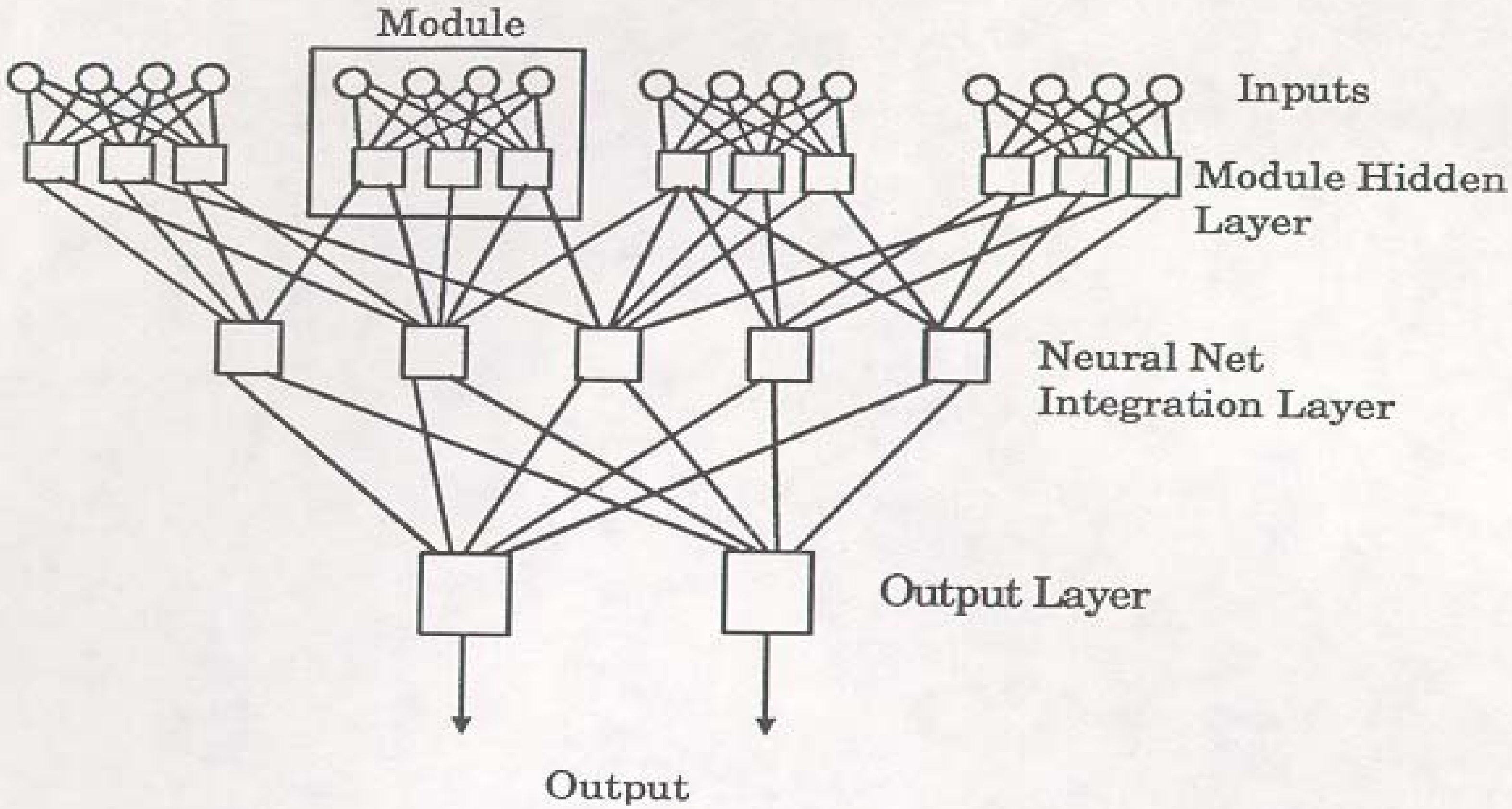


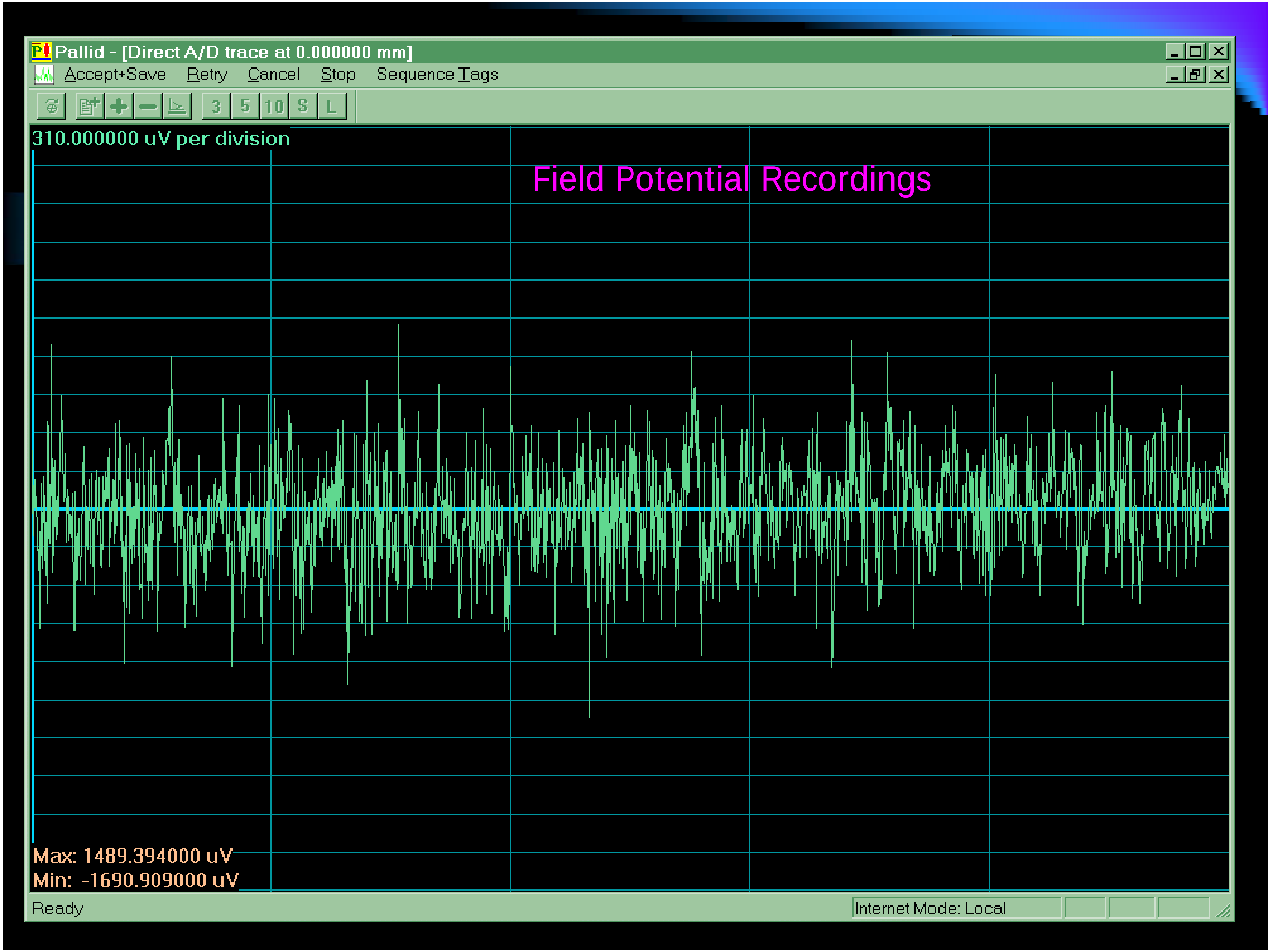
Figure 3.5 - An Example of a Modular Neural Network
The boxed area highlights one module. The Module Hidden Layer and the Neural Net Layer are not fully connected just for clarity. Normally these two layers would be fully connected Only the input layer to the Module Hidden Layer are modularized.

We developed a system to analyze spontaneous activity within the Globus Pallidus of Parkinson’s patients and able to:

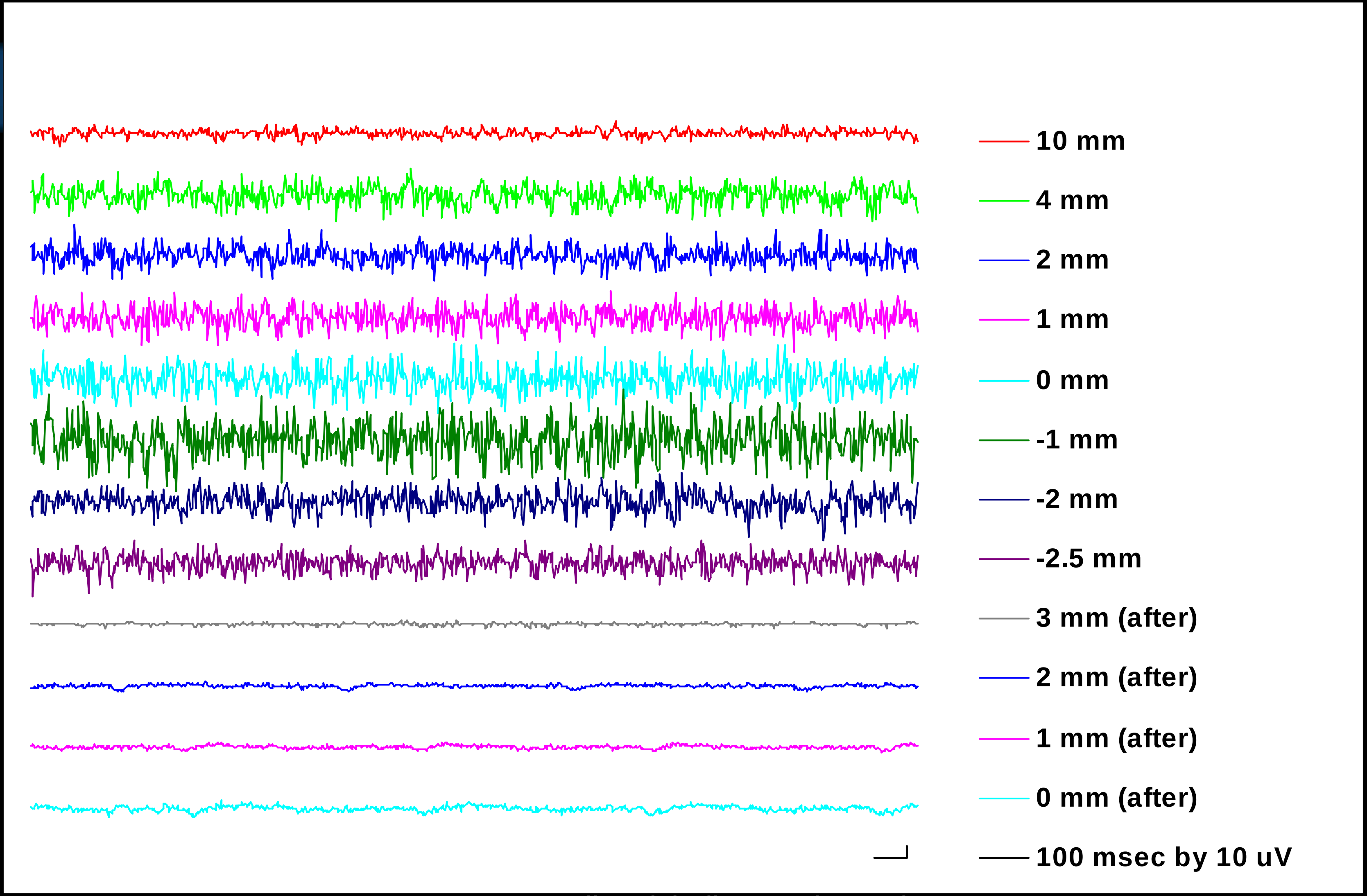
- Rate the degree to which proposed lesions at specific locations along the current surgical tract are expected to relieve Parkinsonian symptoms
- Rate the degree to which proposed lesions at specific locations along the current surgical tract are expected to cause unwanted effects such as scotoma and/or dysarthria

Localization Methods

- Imaging
- Stimulation testing
- Recordings of spontaneous activity
  - mapping boundaries of pallidum?
  - extends duration of procedure


Pallid - [Direct A/D trace at 0.000000 mm]
Accept+Save Retry Cancel Stop Sequence Tags
3 5 10 S L
310.000000 uV per division
Field Potential Recordings
Max: 1489.394000 uV
Min: -1690.909000 uV
Ready
Internet Mode: Local

Activity Recordings



Patient Data: Efficacy Assessment

- patients were examined by neurologists and neurosurgeons before and after pallidotomy
- bradykinesia, rigidity, tremor, and dyskinesia rated on 5-point scales
- “after” results taken as close to six months after operation as possible
- improvement mapped to a 0-5 scale
- “best” improvement (of bradykinesia, rigidity, tremor, dyskinesia) and “average” improvement used to train network

Patient Data: Deleterious Outcomes

Rating	Deleterious Outcome
5	death
4	stroke, meningitis
3.5	confusion, hallucinations; difficulty swallowing
3	dysarthria; measurable field cut
2.5	diplopia
2	slowed speech; visual disturbance of lesser severity than a measurable field cut or diplopia
1	decreased speech volume
0	no hazardous outcome noted

Patient Data: Deleterious Outcomes

Rating	Duration of Deleterious Outcome
5	greater than one year
4	6-12 months
3	1-6 months
2	1-4 weeks
1	up to one week
0	no hazardous outcome detected

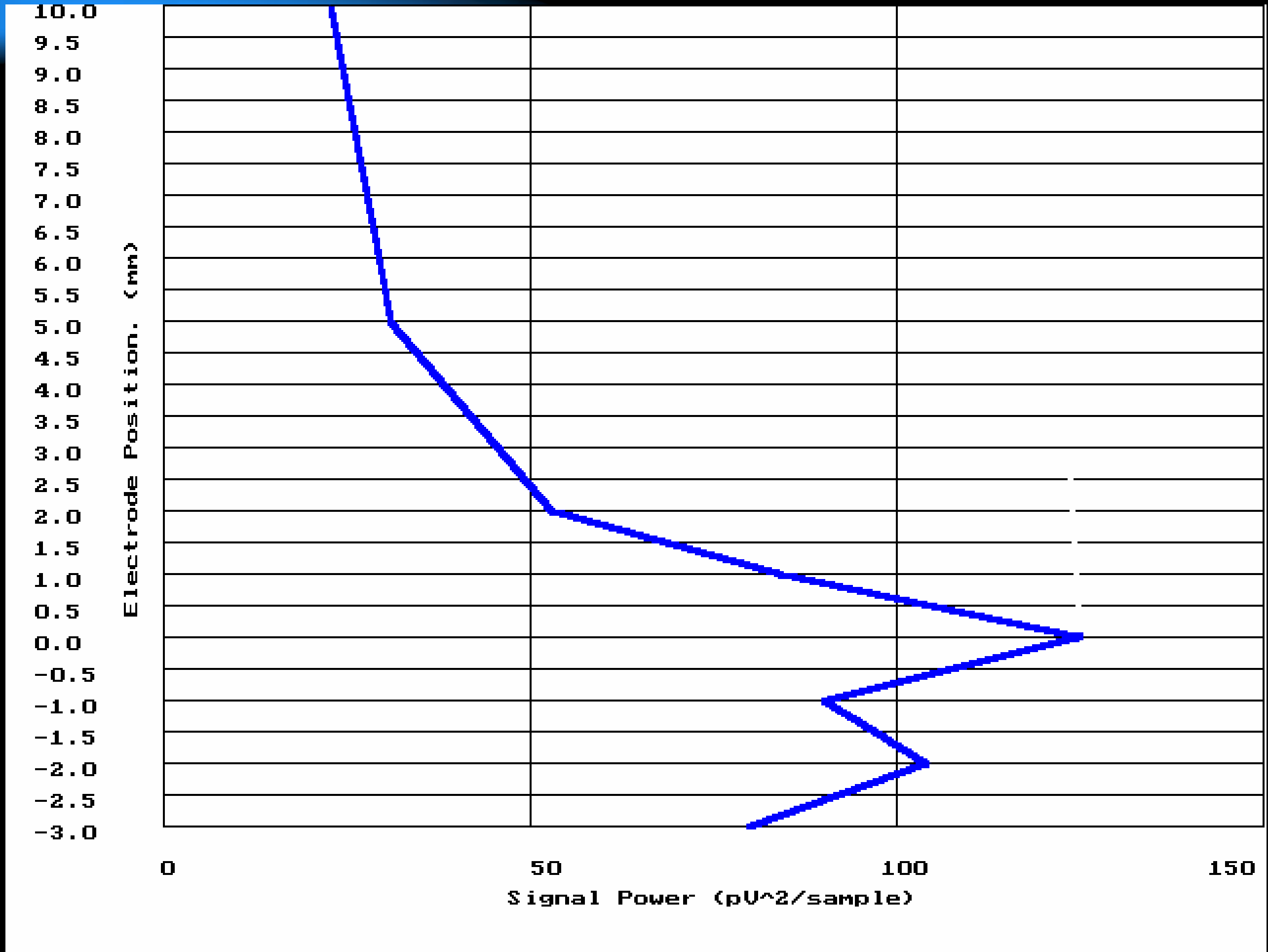
Patient Data: Multiple Data Segments

- Recordings at each site often contain more samples than are needed for analysis techniques.
- When "extra" data existed, the network was trained with up to 10 different views of the analysis results for each patient.

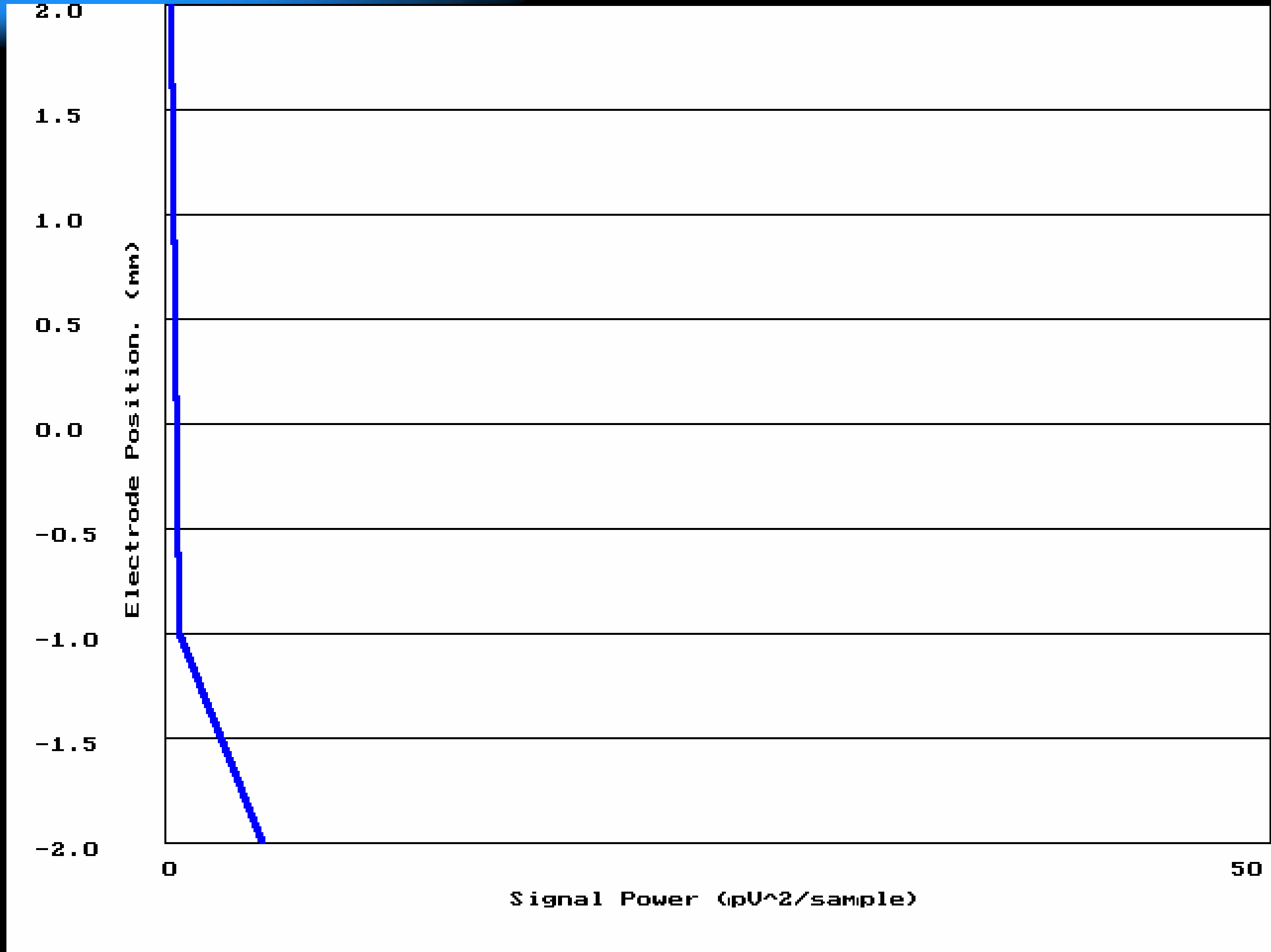
Analysis Methods: "Toolkit"

- average power (already in use on-line)
- frequency-based
  - Fourier analysis
  - wavelets
- complexity measures (used off-line)
  - fractal dimension
  - entropy
- moments
- pre-operative information

Analysis Methods: power analysis

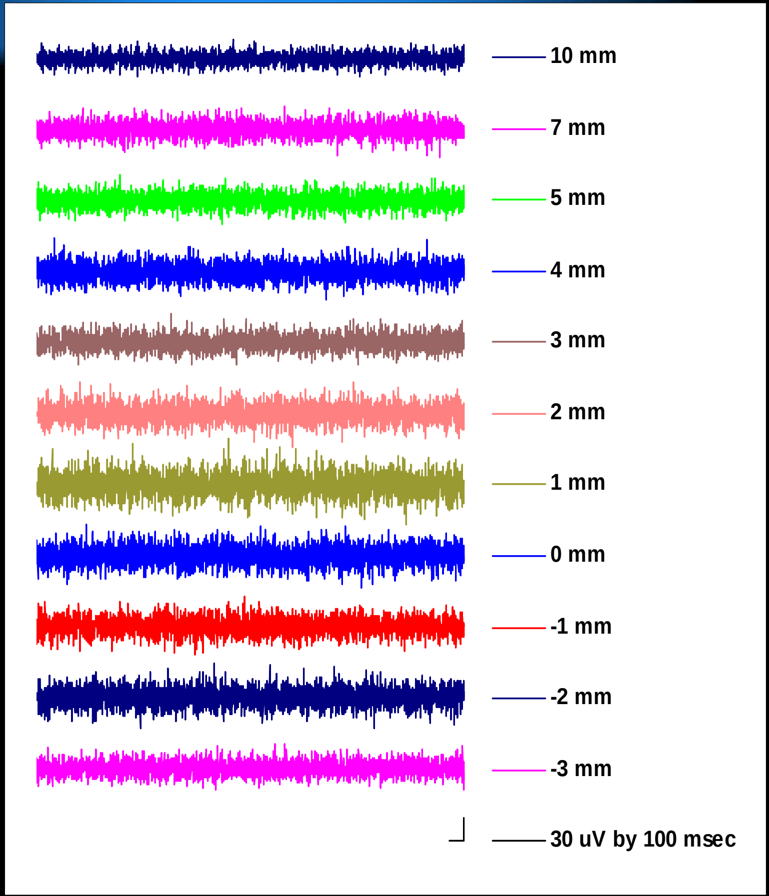


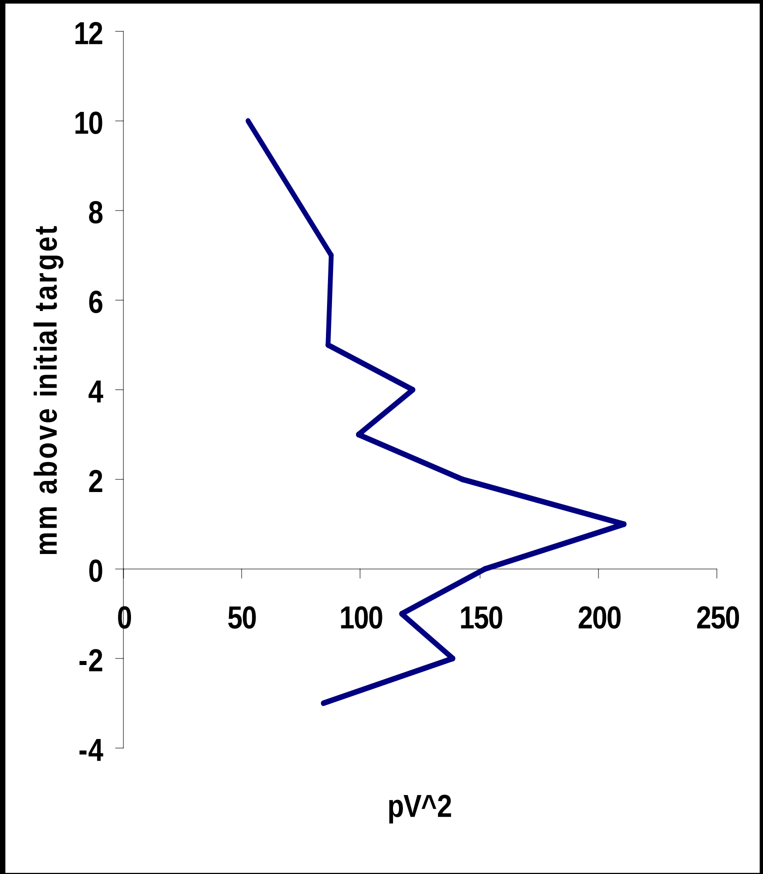
Analysis Methods: power analysis



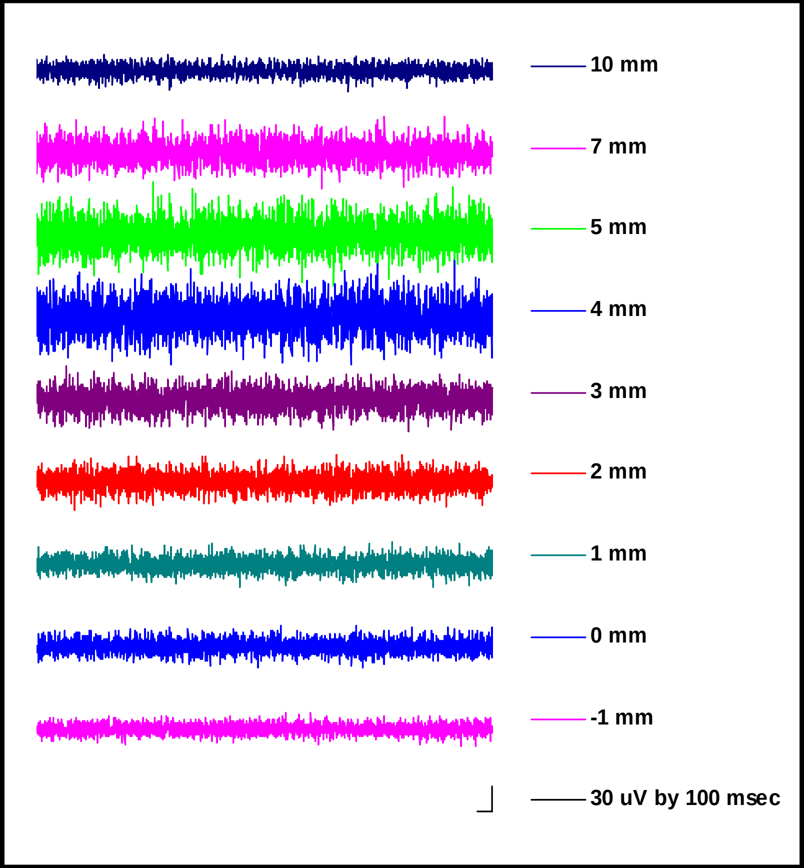
Evangelia Micheli-Tzanakou, PhD

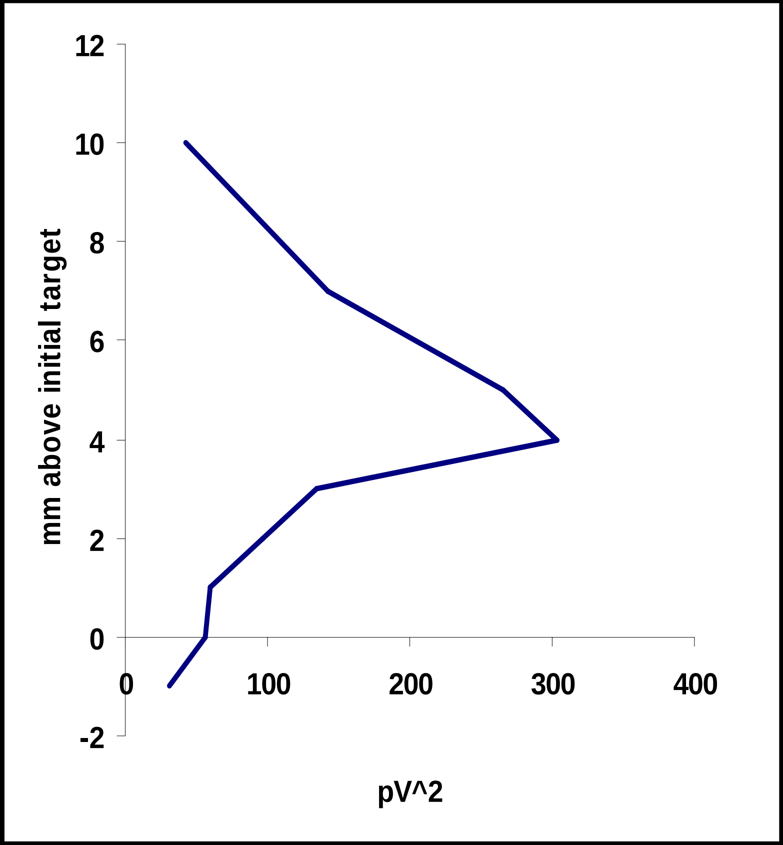
AT, before lesioning





VQ, right side, before





Evangelia Micheli-Tzanakou, PhD

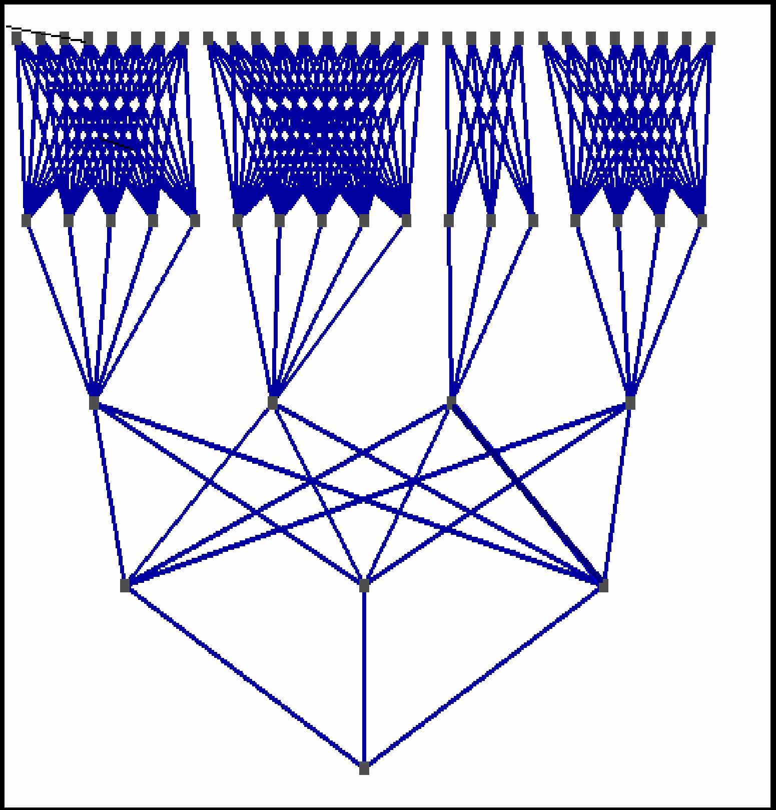
Neural Networks: ALOPEX

- $W_i(n) = W_i(n-1) + \delta_i(n)$
- $\delta_i(n) = \pm\delta$ with probability $p_i(n)$
- $\delta_i(n) = \mp\delta$ with probability $1-p_i(n)$

$$p_i(n) = \frac{1}{1 + \exp\left(\frac{\Delta W_i \Delta R}{\text{temperature}(n)}\right)}$$

Neural Networks: Architecture

- inputs encode results of different analysis techniques, possible lesion locations
- output represents estimated efficacy or hazard



Data Used

- **Obtained before**
- **During and**
- **After the operation**

Artificial Data: Why?

- patient data describe locations that were lesioned
- no patient data available for locations that were *not* lesioned
- network trained only with data from "good" locations and will not recognize "bad" locations

Artificial Data: Types

- Lesion too high
  - no benefit, minimal hazard if 3mm above highest actual lesion
- Lesion too low
  - 2 mm below lowest actual lesion: unknown benefit, risk varies with size
  - >2 mm below lowest actual lesion: no benefit, severe hazard
- Subsets and combinations

Artificial Data

Description	average impr.	best impr.	hazard	hazard duration	location
actual lesion and outcome	**1.17**	**1.50**	**0.00**	**0.00**	**{2, 3}**
zero-benefit lesions	0.00	0.00	0.00	0.00	{7}
	0.00	0.00	0.00	0.00	{6}
	0.00	0.00	0.00	0.00	{6, 7}
	1.17	1.50	0.00	0.00	{2, 3, 7}
	1.17	1.50	0.00	0.00	{2, 3, 6}
	1.17	1.50	0.00	0.00	{2, 3, 6, 7}

Artificial Data: Balancing Act

- The available pool of artificial data far exceeds the amount of actual outcome data
- Only a portion of available artificial data was used
- Final training set included between 1-4 items of artificial data for each case.
- “Too low” lesions
  - rejected by neurosurgeons because of side effects
  - associated by network with increased side effects

Results: Plausibility

- "Too high" lesions
  - additional lesions near the target area have a greater benefit than those farther away
  - additional lesions frequently associated with *reduction* in overall benefit

Results: Flexibility

- Unusually low target
- Unusually high target
- Unusually high hazard area
- Unusually modest benefit

Results: Flexibility

- Unusually low target
  Actual: (-3, -2, -1, 0, 1, 2) with no adverse effect
  Network: (-4, -3, -2), hazard rating < 0.6
- Unusually high target
- Unusually high hazard area
- Unusually modest benefit

Results: Flexibility

- Unusually low target
  Actual: (-3, -2, -1, 0, 1, 2) with no adverse effect
  Network: (-4, -3, -2), hazard rating < 0.6
- Unusually high target
  Actual: (2, 4, 5) with no adverse effect
  Network: (2, 3, 4), (3, 4, 5), (2, 4, 5), and (3, 4, 6)
- Unusually high hazard area

- Unusually modest benefit

Results: Flexibility

- Unusually low target
  Actual: (-3, -2, -1, 0, 1, 2) with no adverse effect
  Network: (-4, -3, -2), hazard rating < 0.6
- Unusually high target
  Actual: (2, 4, 5) with no adverse effect
  Network: (2, 3, 4), (3, 4, 5), (2, 4, 5), and (3, 4, 6)
- Unusually high hazard area
  Actual: patient saw flashes of light at 0 mm
  Network: Hazard increased for combinations with 0mm
- Unusually modest benefit

Results: Flexibility

- Unusually low target
  Actual: (-3, -2, -1, 0, 1, 2) with no adverse effect
  Network: (-4, -3, -2), hazard rating < 0.6
- Unusually high target
  Actual: (2, 4, 5) with no adverse effect
  Network: (2, 3, 4), (3, 4, 5), (2, 4, 5), and (3, 4, 6)
- Unusually high hazard area
  Actual: patient saw flashes of light at 0 mm
  Network: Hazard increased for combinations with 0mm
- Unusually modest benefit
  Actual: lesions at (1, 2, 3) helped for only a short time
  Network: *no* beneficial combination of lesions found

Comparison: Hazard

- 15 different cases reviewed
- 5 cases had hazardous outcome
- Under previous method, none of these were predicted
- Network identified 2/5 cases as hazardous
- Also identified 1/10 “safe” cases as hazardous

Comparison: Hazard

- 15 different cases reviewed
- 5 cases had hazardous outcome
- Under previous method, none of these were predicted
- Network identified 2/5 cases as hazardous
- Also identified 1/10 “safe” cases as hazardous
- ***Recognized 40% of hazards that the previous method missed***

Comparison: Efficacy

- 15 different cases reviewed
- Network identified the 1 site not lesioned because of low expected efficacy as having marginal benefit (maximum = 1.86)
- Network identified 1 site which was lesioned, producing no benefit, as having *no combination of standard lesions* which could produce any benefit.
- Two sites which produced good results when lesioned were rejected by network

Comparison: Efficacy

- 15 different cases reviewed
- Network identified the 1 site not lesioned because of low expected efficacy as having marginal benefit (maximum = 1.86)
- Network identified 1 site which was lesioned, producing no benefit, as having *no combination of standard lesions* which could produce any benefit.
- Two sites which produced good results when lesioned were rejected by network
- ***Correctly identified 100% of low-benefit sites***

Conclusions

- Neural networks trained with data obtained by a variety of common analysis methods produce more accurate assessments of surgical outcome than do current power-based techniques.
- Networks trained with data derived from wavelet analysis, entropy, and fractal dimension give more accurate results than those which use Fourier analysis, statistical moments, or power content.

A DATABASE IMAGE MANAGEMENT SYSTEM WITH AUTOMATED CLASSIFICATION OF RETINAL ABNORMALITIES

Goals:

- **Digital Image storage/retrieval**
- **Image Processing**
- **Classification of retinal diseases**

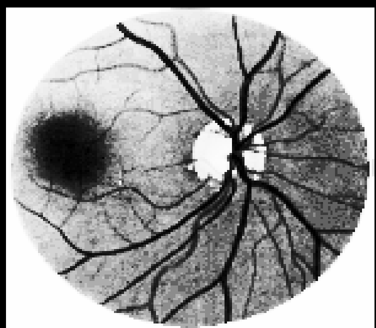
Involved difficulties:

Image Storage

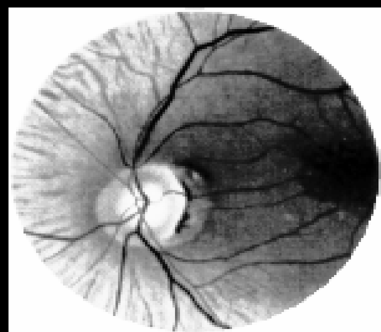
- *Variable data sizes*
- *Multiple data types*
- *Requirement to store different data types in the same file*
- *Reliability of data storage/retrieval*
- *Remote data access*
- *Data compression*

Image Classification

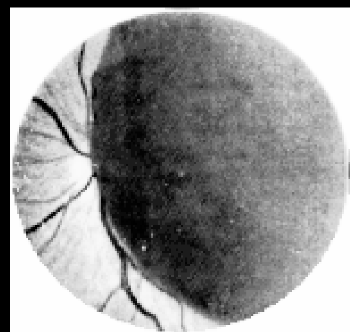
- Variety of diseases
- Different image sources, image qualities, and spatial image characteristics



Normal



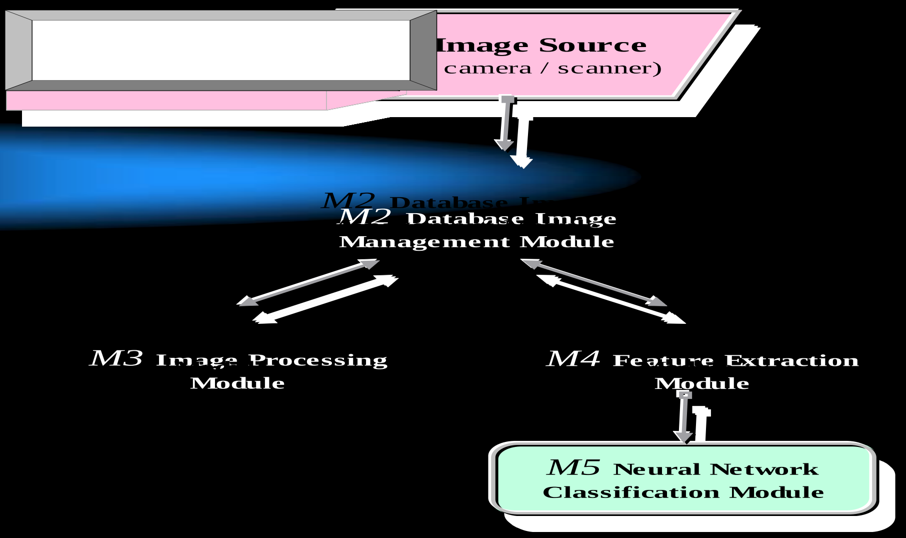
Arteriosclerosis



Hemorrhage

Evangelia Micheli-Tzanakou, PhD

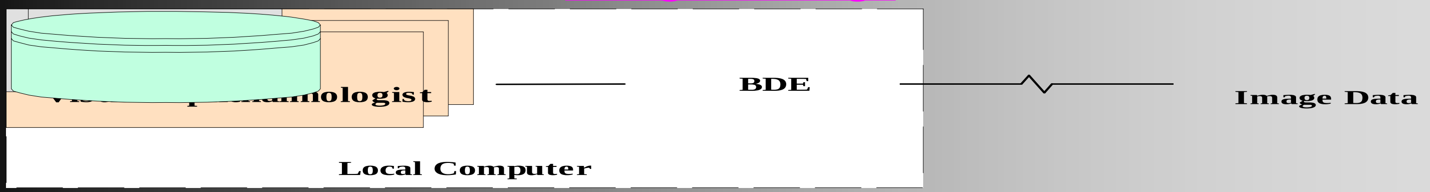
System Components



M1: Image Source

- Scanner (150dpi)
- Digital Camera

M2: Image Storage



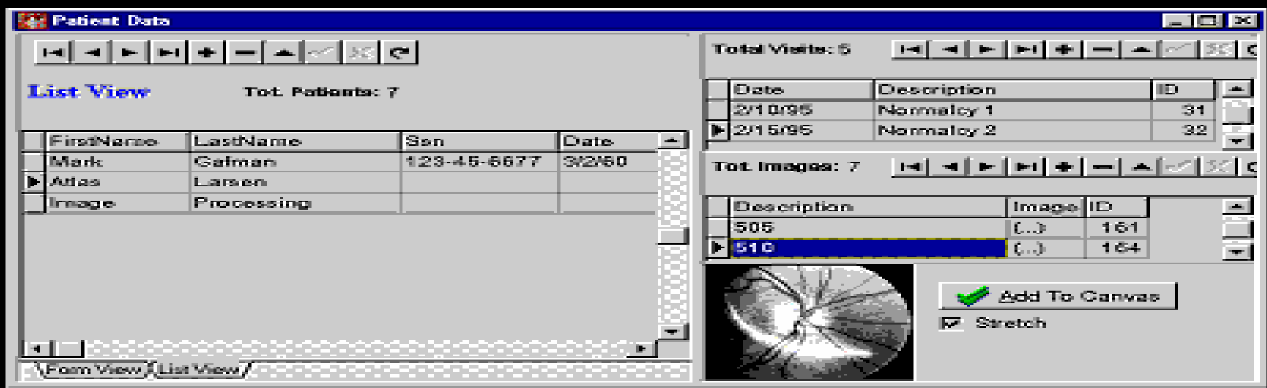
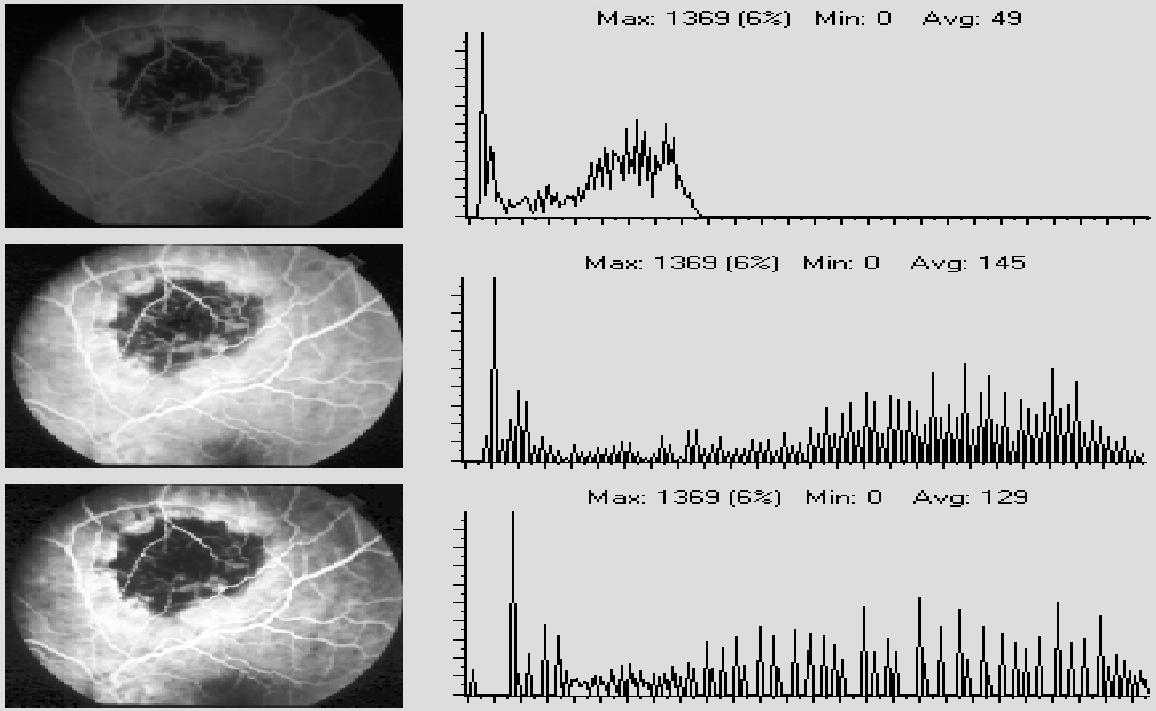


Image Processing

1. Image Histogram functions
    1.1. Histogram Equalization
    1.2. Histogram Stretch
2. Image compression/decompression based on a Gaussian Pyramid
3. Image orientation, and center of mass
4. Image clustering
5. Determination of the best fit ellipse and rectangle based on a given histogram range
6. A set of convolution filters, which include
    6.1. Low-pass, high-pass filters
    6.2. Gaussian and Laplassian filters
    6.3. Median Filters
    6.4. Several other filters with predefined kernels
    6.5. Ability to specify custom filter kernels

Histogram Functions

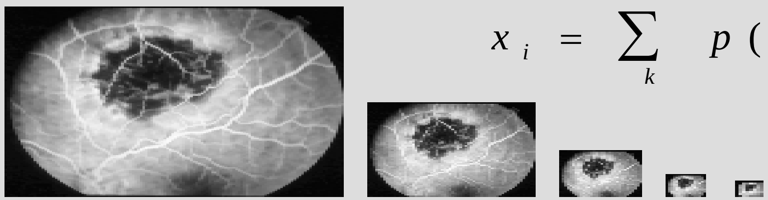


Original

Stretched

Equalized

Image Gaussian Pyramid Compression



$$x_i = \sum_k p(k - 2n)\, x_{i-1}(k)$$

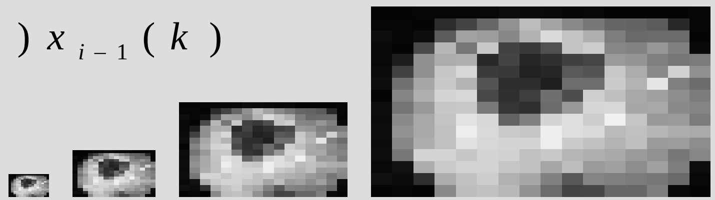


Image Clustering

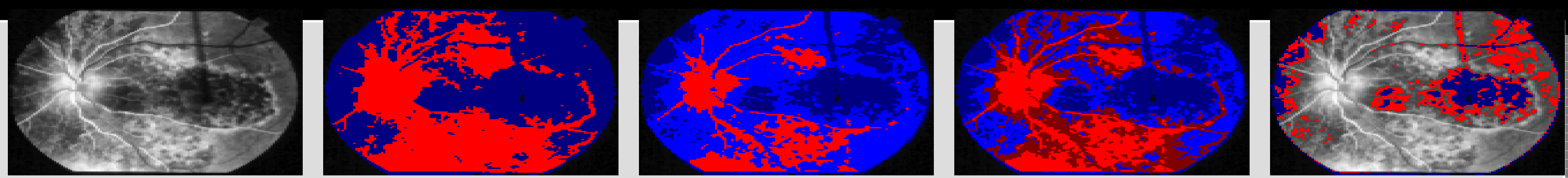


Image Orientation, Best-fit ellipse, Center of Mass, Bounding Rectangle

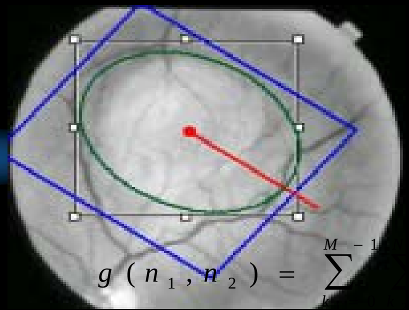
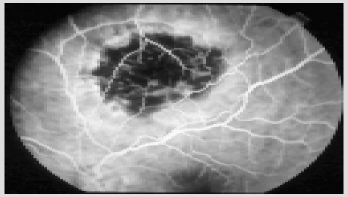
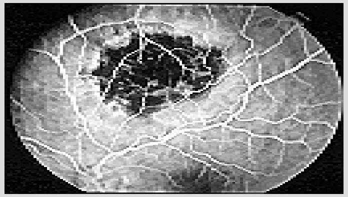
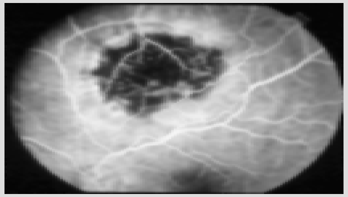


Image Filters





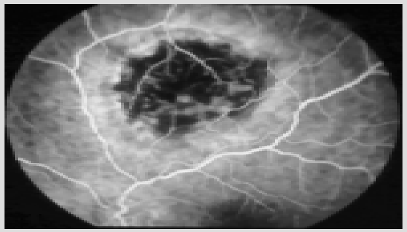
$$\begin{matrix} 0 & -1 & 0 \\ -1 & 5 & -1 \\ 0 & -1 & 0 \end{matrix}$$



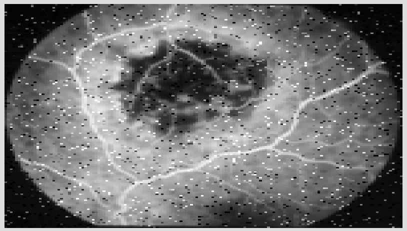
$$\begin{matrix} 1 & 1 & 1 \\ 1 & 4 & 1 \\ 1 & 1 & 1 \end{matrix}$$

$$\begin{matrix} -1 & -1 & -1 \\ -1 & 8 & -1 \\ -1 & -1 & -1 \end{matrix}$$

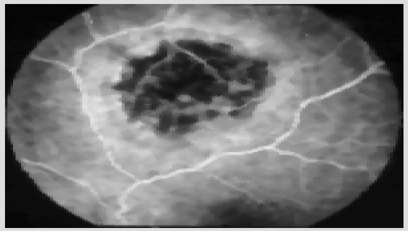
Median Filter



Original



Added noise



Filtered

Feature Extraction Methods

- **Central and Invariant Moments**
- **F-Core**
- **Wavelet Histogram**

$$\mu_{p,q} = \int\int (x - \overline{x})^p (y - \overline{y})^q f(x,y)\,dx\,dy$$

$$\phi_1 = \mu_{2,0} + \mu_{0,2}$$

$$\phi_2 = (\mu_{2,0} - \mu_{0,2})^2 + 4\mu_{1,1}^2$$

$$\phi_3 = (\mu_{3,0} - 3\mu_{1,2})^2 + (\mu_{0,3} - 3\mu_{2,1})^2$$

$$\phi_4 = (\mu_{3,0} + \mu_{1,2})^2 + (\mu_{0,3} + \mu_{2,1})^2$$

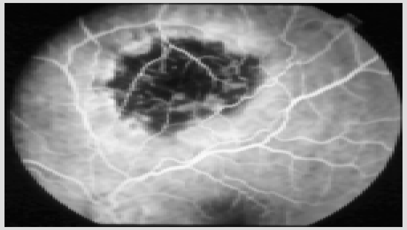
$$\phi_5 = (\mu_{3,0} - 3\mu_{1,2})(\mu_{0,3} + \mu_{1,2})[(\mu_{3,0} + \mu_{1,2})^2 - (\mu_{2,1} + \mu_{0,3})^2] + (\mu_{0,3} - 3\mu_{2,1})(\mu_{0,3} + \mu_{2,1})[(\mu_{0,3} + \mu_{2,1})^2 - (\mu_{1,2} + \mu_{3,0})^2]$$

$$\phi_6 = (\mu_{2,0} - \mu_{0,2})[(\mu_{3,0} + \mu_{1,2})^2 - (\mu_{2,1} + \mu_{0,3})^2] + 4\mu_{1,1}(\mu_{0,3} + \mu_{1,2})(\mu_{0,3} + \mu_{2,1})$$

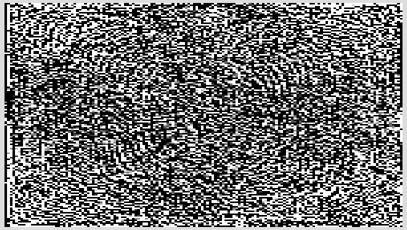
$$\phi_7 = (3\mu_{2,1} - \mu_{0,3})(\mu_{3,0} + \mu_{1,2})[(\mu_{3,0} + \mu_{1,2})^2 - 3(\mu_{2,1} + \mu_{0,3})^2] + (\mu_{3,0} - 3\mu_{2,1})(\mu_{0,3} + \mu_{2,1})[(\mu_{0,3} + \mu_{2,1})^2 - 3(\mu_{1,2} + \mu_{3,0})^2]$$

Fourier Transform

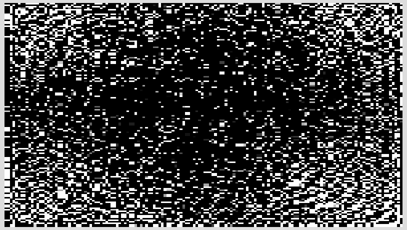
$$F(u,v) = \frac{1}{NM}\sum_{x=0}^{N-1}\sum_{y=0}^{M-1} f(x,y)\, e^{-j2\pi\left(\frac{u}{N}+\frac{v}{M}\right)}$$



Original



Real Coefficients



Imaginary Coeff.

64 x 64 pixels image => 2 x 4096 coeff.

Image power spectrum:

Micheli-Tzanakou and Binge, 1989: F-Core algorithm

Variation of the F-Core algorithm:

1. Compress image using Gaussian Pyramid to 32x32 pixels.
2. Apply the FFT (2x1024 coefficients).
3. Compute the power spectrum (1024 coefficients).
4. Sort coefficients, and store the top 5% (50 coefficients).
5. Save every other feature of the resulting 50 coefficients array.

Wavelets

$$\phi_{s,\tau}(x) = 2^{-\frac{s}{2}}\phi(2^{-s}x - \tau)$$

Scaling function:

$$W(x) = \sum_{k=0}^{N-1}(-1)^k c_{k+1}\phi(2x+k)$$

Haar scaling function:

$$\Phi(t) = 1, \quad \text{if } \ 0 \le t \le 1$$
$$\qquad 0, \quad \text{otherwise}$$

The scalar: $\frac{1}{\sqrt{2}}$ Filter: $\{\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\}$

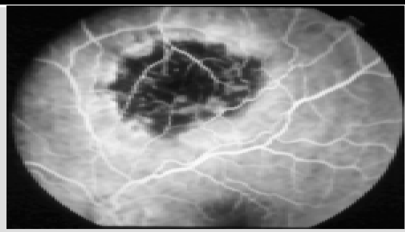
Δ_1 - wavelet coefficient matrix

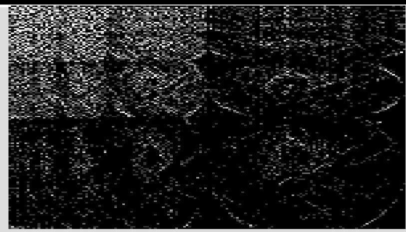
Δ_2 - scaling

$$\begin{bmatrix} y_1\\ y_2\\ y_3\\ y_4\\ y_5\\ y_6\\ y_7\\ y_8\\ y_9\\ y_{10}\\ y_{11}\\ y_{12}\\ y_{13}\\ y_{14}\\ y_{15}\\ y_{16} \end{bmatrix} \xrightarrow{\Delta_1} \begin{bmatrix} s_1\\ d_1\\ s_2\\ d_2\\ s_3\\ d_3\\ s_4\\ d_4\\ s_5\\ d_5\\ s_6\\ d_6\\ s_7\\ d_7\\ s_8\\ d_8 \end{bmatrix} \xrightarrow{\Delta_2} \begin{bmatrix} s_1\\ s_2\\ s_3\\ s_4\\ s_5\\ s_6\\ s_7\\ s_8\\ d_1\\ d_2\\ d_3\\ d_4\\ d_5\\ d_6\\ d_7\\ d_8 \end{bmatrix} \xrightarrow{\Delta_1} \begin{bmatrix} S_1\\ D_1\\ S_2\\ D_2\\ S_3\\ D_3\\ S_4\\ D_4\\ d_1\\ d_2\\ d_3\\ d_4\\ d_5\\ d_6\\ d_7\\ d_8 \end{bmatrix} \xrightarrow{\Delta_2} \begin{bmatrix} S_1\\ S_2\\ S_3\\ S_4\\ D_1\\ D_2\\ D_3\\ D_4\\ d_1\\ d_2\\ d_3\\ d_4\\ d_5\\ d_6\\ d_7\\ d_8 \end{bmatrix} \cdots \begin{bmatrix} J_1\\ J_2\\ R_1\\ R_2\\ D_1\\ D_2\\ D_3\\ D_4\\ d_1\\ d_2\\ d_3\\ d_4\\ d_5\\ d_6\\ d_7\\ d_8 \end{bmatrix}$$

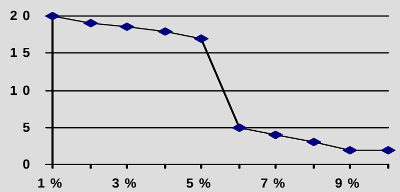
Advantages over FFT:

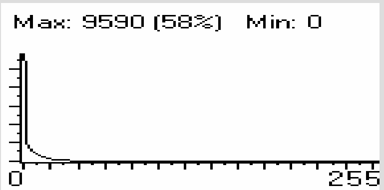
1. Can approximate functions defined in finite domains
2. Can be applied to sharp discontinuities





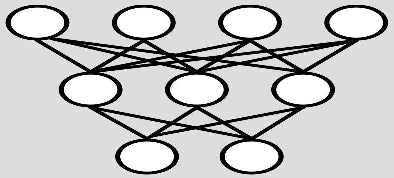
Wavelet Histogram

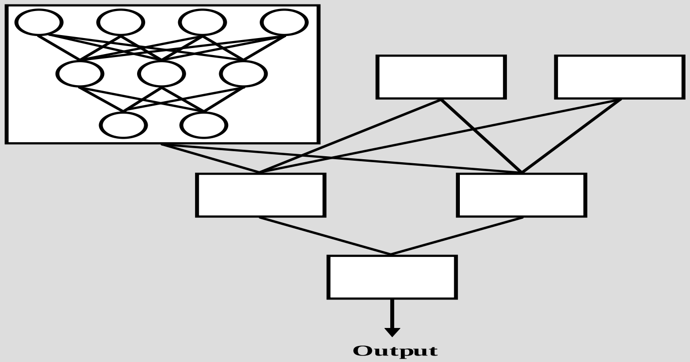




1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
Minimum values of each feature within all templates:																							
482	5	6	6	5	5	6	5	5	4	5	2	3	3	2	1	1	1	1	1	1	1	0	0
589	84	63	45	40	39	30	29	25	27	27	21	21	20	20	20	19	17	16	18	14	14	14	14

Modular Neural Networks





Template Clustering

1	0	0	0
0	1	0	0
0	0	1	0
0	0	0	1

1	0
1	0
0	1
0	1

ALOPEX optimization

Tzanakou & Harth, 1973.

$$x_i(n) = x_i(n-1) \pm \gamma \cdot \Delta x_i(n) \cdot \Delta E(n) + r_i(n)$$

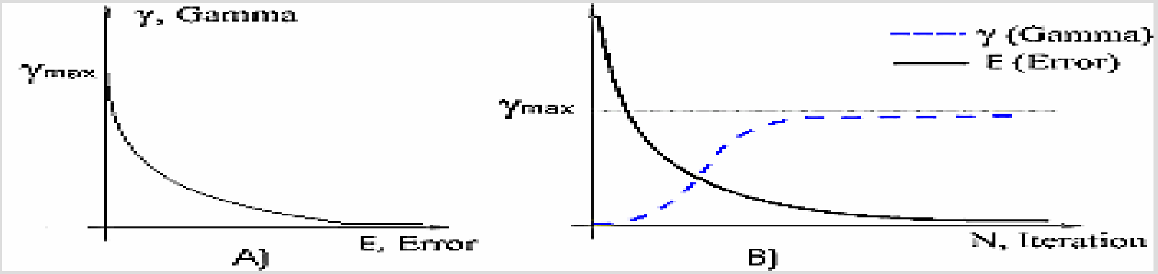
$$\Delta E(n) = E(n-1) - E(n-2)$$

$$\Delta x_i(n) = x_i(n-1) - x_i(n-2)$$

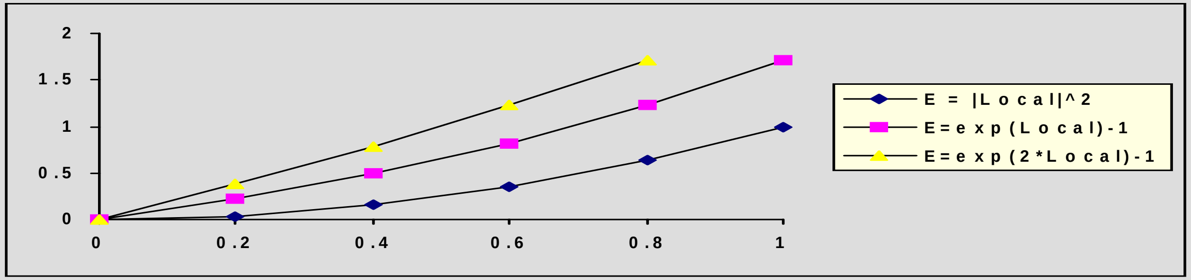
γ - Learning rate modulator

$r_i(n)$ - Gaussian noise

ΔW	E	ΔE	$-\Delta W \; \Delta E$	W (new)
> 0	↑	> 0	< 0	decreased
> 0	↓	< 0	> 0	increased
< 0	↑	> 0	< 0	decreased
< 0	↓	< 0	> 0	increased
= 0			= 0	remain unchanged



$$E_i' = \left| Out_i^{desired} - Out_i^{observed} \right|$$



Classification Criterion

Class 1

1	0	0	0	0
0	1	0	0	0
0	0	1	0	0

Class 3

0	0	1	0	0
0	1	0	0	0
0	0	0	1	0

Class 5

0	0	0	0	1
0	0	0	1	0
0	0	1	0	0

Class 2

0	1	0	0	0
1	0	0	0	0
0	0	1	0	0
0	0	0	1	0

Class 4

0	0	0	1	0
0	0	0	0	1
0	0	1	0	0
0	1	0	0	0

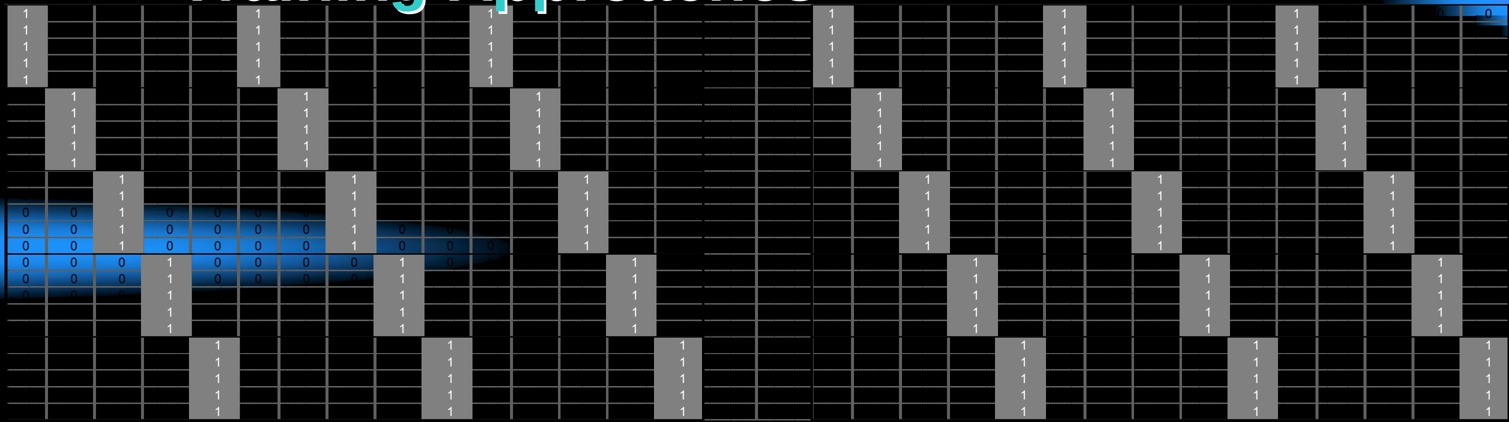
Results

Training convergence: 95%

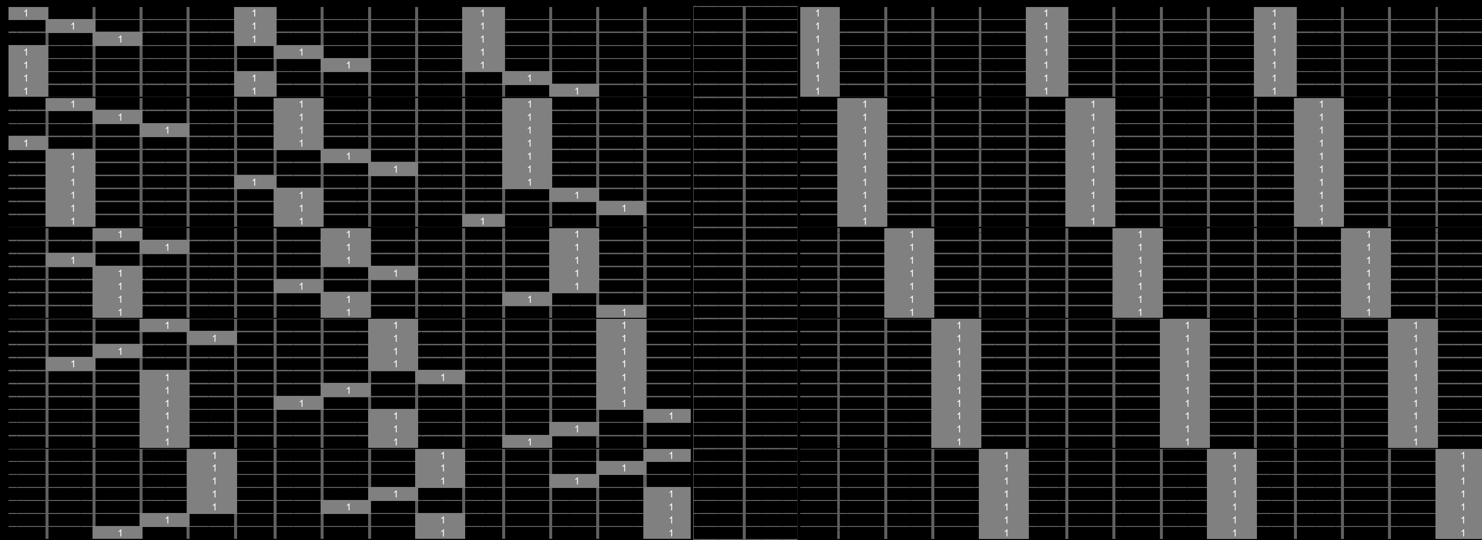
		Moments	Histog	Wave	Joint
Total	Recognized	127	145	125	127
160	Unrecognized	33	15	35	33
	Recognition Rate	79.38%	90.63%	78.13%	79.38%

Training Approaches

#1



#2



Comparison of two approaches

		First	Second	Gained	Lost
Total	Recognized	120	127	20	13
160	Unrecognized	40	33		
		75.00%	79.38%	12.50%	8.13%

Conclusions

- *The goal of uniform image storage/retrieval in a database format is achieved*
- *The image processing tools were successfully incorporated in the system*
- *The system classification of the retinal diseases proved to be satisfactory*

Future Improvement

- Using compression to minimize space that images allocate in the databases (GIF, TIFF, JPEG).
- Incorporation of additional image processing tools (more filters).
- Increase image classification accuracy by applying additional feature extraction methods, and enhancing existing methods.
- Improving ALOPEX training parameters to achieve faster convergence.

References

Bergen, J.R. (1991). Theories of visual texture perception. In Regan, D., editor, *Spatial Vision*, volume 10 of *Vision* and *Visual Dysfunction*, Macmillan.

Bergen, J.R. and Adelson, E.H. (1986). Visual texture segmentation and early vision. *Nature*, 333 (6171): 363-364.

Bergen, J.R. and Landy, M.S. (in press). Computational modeling of visual texture segregation.

Caelli, T.M. (1985). Three processing characteristics of visual texture segmentation. *Spatial Vision*, 1(1):19-30

Julesz, B. and Bergen, J.R. (1983). Textons, the fundamental elements in pre-attentive vision and perception of textures. *Bell Syst. Tech. J.* , 62, 1619-1645

Sutter, A., Beck, J., and Graham, N. (1989(. Contrast and spatial variable in texture segregation: testing a simple spatial frequency channels model. *Perception and Psychophysics*, 46(4):312-332.

Treisman, A.M. and Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology*, 12, 97-136.

References

- Hu, M-K, “Visual Pattern Recognition by Moment Invariants”, IRE Trans. in Information Theory, Feb.1962, IT-8:179-187
- Hjorth, B. EEG analysis based on time domain properties. *Electroencephalography and Clinical Neurophysiology*, Vol. 29, pp. 306-310., 1970.
- Le Mehaute, Alain , Fractal Geometries, CRC Press 1990
- Wavelet Applications in Signal and Image Processing IV, 6-9 August, 1996, SPIE Proceedings
- Richard Voss – Florida Atlantic University http://www.math.fau.edu/voss/fdds.gif or fdmr.gif
- David G. Green “Fractals and Scale” http://life.csu.edu.au/complex/tutorials/tutorial3.html
- Jean-Pierre Louvet - http://graffiti.cribx1.u-bordeaux.fr/MAPBX/louvet/history.htm
- Jean-Pierre Louvet - http://graffiti.cribx1.u-bordeaux.fr/MAPBX/louvet/jp11a.html

References

- Micheli-Tzanakou, E., Uyeda, E., Ray, R., Sharma, A., Ramanujan, R., Doug, J., “Comparison of Neural Network Algorithms for Face Recognition”, Simulation, v. 64, no. 1, July 1995, pp. 15-27.
- Mallat, S., “A theory of Multiresolution Signal Decomposition: the Wavelet Representation”, IEEE Trans on Pattern Analysis and Mach Intel, vol. 11, no. 7, July 1989, pp. 674-693
- Hu, M., “Visual Pattern Recognition by Moment Invariants”, IRE Transactions on Information Theory, no 8, Feb 1962, pp. 179-187

•Daubechies, I., "Orthonormal Bases of Compactly Supported Wavelets," Comm. Pure Appl. Math., Vol 41, 1988, pp. 906-966.

•Daubechies, I., “Ten Lectures on Wavelets”, SIAM, Philadelphia, PA, 1992

References

- Binge, G., Micheli-Tzanakou, E., “A Hybrid Image Reconstruction Technique”, Proc. Of the 13th Northeast Bioeng. Conf., v.I, 19-22, Phil, PA, 1987.
- Burt, Peter J. and Edward H. Adelson, "The Laplacian Pyramid as a Compact Image Code," IEEE Trans. on Communications, vol. com-31, no. 4, April 1983.

- Binge, G., Micheli-Tzanakou, E., “A Hybrid Image Reconstruction Technique”, Proc. Of the 13th Northeast Bioeng. Conf., v.I, 19-22, Phil, PA, 1987.

- Aleynikov, S., Michelli-Tzanakou, E., “Design and Implementation of Modular Neural Networks based on the ALOPEX Algorithm”, Virtual Intelligence, Proc., SPIE, v.2878, p81-93, Nov, 1996.

References

- Schepers et al, Four methods to estimate the fractal dimension from self-affine signals, IEEE engineering in medicine and biology, 1992, 0739-5175
- http://www.nlm.nih.gov/medlineplus/ency/article/003929.htm
- http://faculty.washington.edu/chudler/ap.html
- http://www2.crl.go.jp/jt/a134/yoshi/myo/
- Micheli-Tzanakou, E., Uyeda, E., Ray, R., Sharma, A., Ramanujan, R., Doug, J., "Comparison of Neural Network Algorithms for Face Recognition", Simulation, v. 64, no. 1, July 1995, pp. 15-27.

- Micheli-Tzanakou, E: Supervised and Unsupervised Pattern Recognition-Feature Extraction in Computational Intelligence, CRC Press, 2000.