

Optical Character Recognition for Handwritten Characters



National Center for Scientific Research
"Demokritos" Athens - Greece



Institute of Informatics and
Telecommunications



Computational Intelligence Laboratory
(CIL)

Giorgos Vamvakas

Outline

- ❑ Handwritten OCR systems
- ❑ CIL - Greek Handwritten Character Database
- ❑ Proposed OCR Methodology
- ❑ Experimental Results
- ❑ Experiments on Historical Documents
- ❑ Future Work

OCR Systems

□ OCR systems consist of four major stages :

- Pre-processing
- Segmentation
- Feature Extraction
- Classification
- Post-processing

Pre-processing

□ The raw data is subjected to a number of preliminary processing steps to make it usable in the descriptive stages of character analysis. Pre-processing aims to produce data that are easy for the OCR systems to operate accurately. The main objectives of pre-processing are :

- Binarization
- Noise reduction
- Stroke width normalization
- Skew correction
- Slant removal

Binarization



□ Document image binarization (thresholding) refers to the conversion of a gray-scale image into a binary image. Two categories of thresholding:

- Global, picks one threshold value for the entire document image which is often based on an estimation of the background level from the intensity histogram of the image.
- Adaptive (local), uses different values for each pixel according to the local area information

Noise Reduction - Normalization

❑ Noise reduction improves the quality of the document. Two main approaches:

- Filtering (masks)
- Morphological Operations (erosion, dilation, etc)

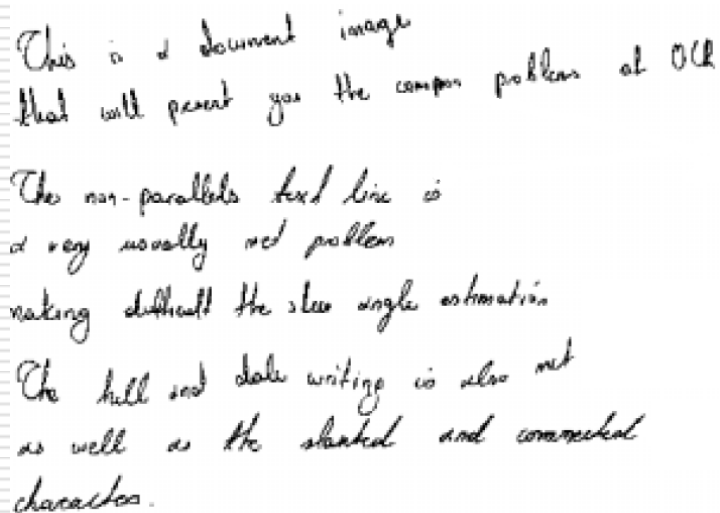


❑ Normalization provides a tremendous reduction in data size, thinning extracts the shape information of the characters.



Skew Correction

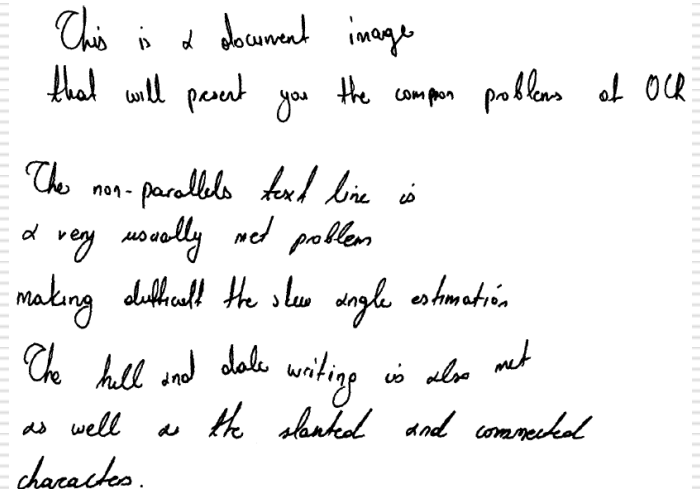
□ Skew Correction methods are used to align the paper document with the coordinate system of the scanner. Main approaches for skew detection include **correlation**, **projection profiles**, **Hough transform**.



This is a document image
that will present you the common problems of OCR

The non-parallel text line is
a very usually not problem
making difficult the skew angle estimation

The full and date writing is also not
as well as the slanted and connected
characters.



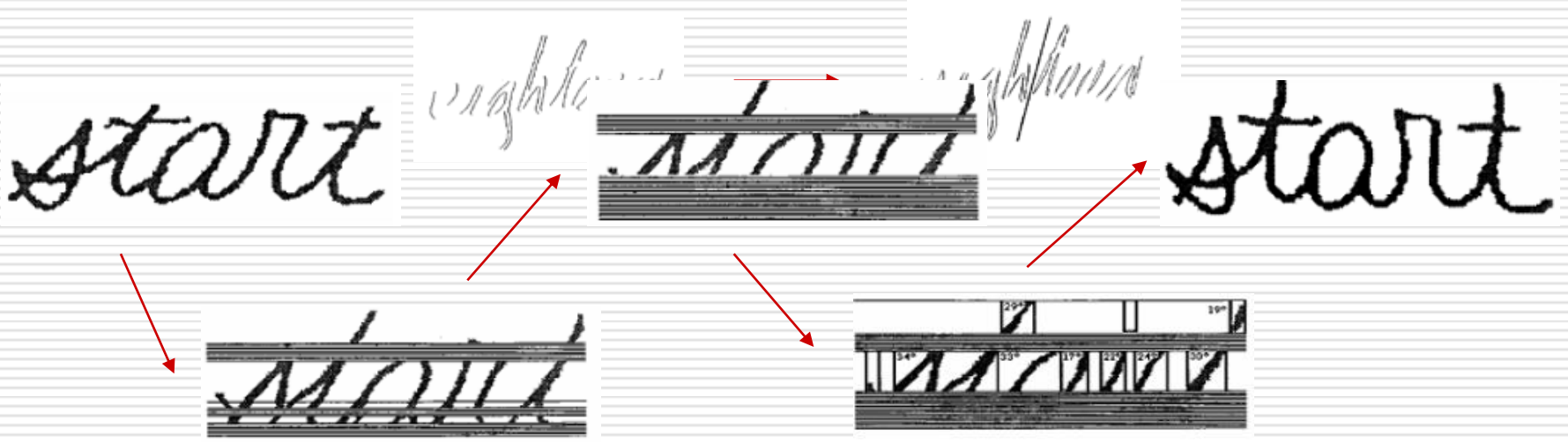
This is a document image
that will present you the common problems of OCR

The non-parallel text line is
a very usually not problem
making difficult the skew angle estimation

The full and date writing is also not
as well as the slanted and connected
characters.

Slant Removal

- ❑ The slant of handwritten texts varies from user to user. Slant removal methods are used to normalize the all characters to a standard form.
- ❑ Popular deslanting techniques are:
 - Bozinovic – Shrihari Method (BSM).
 - Calculation of the average angle of near-vertical elements



Slant Removal

□ Entropy

- The dominant slope of the character is found from the slope corrected characters which gives the minimum entropy of a vertical projection histogram. The vertical histogram projection is calculated for a range of angles $\pm R$. In our case $R=60$, seems to cover all writing styles. The slope of the character, a_m , is found

from:

$$\alpha_m = \min_{a \in \pm R} H \quad H = - \sum_{i=1}^N p_i \log p_i$$

- The character is then corrected by a_m using:

$$x' = x - y \tan(a_m) \quad y' = y$$

- ## Word Extraction 2 (RLSA)

and all hearts merry, but none
more glad than ours.

This is 2nd document image

action 2 (RLSA)

that will present you the common problems of OCR



This is a document image

that will present you the common problems of OCR

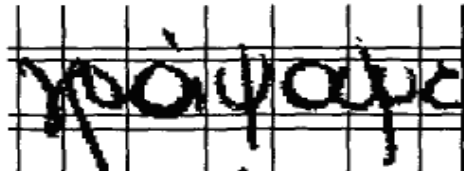
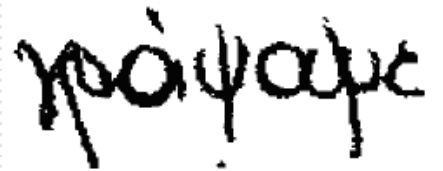
The hull and date writings is also not

as well as the slanted and connected characters.

Segmentation

□ Explicit Segmentation

□ In explicit approaches one tries to identify the smallest possible word segments (primitive segments) that may be smaller than letters, but surely cannot be segmented further. In implicit approaches the words are recognized entirely without segmenting them into letters. This is most effective later in the recognition process these primitive segments are assembled into letters based on input from the character recognizer. The advantage of the first strategy is that it is robust and quite straightforward, but is not very flexible.



Feature Extraction

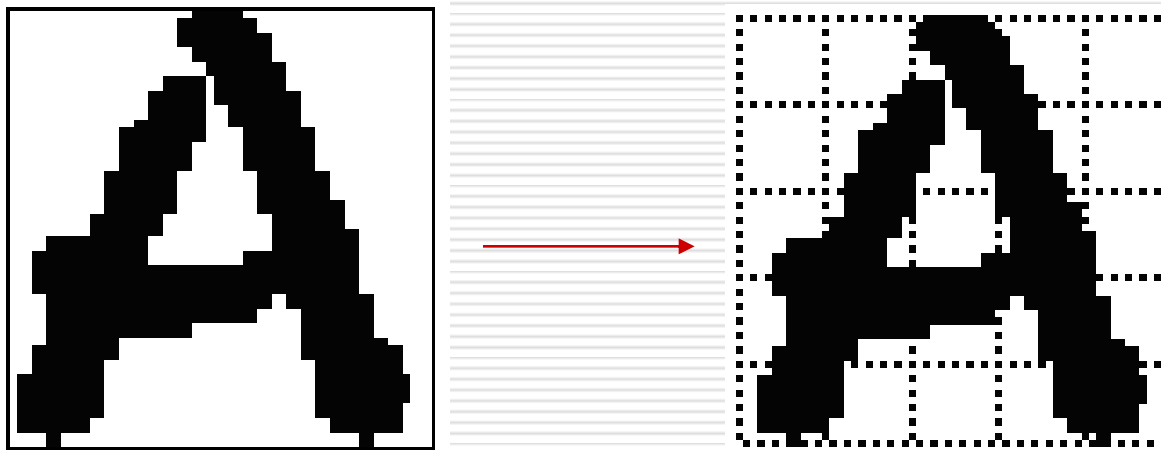
- ❑ In feature extraction stage each character is represented as a feature vector, which becomes its identity. The major goal of feature extraction is to extract a set of features, which maximizes the recognition rate with the least amount of elements.
- ❑ Due to the nature of handwriting with its high degree of variability and imprecision obtaining these features, is a difficult task. Feature extraction methods are based on 3 types of features:
 - Statistical
 - Structural
 - Global transformations and moments

Statistical Features

- ❑ Representation of a character image by statistical distribution of points takes care of style variations to some extent.
- ❑ The major statistical features used for character representation are:
 - Zoning
 - Projections and profiles
 - Crossings and distances

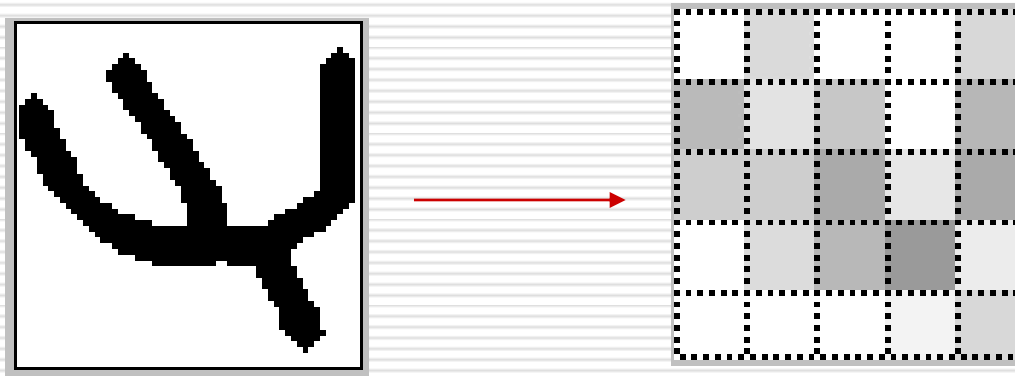
Zoning

- The character image is divided into $N \times M$ zones. From each zone features are extracted to form the feature vector. The goal of zoning is to obtain the local characteristics instead of global characteristics



Zoning – Density Features

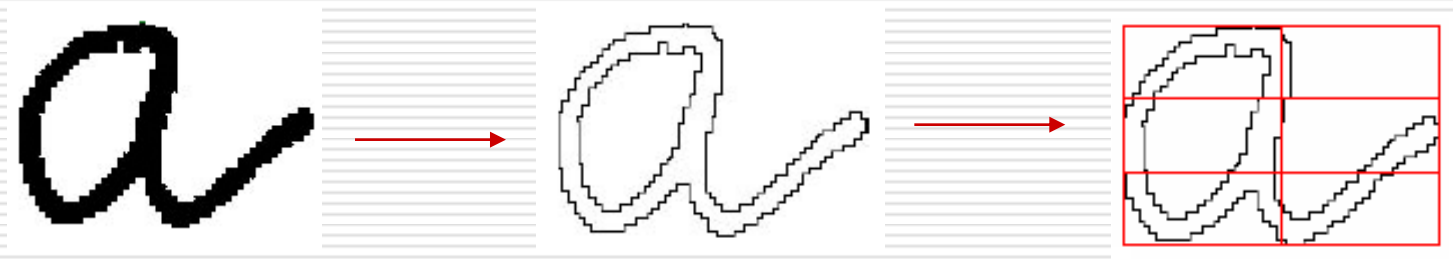
- The number of foreground pixels, or the normalized number of foreground pixels, in each cell is considered a feature.



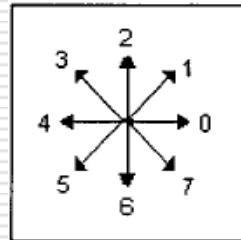
Darker squares indicate higher density of zone pixels.

Zoning – Direction Features

- Based on the contour of the character image



- For each zone the contour is followed and a directional histogram is obtained by analyzing the adjacent pixels in a 3x3 neighborhood

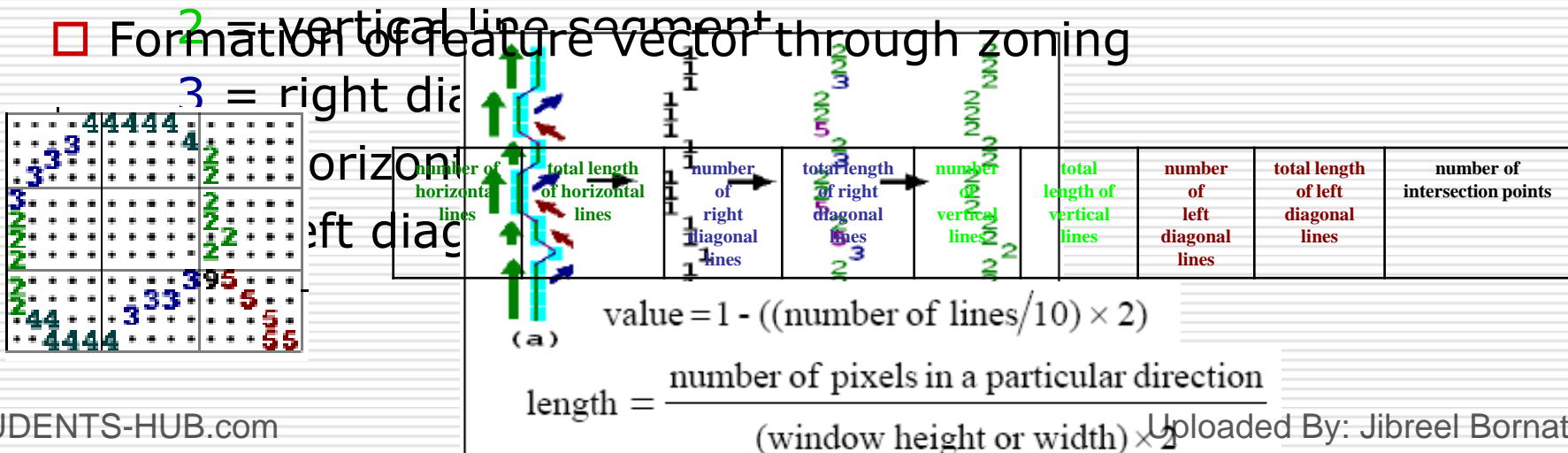


Zoning – Direction Features

- Based on the skeleton of the character image

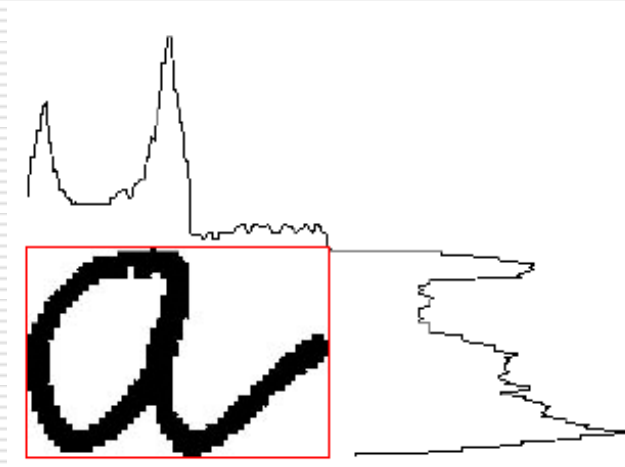


- Distinguish individual line segments
- Labeling line segment information
- Line type segmentalization coded with a direction number
- Formation of feature vector through zoning



Projection Histograms

- ❑ The basic idea behind using projections is that character images, which are 2-D signals, can be represented as 1-D signal. These features, although independent to noise and deformation, depend on rotation.
- ❑ Projection histograms count the number of pixels in each column and row of a character image. Projection histograms can separate characters such as "m" and "n".



Profiles

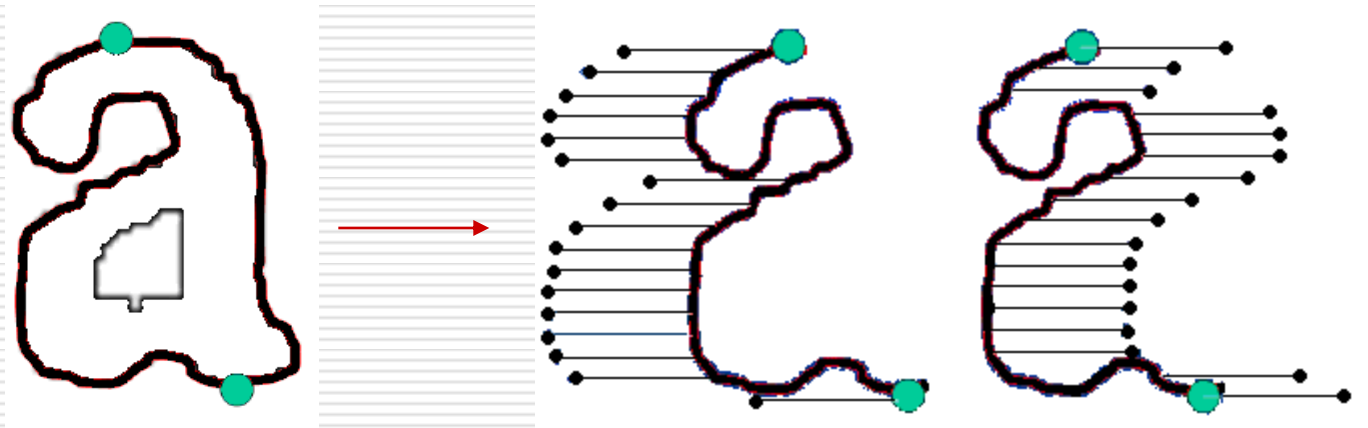
□ The profile counts the number of pixels (distance) between the bounding box of the character image and the edge of the character. The profiles describe well the external shapes of characters and allow to distinguish between a great number of letters, such as “p” and “q”.



Profiles

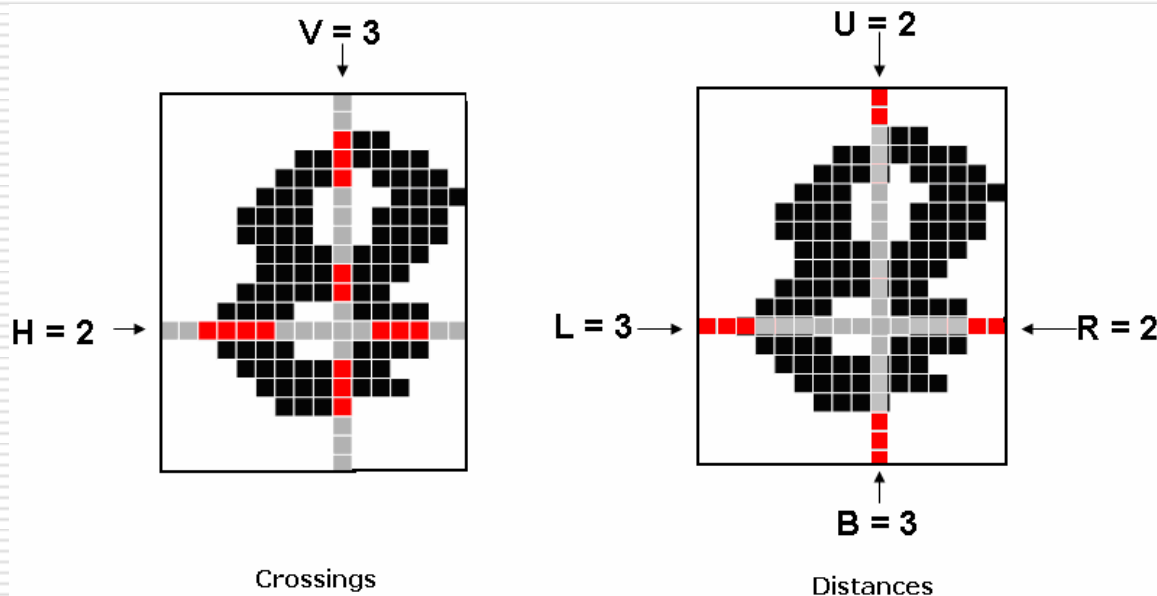
□ Profiles can also be used to the contour of the character image

- Extract the contour of the character
- Locate the uppermost and the lowermost points of the contour
- Calculate the in and out profiles of the contour



Crossings and Distances

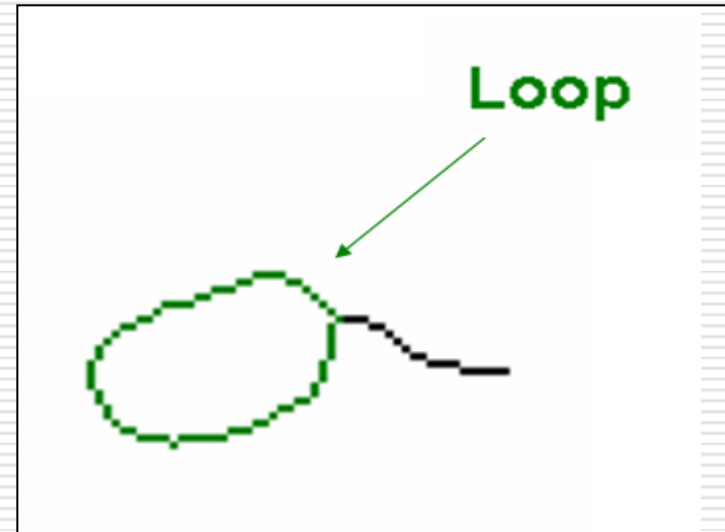
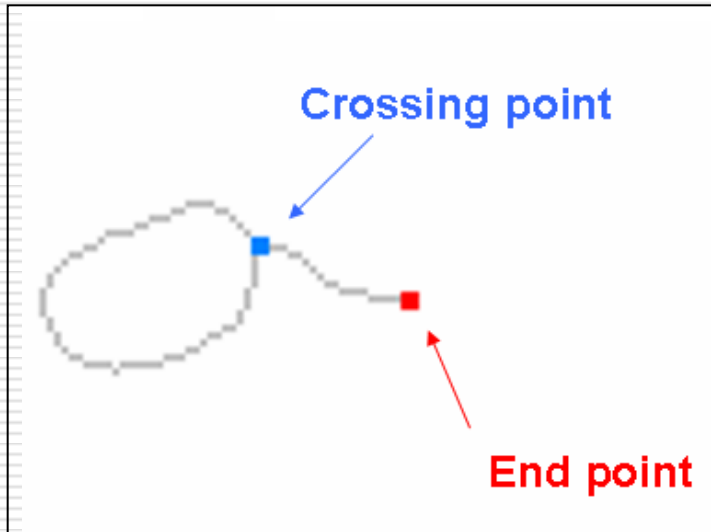
□ **Crossings** count the number of transitions from background to foreground pixels along vertical and horizontal lines through the character image and **Distances** calculate the distances of the first image pixel detected from the upper and lower boundaries, of the image, along vertical lines and from the left and right boundaries along horizontal lines



Structural Features

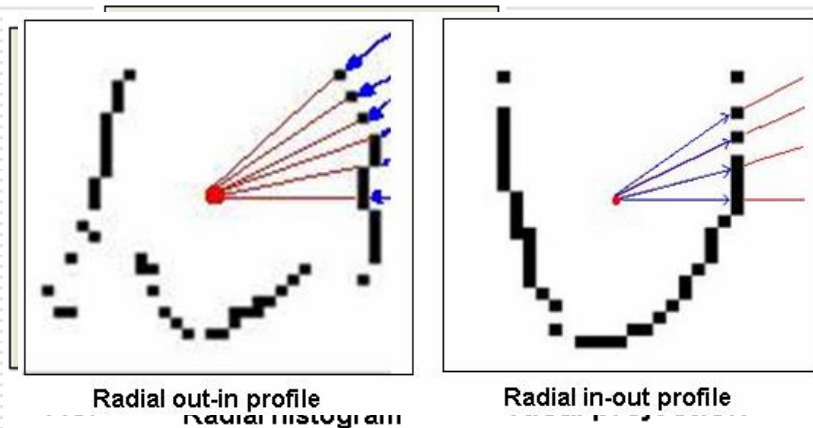
- ❑ Characters can be represented by structural features with high tolerance to distortions and style variations. This type of representation may also encode some knowledge about the structure of the object or may provide some knowledge as to what sort of components make up that object.
- ❑ Structural features are based on topological and geometrical properties of the character, such as aspect ratio, cross points, loops, branch points, strokes and their directions, inflection between two points, horizontal curves at top or bottom, etc.

Structural Features



Structural Features

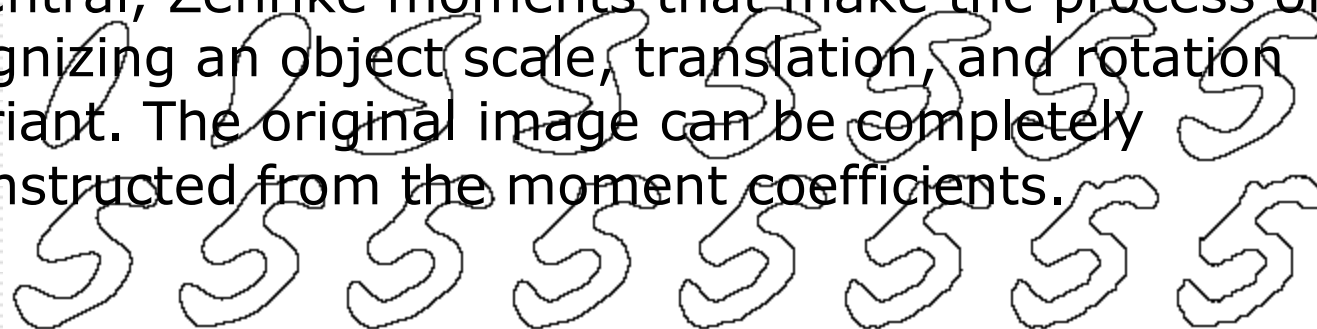
- ❑ A structural feature extraction method for recognizing Greek handwritten characters [Kavallieratou *et.al* 2002]
- ❑ Three types of features:
 - Horizontal and Vertical projection histograms
 - Radial histogram
 - Radial out-in and radial in-out profiles



Global Transformations - Moments

□ The Fourier Transform (FT) of the contour of the image is calculated. Since the first n coefficients of the FT can be used in order to reconstruct the contour, then these n coefficients are considered to be a n -dimensional feature vector that represents the character.

□ Central, Zenrike moments that make the process of recognizing an object scale, translation, and rotation invariant. The original image can be completely reconstructed from the moment coefficients.



Classification

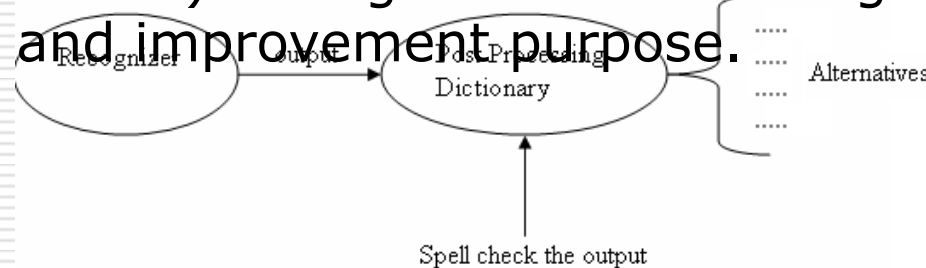
□ k-Nearest Neighbour (k-NN) , Bayes Classifier, Neural Networks (NN), Hidden Markov Models (HMM), Support Vector Machines (SVM), etc

There is no such thing as the "best classifier". The use of classifier depends on many factors, such as available training set, number of free parameters etc.

Post-processing

□ Goal : the incorporation of context and shape information in all the stages of OCR systems is necessary for meaningful improvements in recognition rates.

□ The simplest way of incorporating a context-dependent lexicon is the use of a dictionary for rules (lexicon) or mistakes (approaches) during or after the recognition stage for verification and improvement purpose.



□ Drawback : Unrecoverable OCR decisions.

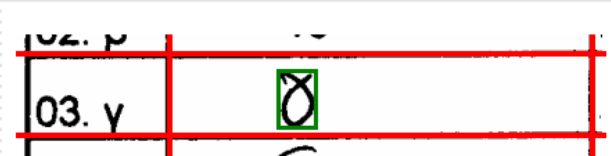
CIL- Greek Handwritten Character Database

□ Each form consists of 56 Greek handwritten characters:

- 24 upper-case
- 24 lower-case
- the final "ς"
- the accented vowels "ᾱ", "ῆ", "ῑ", "ῖ", "ῑ", "ῡ"

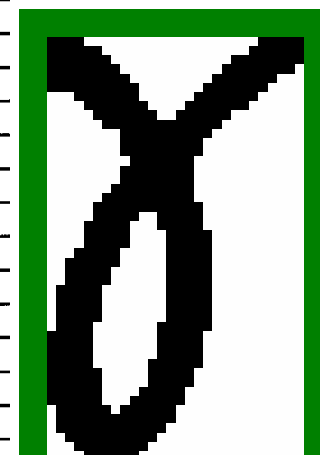
□ The steps led to the Greek handwritten character database are:

- Line detection using Run Length Smoothing Algorithm (RLSA)
- Character extraction



01. α	α	32. ω	ω
02. β	β	33. Α	Α
03. γ	γ	34. Β	Β
04. δ	δ	35. Γ	Γ
05. ε	ε	36. Δ	Δ
06. ζ	ζ	37. Ε	Ε
07. η	η	38. Ζ	Ζ
08. θ	θ	39. Η	Η

15. ο	ο	46. =	=
16. π	π	47. ο	ο
17. ρ	ρ	48. π	π
18. σ	σ		
19. τ	τ		
20. υ	υ		
21. φ	φ		
22. χ	χ		
23. ψ	ψ		
24. ω	ω		
25. ς	ς		
26. ᾱ	ᾱ		
27. ῆ	ῆ		
28. ῑ	ῑ		
29. ῖ	ῖ		
30. ῑ	ῑ		
31. ῡ	ῡ		



CIL- Greek Handwritten Character Database

□ CIL Database:

- 125 Greek writers
- 5 forms per writer
- 625 variations of each character led to an overall of 35,000 isolated and labeled Greek handwritten characters



Proposed OCR Methodology

□ Pre-processing :

- Image size normalization



- Slope correction



□ Feature Extraction

Feature Extraction

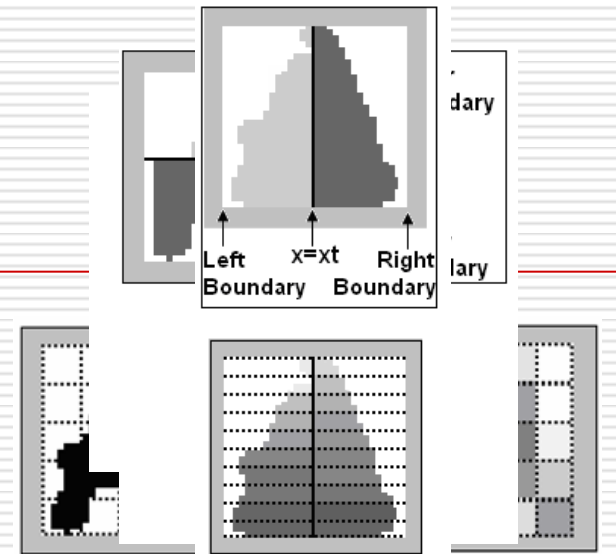
□ Two types of features :

- Features based on zones:

- ✓ The character image is divided into horizontal and vertical zones and the density of character pixels is calculated for each zone

- Features based on character projection profiles:

- ✓ The centre mass (x_t, y_t) of the image is first found
 - ✓ Upper/ lower profiles are computed by considering for each image column, the distance between the horizontal line $y = y_t$ and the closest pixel to the upper/lower boundary of the character image. This ends up in two zones depending on y_t . Then both zones are divided into vertical blocks. For all blocks formed we calculate the area of the upper/lower character profiles.
 - ✓ Similarly, we extract the features based on left/right profiles.



Experimental Results

□ The CIL Database was used

- 56 characters
- 625 variations of each character
- 35,000 isolated and labeled Greek handwritten characters

□ 10 pairs of classes were merged, due to size normalization step, resulting to a database of 28,750 characters.

	Upper-case	Lower-case
1	Ε	ε
2	Θ	θ
3	Κ	κ
4	Ο	ο
5	Π	π
6	Ρ	ρ
7	Τ	τ
8	Φ	φ
9	Χ	χ
10	Ψ	ψ

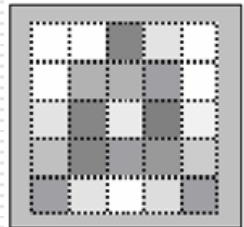
Experimental Results

- 1/5 of each class was used for testing and 4/5 for training
- Character images normalized to a 60x60 matrix

- Features

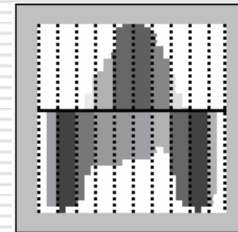
- Based on Zones

- ✓ 5 horizontal and 5 vertical zones => 25 features →



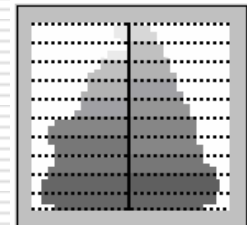
- Based on Upper and Lower profiles

- ✓ 10 vertical zones => 20 features →



- Based on Left and Right profiles

- ✓ 10 horizontal zones => 20 features



- Total Number of features
 $25 + 20 + 20 = 65$

Experimental Results

- The Greek handwritten character database was used:
 - Euclidean Minimum Distance Classifier (EMDC)
 - Support Vector Machines (SVM)

Pre-processing	Features			Number of features	Classifier		Recognition Rate (%)
	Slope Correction	Kavallieratou 2002	Hybrid		EMDC	SVM	
			Zones Projections				
		✓		280	✓		81.36%
✓		✓		280	✓		81.20%
			✓	25	✓		85.94%
				40	✓		76.80%
			✓	65	✓		83.44%
✓			✓	25	✓		85.36%
✓				40	✓		78.46%
✓			✓	65	✓		84.55%
		✓		280		✓	87.52%
✓		✓		280		✓	88.62%
			✓	25		✓	88.29%
				40		✓	87.56%
			✓	65		✓	90.12%
✓			✓	25		✓	88.48%
✓				40		✓	87.75%
✓			✓	65		✓	91.61%

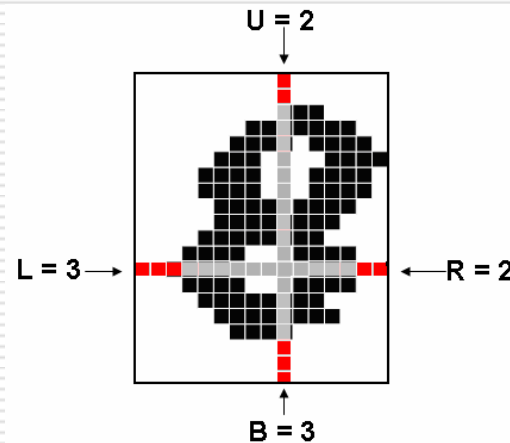
Experimental Results

□ Dimensionality Reduction

- Three types of features

- ✓ our features
- ✓ distance features
- ✓ profile features

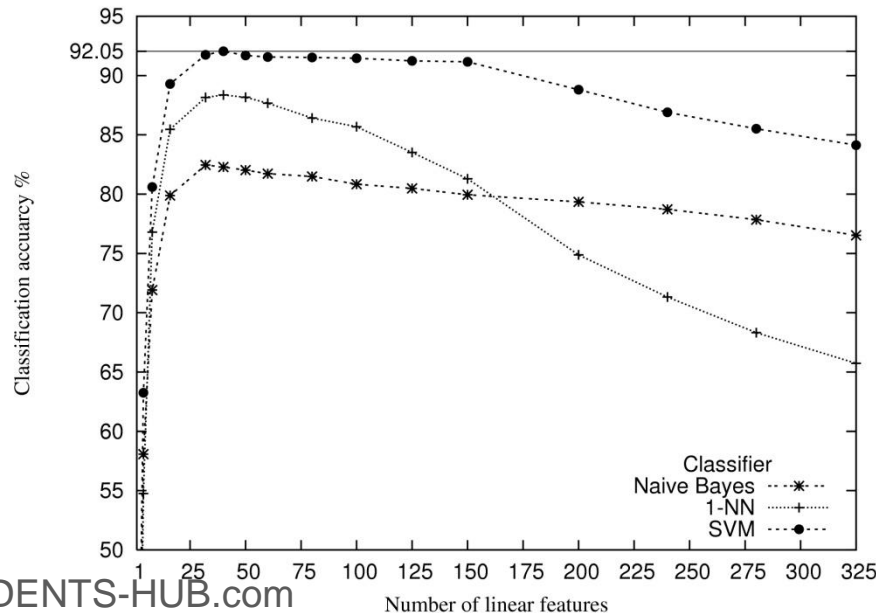
→ 325 features



Experimental Results

□ Dimensionality Reduction

Linear Discriminant Analysis (LDA) method is employed, according to which the most significant linear features are those where the samples distribution has important overall variance while the samples per class distributions have small variance




- Recognition Rate = 92.05%
- Number of features = 40

Experiments on Historical Documents

- 12 Documents
- 11,963 “characters” using connected component labelling
- Size normalization to a 60x60 matrix

e.g. 

- “Database” has 4,503 characters (*lower-case Greek handwritten characters, that is “α”, “β”, “γ”, ... , “ω” and “ς”*)

e.g. 

Publications

- G. Vamvakas, B. Gatos, I. Pratikakis, N. Stamatopoulos, A. Roniotis and S.J. Perantonis, "**Hybrid Off-Line OCR for Isolated Handwritten Greek Characters**", *The Fourth IASTED International Conference on Signal Processing, Pattern Recognition, and Applications (SPPRA 2007)*, ISBN: 978-0-88986-646-1, pp. 197-202, Innsbruck, Austria, February 2007.

- G. Vamvakas, N. Stamatopoulos, B. Gatos, I. Pratikakis and S.J. Perantonis, "**Standard Database and Methods for Handwritten Greek Character Recognition**", accepted for publication in the proc. of the 11th Panhellenic Conference on Informatics (PCI 2007), Patras, May 2007.

- "**An Efficient Feature Extraction and Dimensionality Reduction Scheme for Isolated Greek Handwritten Character Recognition**", 9th International Conference on Document Analysis and Recognition (ICDAR 2007), Curitiba, Brazil, September 2007. Waiting...

Future Work

- ❑ Creating new hierarchical classification schemes based on rules after examining the corresponding confusion matrix.
- ❑ Exploiting new features to improve the current performance.

