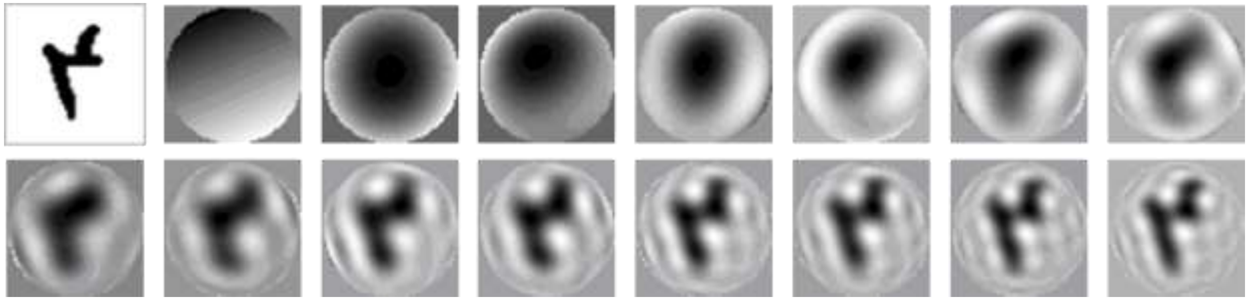
IN THE NAME OF ALLAH

Neural Networks

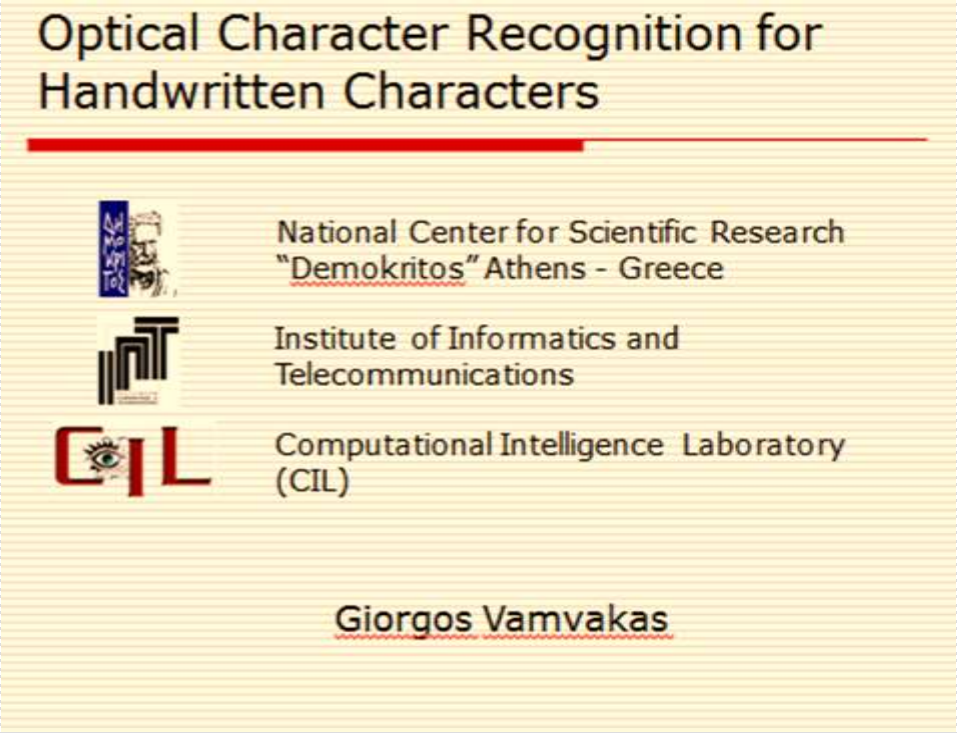
Feature Extraction: Case Study

Optical Character Recognition for Handwritten Characters



Shahrood University of Technology
Hossein Khosravi

Reference



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

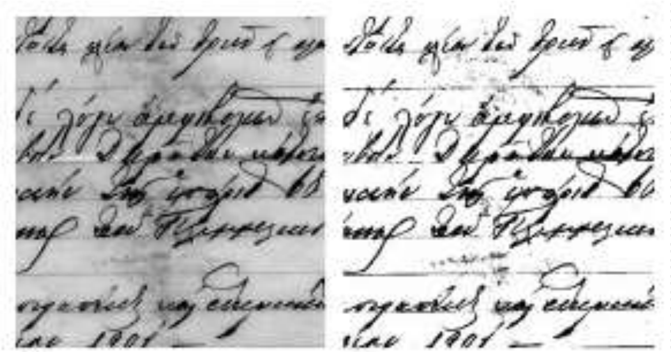
OCR Systems

- 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

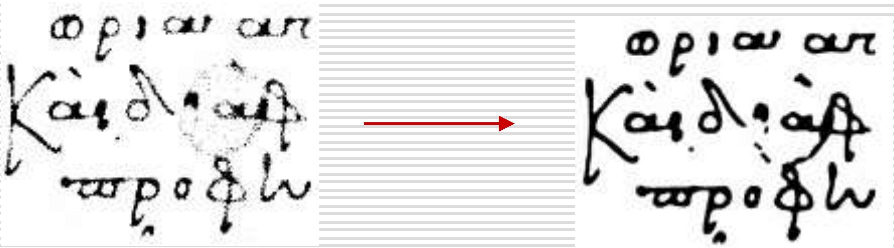


- Document image binarization (thresholding) refers to the conversion of a gray-scale image into a binary image. Two categories of thresholding:
    - Global, picks one threshold value for the entire document image which is often based on an estimation of the background level from the intensity histogram of the image.
    - Adaptive (local), uses different values for each pixel according to the local area information

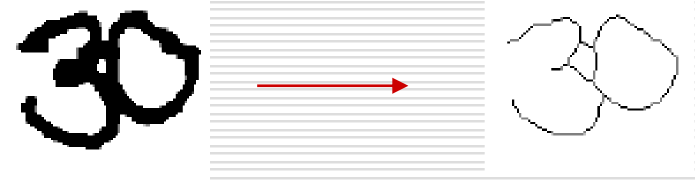
Noise Reduction - Normalization

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

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

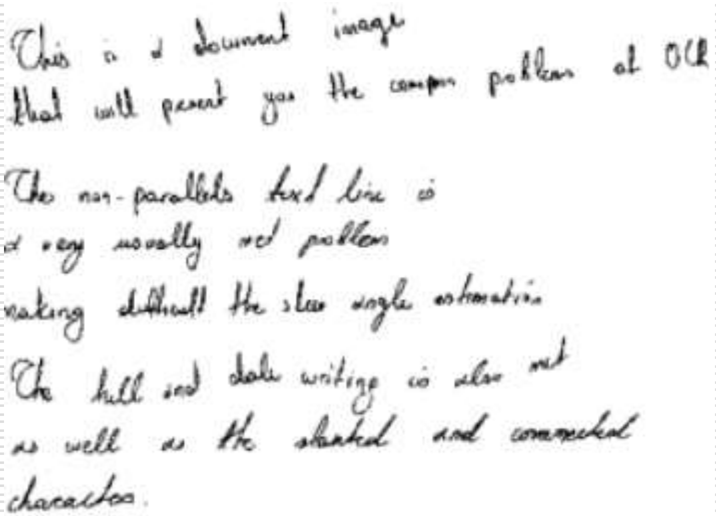


□ Normalization provides a tremendous reduction in data size, thinning extracts the shape information of the characters.



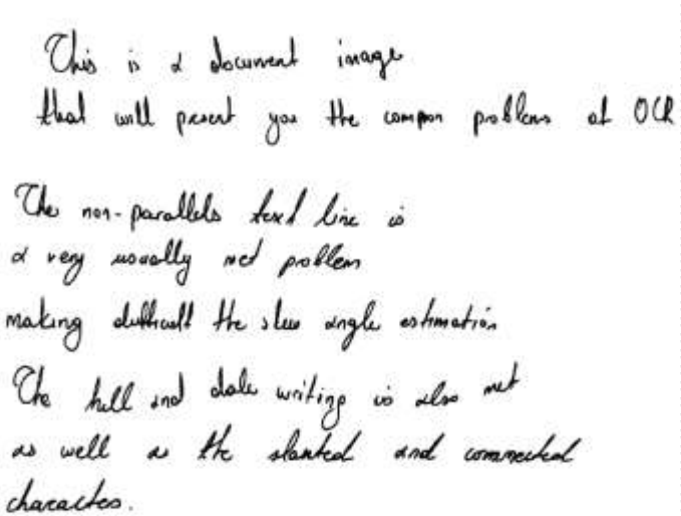
Skew Correction

□ 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-parallels text line is
a very usually met problem
making difficult the skew angle estimation
The hill and dale writing is also met
as well as the slanted and connected
characters.



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

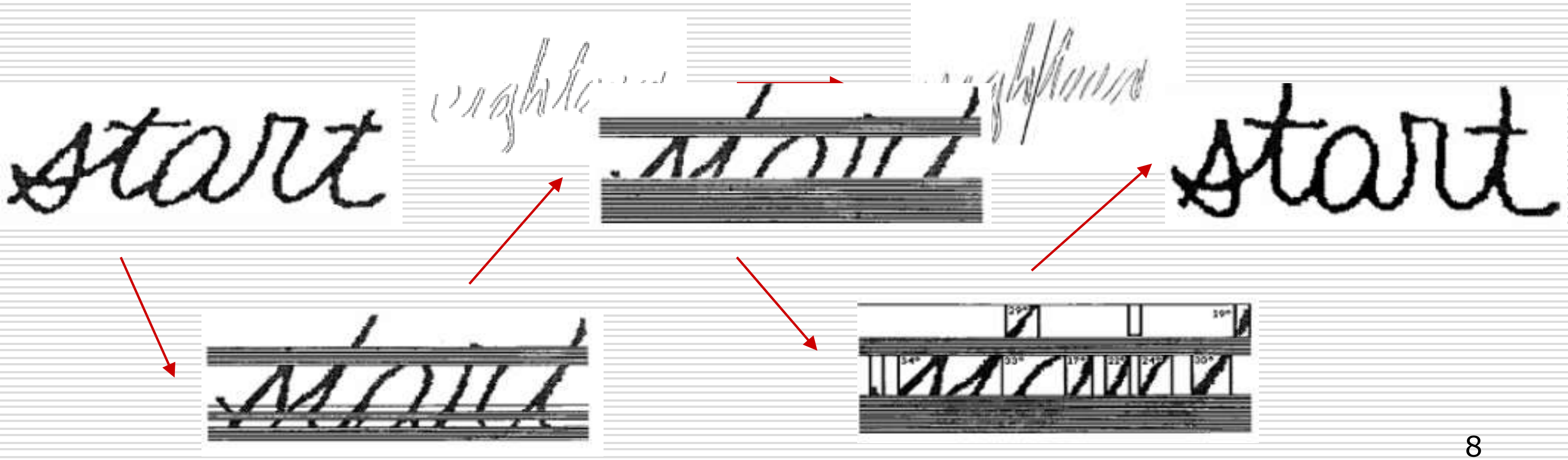
The non-parallels text line is
a very usually met problem
making difficult the skew angle estimation
The hill and dale writing is also met
as well as the slanted and connected
characters.

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.

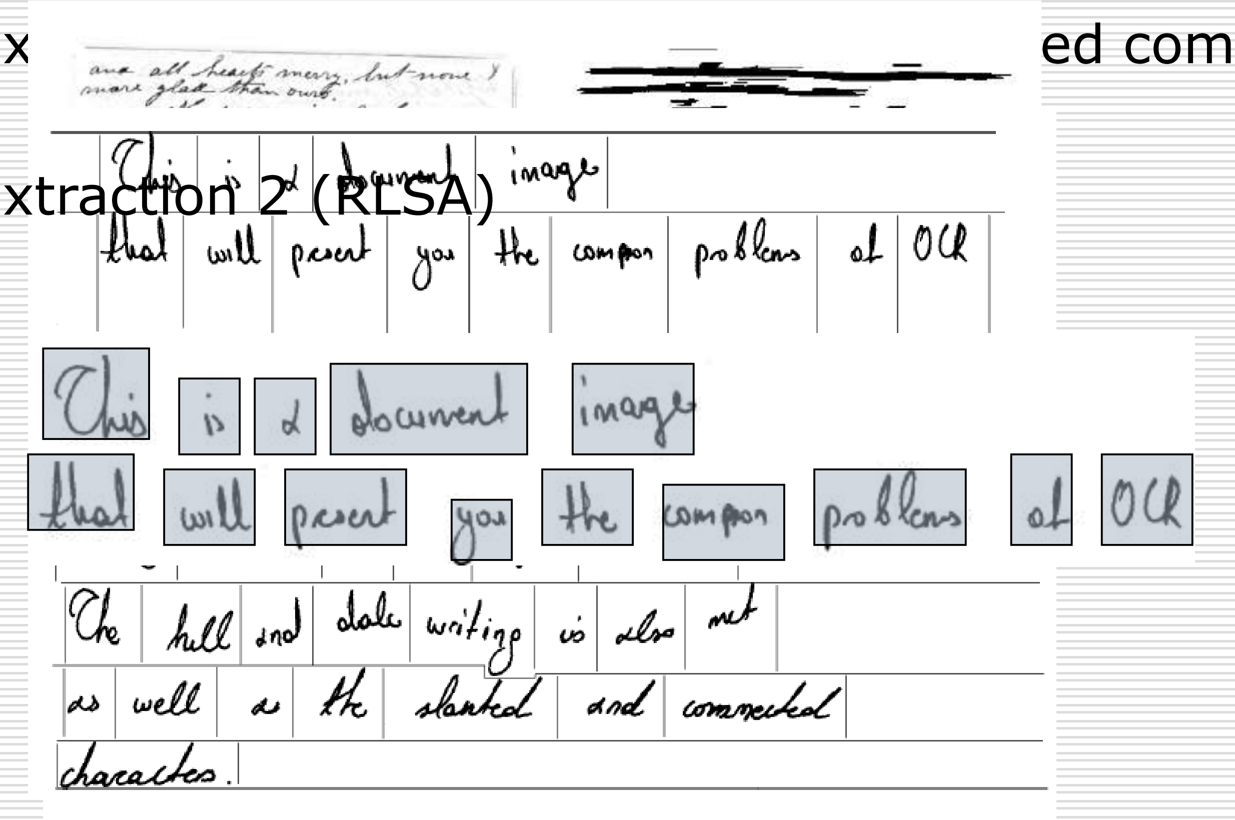
□ Popular deslanting techniques are:

- Bozinovic – Shrihari Method (BSM).
- Calculation of the average angle of near-vertical elements



Segmentation

- □ Text Line Detection (Hough Transform, projections, smearing)
- □ Word Ex ed component analysis)
- □ Word Extraction 2 (RLSA)


This is a document image
that will present you the common problems of OCR
The hill and dale writing is also met
as well as the slanted and connected
characters.

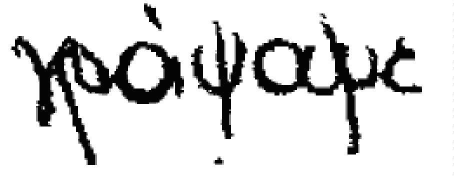
Segmentation

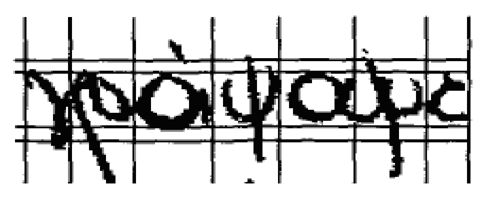
- □ Explicit Segmentation

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. 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.

- □ Implicit Segmentation

In implicit approaches the words are recognized entirely without segmenting them into letters. This is most effective and viable only when the set of possible words is small and known in advance, such as the recognition of bank checks and postal address





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

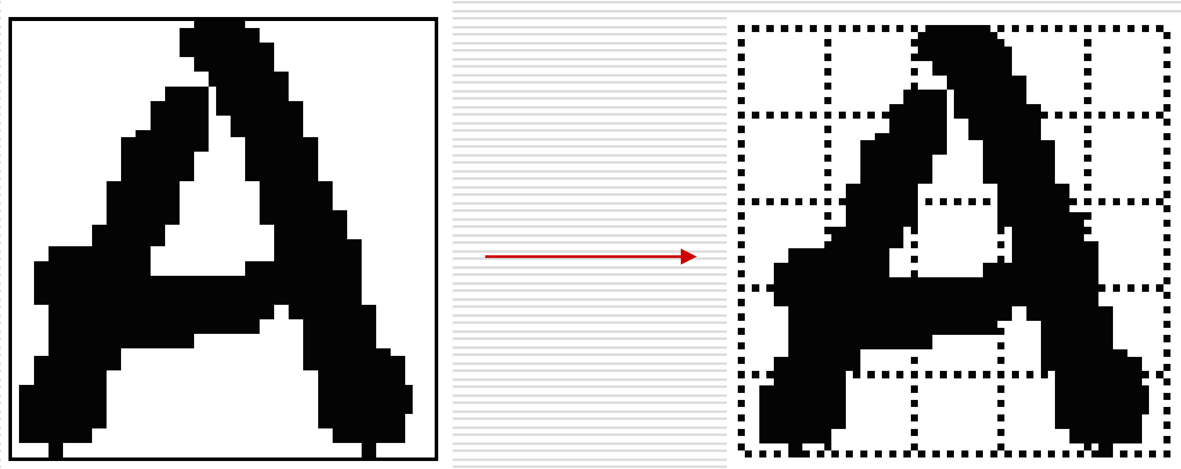
□ Representation of a character image by statistical distribution of points takes care of style variations to some extent.

□ The major statistical features used for character representation are:

- Zoning
- Projections and profiles
- Crossings and distances

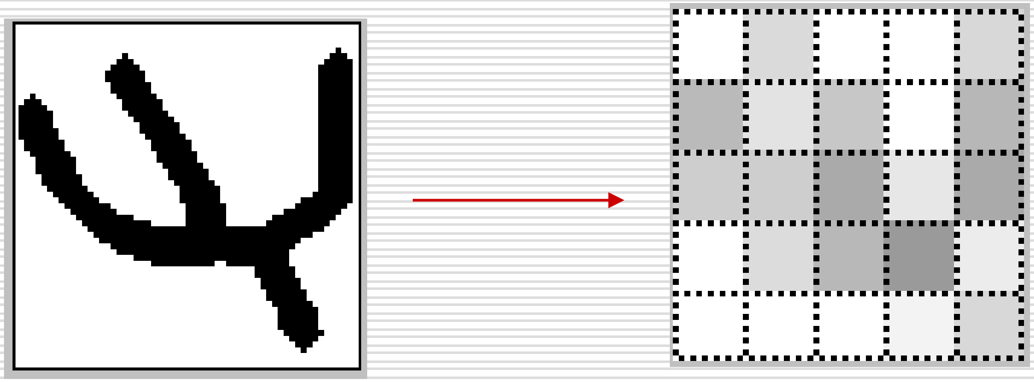
Zoning

- The character image is divided into NxM 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

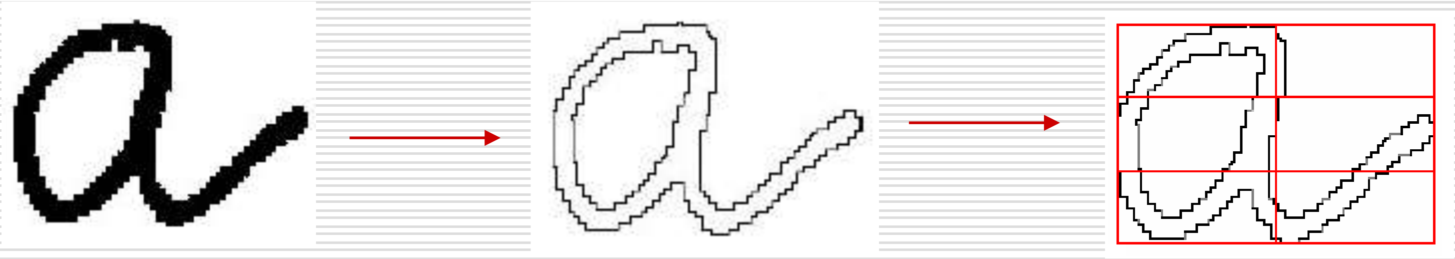
- The number of foreground pixels, or the normalized number of foreground pixels, in each cell is considered a feature.



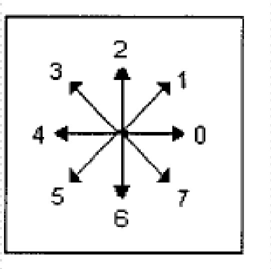
Darker squares indicate higher density of zone pixels.

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 normalization

□ Formation of feature vector through zoning

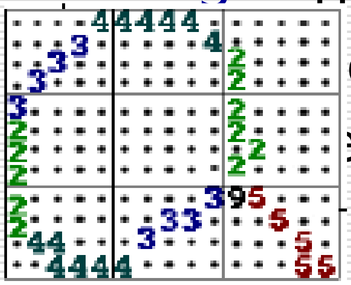
- Line segments are coded with a direction number

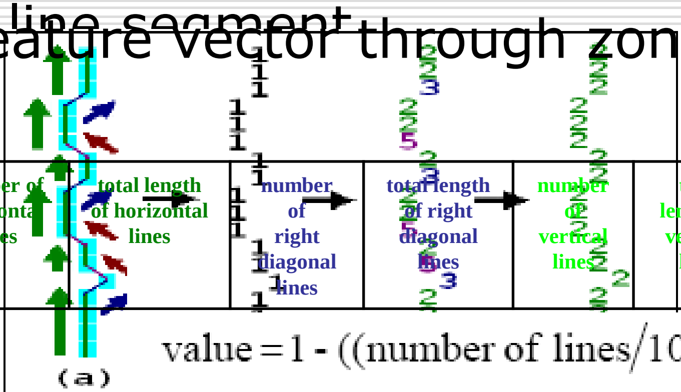
2 = vertical line segment

3 = right dia

orizont

eft diag





number of horizontal lines	total length of horizontal lines	number of right diagonal lines	total length of right diagonal lines	number of vertical lines	total length of vertical lines	number of left diagonal lines	total length of left diagonal lines	number of intersection points

$$\text{value} = 1 - ((\text{number of lines}/10) \times 2)$$

$$\text{length} = \frac{\text{number of pixels in a particular direction}}{(\text{window height or width}) \times 2}$$

Gradient Histogram

$$g(x, y) = [g_x, g_y]^T$$

$$P(x, y) = \tan^{-1}(g_y, g_x)$$

$$A(x, y) = \sqrt{g_x^{\ 2} + g_y^{\ 2}}$$

$$F_\theta = \sum_{x_\theta, y_\theta} A(x, y)$$

- Divide input image into some squares
- Compute gradient for each pixel in each block
- Quantize angles into 16 bins
- Compute 16 features for each block

Filters used in gradient histogram

1	0	-1	Sobel Operators	1	2	1
2	0	-2		0	0	0
1	0	-1		-1	-2	-1
	0	1	Roberts Operators	1	0	
	-1	0		0	-1	

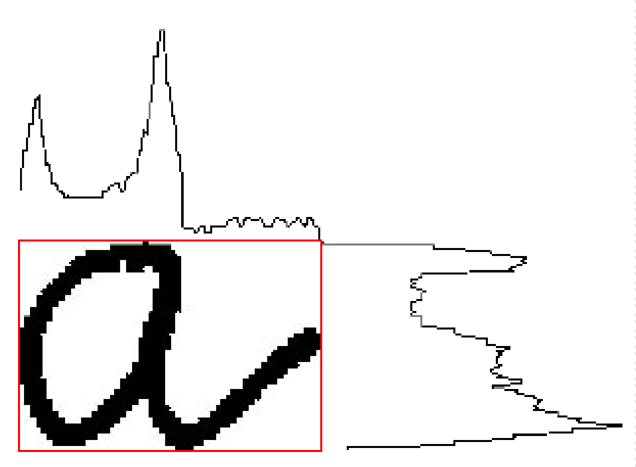
۵	۵	۵		−۳	−۳	−۳		−۳	−۳	۵		۵	−۳	−۳
−۳	۰	−۳		−۳	۰	−۳		−۳	۰	۵		۵	۰	−۳
−۳	−۳	−۳		۵	۵	۵		−۳	−۳	۵		۵	−۳	−۳
Horizontal								Vertical						
−۳	۵	۵		−۳	−۳	−۳		−۳	−۳	−۳		۵	۵	−۳
−۳	۰	۵		۵	۰	−۳		−۳	۰	۵		۵	۰	−۳
−۳	−۳	−۳		۵	۵	−۳		−۳	۵	۵		−۳	−۳	−۳
Right-Diagonal								Left-Diagonal						

Kirsch Operators

Projection Histograms

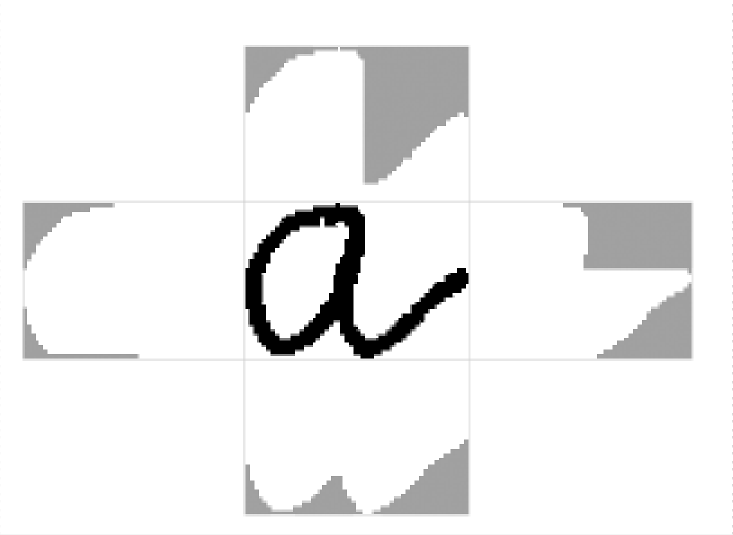
- The basic idea behind using projections is that character images, which are 2-D signals, can be represented as 1-D signal. These features, although are not dependant to noise and deformation, depend on rotation.

- 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” .

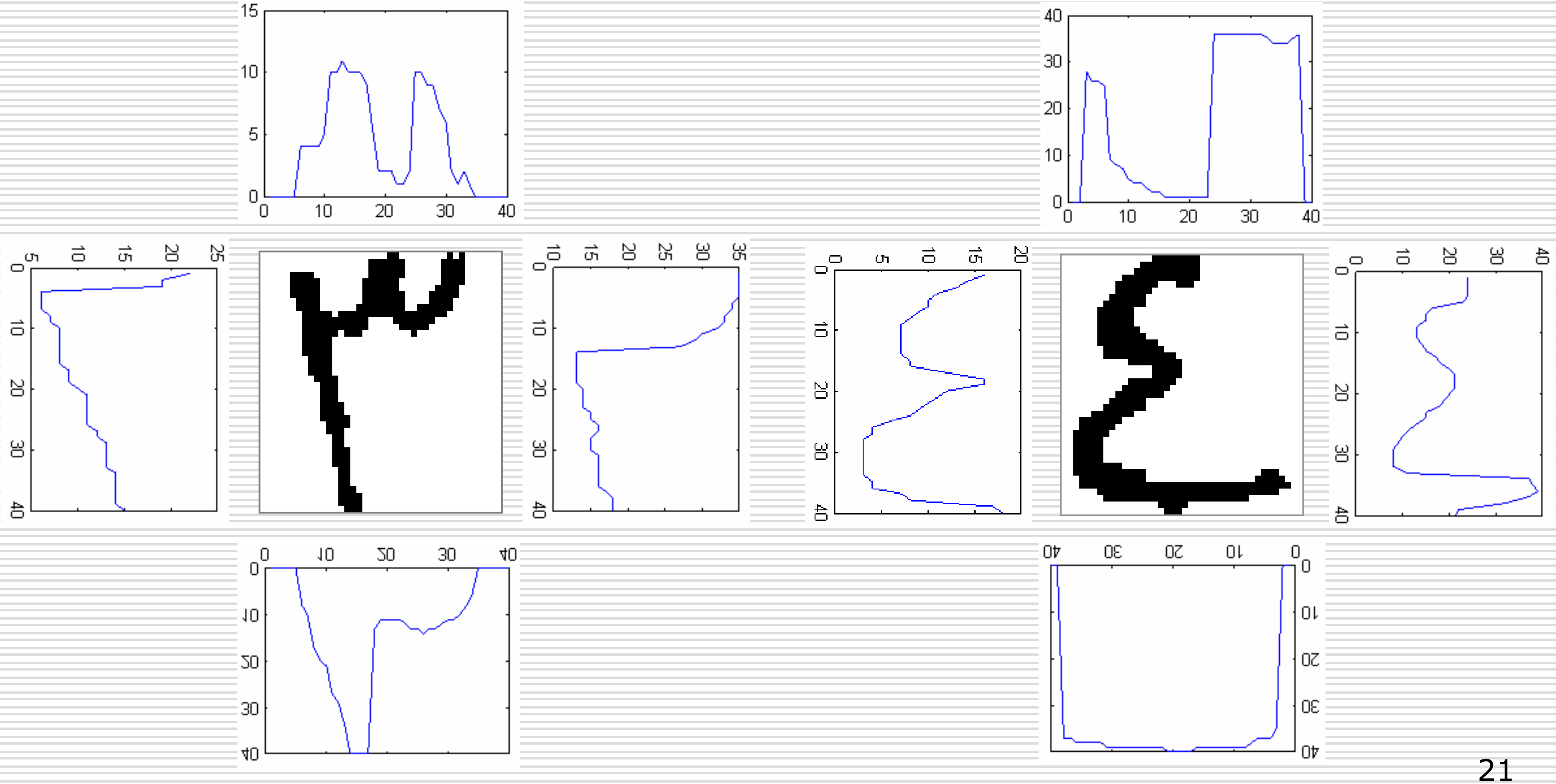


Profiles

- 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”.



Example

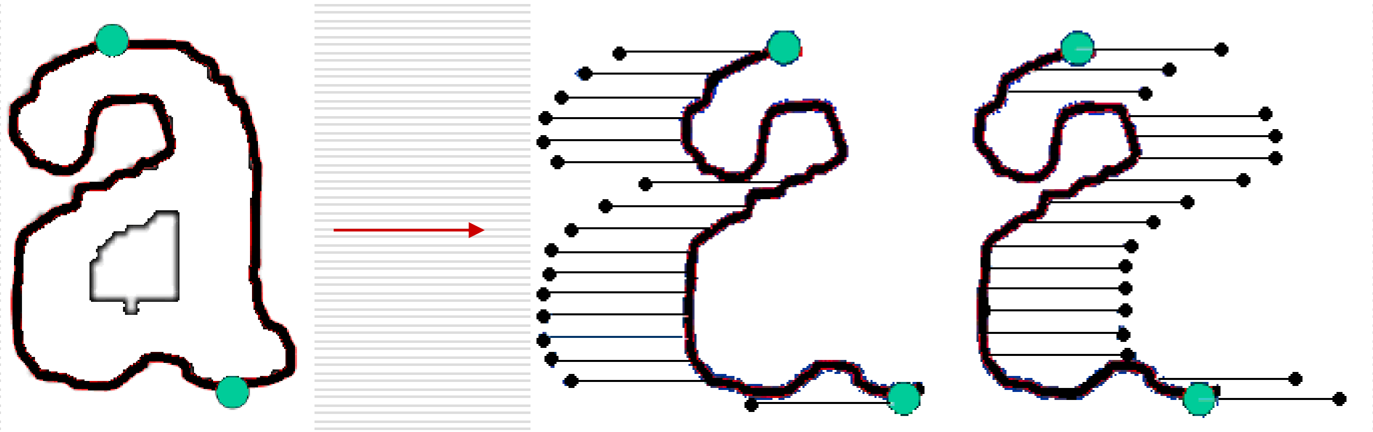


0 5 10 15
0 10 20 30 40
0 10 20 30 40
5 10 15 20 25
0 10 20 30 40
10 15 20 25 30 35
0 5 10 15 20
0 10 20 30 40

Profiles

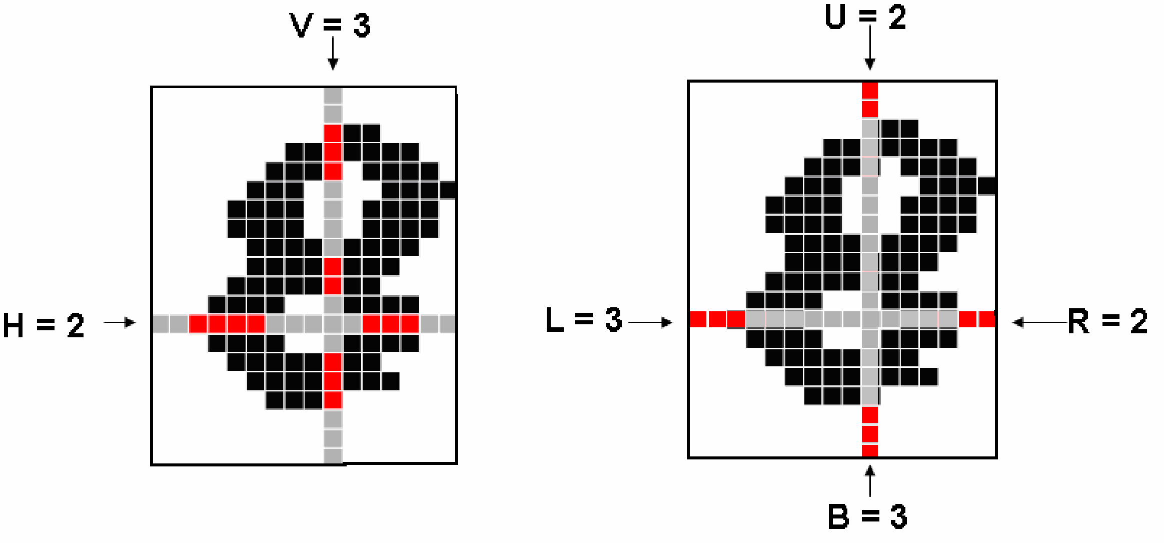
□ Profiles can also be used to the contour of the character image

- 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

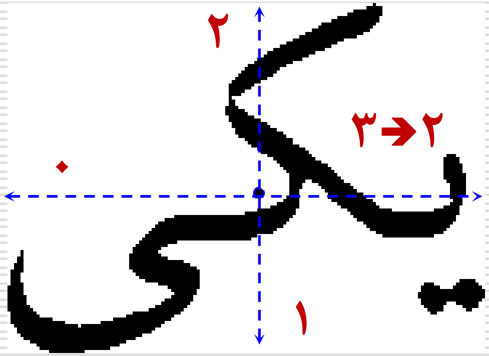
□ **Crossings** count the number of transitions from background to foreground pixels along vertical and horizontal lines through the character image and **Distances** calculate the distances of the first image pixel detected from the upper and lower boundaries, of the image, along vertical lines and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



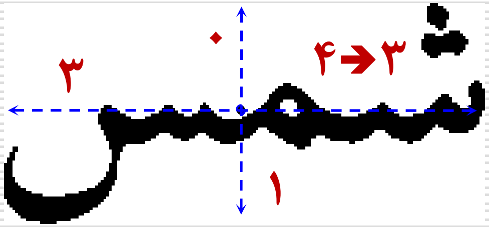
Crossings

Distances

Characteristic Loci, Extended Loci



$(1022)_3 = 2 + 2\times3 + 0\times9 + 1\times27 = 35$



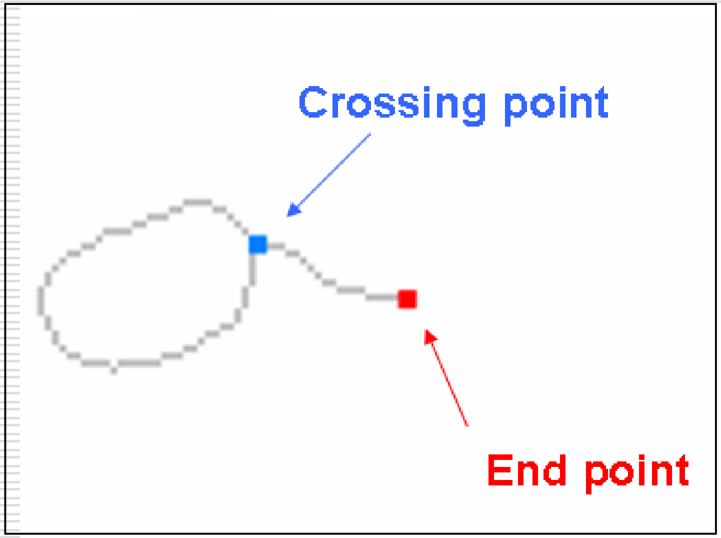
$(1303)_{4\text{و}3} = 3 + 0\times4 + 3\times(4\times3) + 1\times(4\times3\times4) = 87$

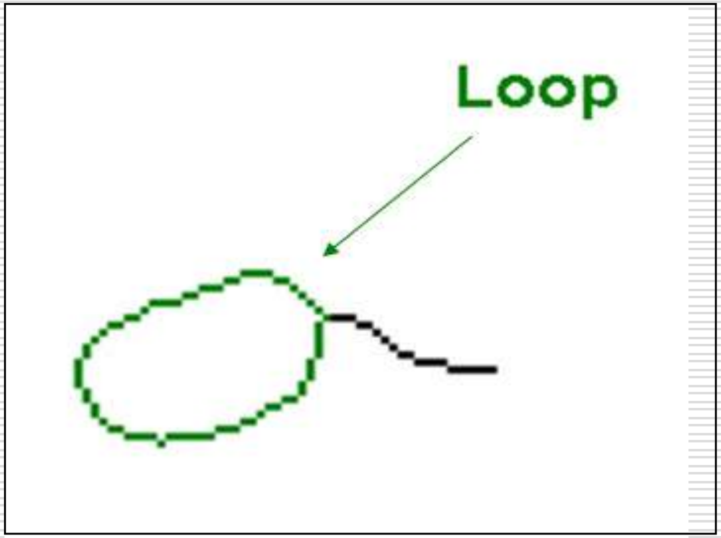
Structural Features

- 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.

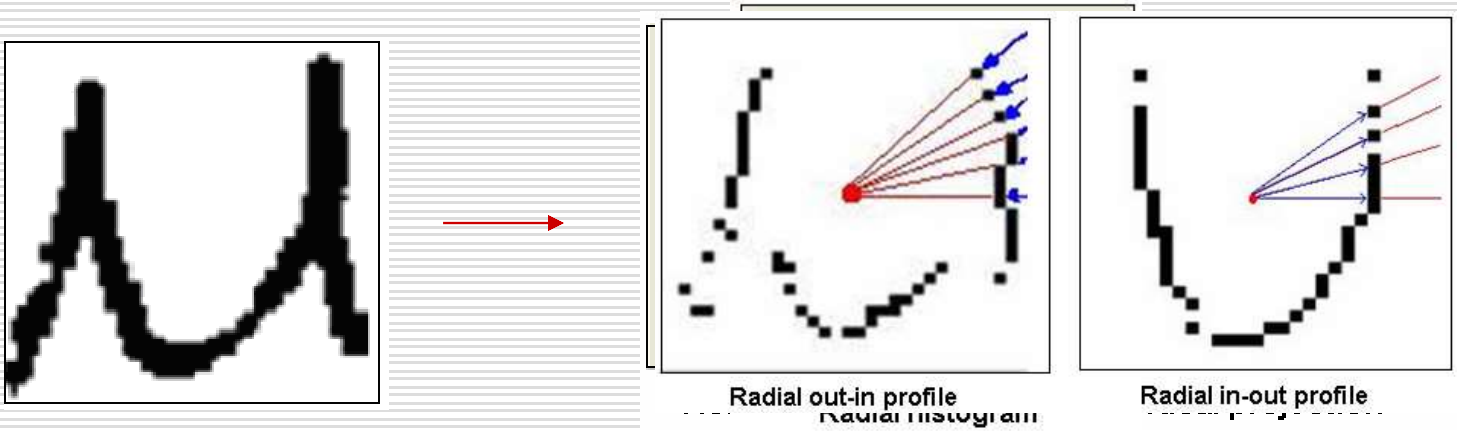
Structural Features





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



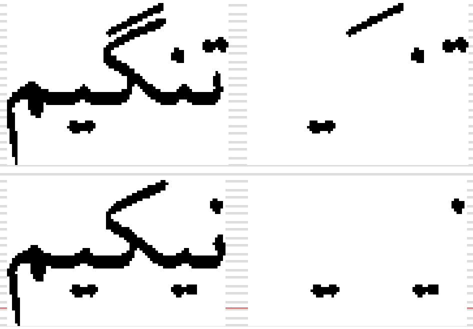
Radial out-in profile

Radial in-out profile

Signs and dots

*

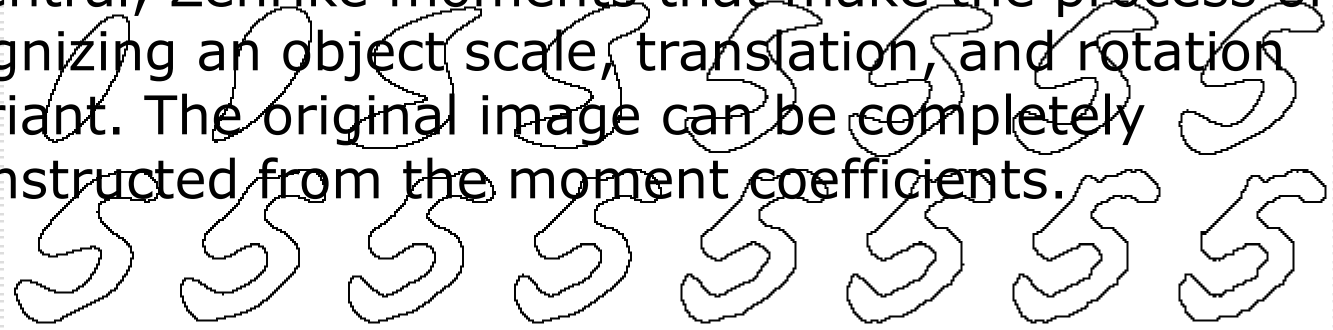
- بازشناسی بر اساس ترتیب و تعداد نقاط
- شگفت: *3u4u1u2u*
- بسیج: *1b2b1m*
- شمس، سمش: *3u*



Global Transformations - Moments

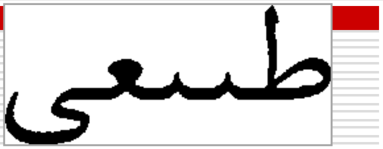
- The Fourier Transform (FT) of the contour of the image is calculated. Since the first *n* coefficients of the FT can be used in order to reconstruct the contour, then these *n* coefficients are considered to be a *n*-dimesional feature vector that represents the character.

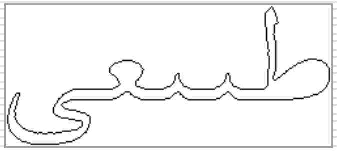
- Central, Zenrike moments that make the process of recognizing an object scale, translation, and rotation invariant. The original image can be completely reconstructed from the moment coefficients.

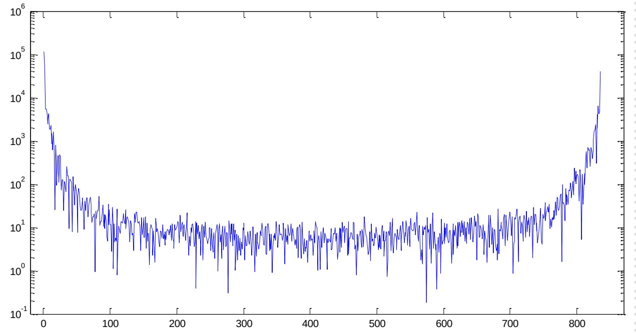


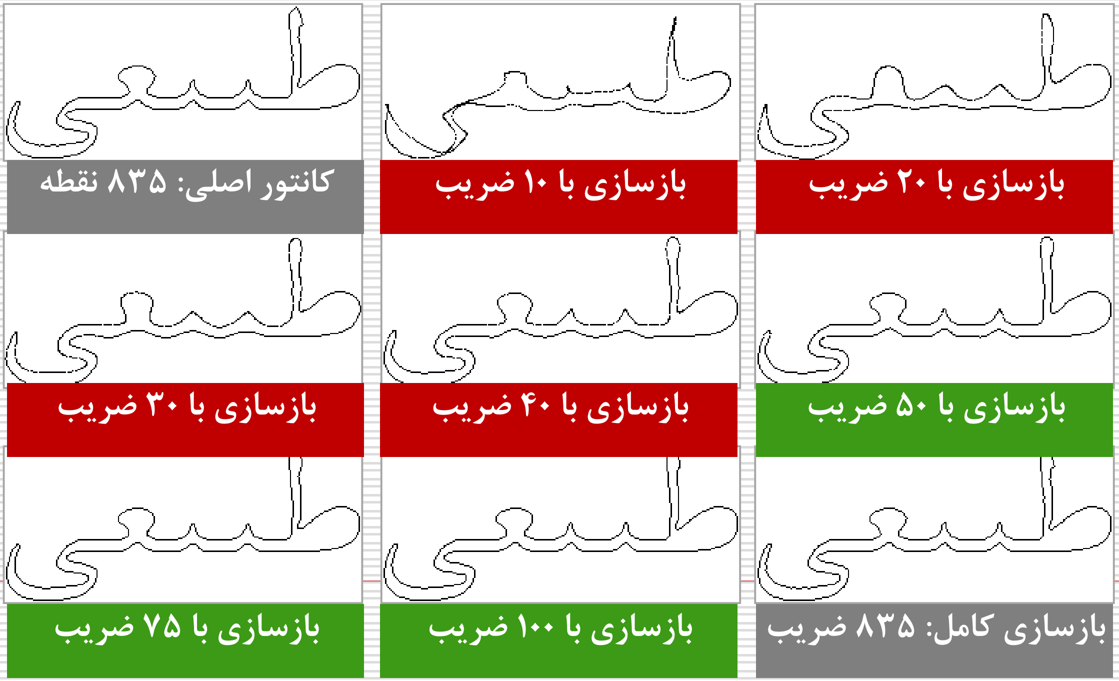
FFT on Contour

Useless when signs and dots are available









کانتور اصلی: ۸۳۵ نقطه

بازسازی با ۱۰ ضریب

بازسازی با ۲۰ ضریب

بازسازی با ۳۰ ضریب

بازسازی با ۴۰ ضریب

بازسازی با ۵۰ ضریب

بازسازی با ۷۵ ضریب

بازسازی با ۱۰۰ ضریب

بازسازی کامل: ۸۳۵ ضریب

گشتاورهای زرنیک

- در سال ۱۹۳۴ توسط آقای زرنیک (فیزیکدان) مطرح شد
- آقای ختن زاد در سال ۱۹۹۰ از اندازهٔ آنها برای بازشناسی تصاویر دو بعدی استفاده کرد
- دارای تبدیل معکوس است.
- امروزه رواج زیادی یافته است
- سرعت استخراج ویژگی بسیار کند است
  - هر چند روشهایی برای استخراج سریع آن مطرح شده است.
- کارایی چندانی در مورد ارقام فارسی ندارد

معرفی گشتاور زرنیک

$$A_{nm} = \frac{n+1}{\pi}\sum_{x}\sum_{y} f(x,y)[V_{nm}(x,y)]^{*}, \qquad x^{2}+y^{2}\le 1$$

$$V_{nm}(x,y) = R_{nm}(x,y)e^{jm\tan^{-1}(\frac{y}{x})}$$

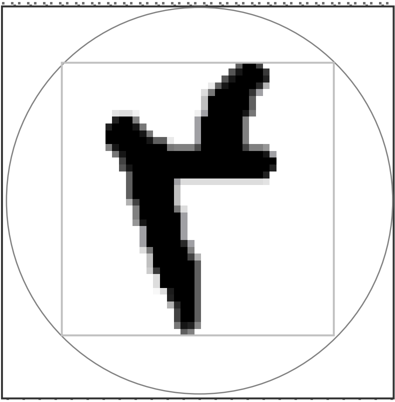
where

$$j=\sqrt{-1}\,, n\ge 0\,, |m|\le n\,, n-|m| \;\; is \;\; even$$

$$R_{nm}(x,y) = \sum_{s=0}^{\frac{n-|m|}{2}} \frac{(-1)^{s}(x^{2}+y^{2})^{\frac{n}{2}-s}(n-s)!}{s!(\frac{n+|m|}{2}-s)!(\frac{n-|m|}{2}-s)!}$$

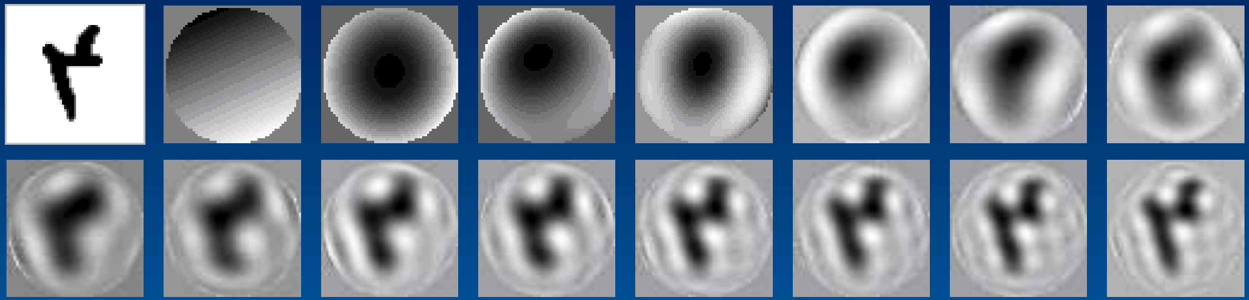
پیش پردازش

- در محاسبه گشتاور زرنیک، x و y باید داخل دایرۀ واحد باشند و نیز تنها قمستی از تصویر را می توان بازسازی کرد که داخل دایرۀ مذکور قرار دارد
- تصویر اصلی را باید داخل یک دایره محاطی قرار داده و اطراف آنرا با زمینه پر کنیم

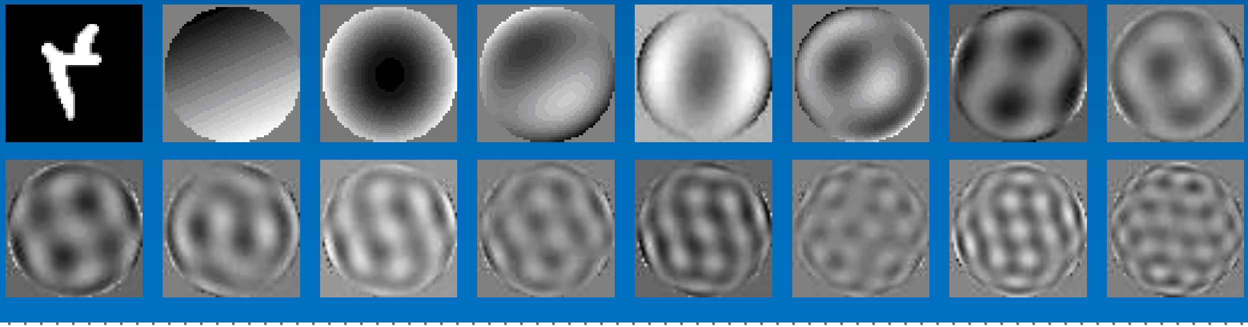


بازسازی تصویر از روی گشتاور های زرنیک

$$f(x,y)=\lim_{N\to\infty}\sum_{n=0}^{\infty}\sum_{m}A_{nm}V_{nm}(x,y)$$



$$I_n(x,y)=\mathrm{Re}\left(\sum_{m}A_{nm}V_{nm}(x,y)\right)$$



استخراج ویژگی با استفاده از گشتاور زرنیک

- استفاده از اندازهٔ گشتاور زرنیک
  - مستقل از چرخش
  - گشتاورهای مرتبهٔ ۱ تا ۱۵ مرسوم است
  - طول بردار ویژگی: ۱۳۵
  - نرخ بازشناسی ارقام: داده های آموزش ۹۴.۸۶٪، داده های آزمایش ۸۹.۸۳٪
- استفاده از مولفهٔ حقیقی گشتاور زرنیک
  - طول بردار ویژگی: ۱۳۵
  - نرخ بازشناسی: داده های آموزش ۹۶.۱۴٪، داده های آزمایش ۹۱.۶۷٪

استخراج ویژگی با استفاده از گشتاور زرنیک

- استفاده از هر دو مولفهٔ حقیقی و موهومی
- طول بردار ویژگی : ۲۷۰
- نرخ بازشناسی: داده های آموزش ۹۹.۱۳٪، داده های آزمایش ۹۷.۱۳٪

270:40:10	0	1	2	3	4	5	6	7	8	9	Sum	RR
0	1935	٠	٠	٠	3	40	2	15	2	3	2000	96.8
1	٠	1986	٠	٠	1	1	4	5	2	1	2000	99.3
2	٠	12	1943	29	6	1	2	2	٠	5	2000	97.2
3	1	٠	79	1886	20	2	1	4	1	6	2000	94.3
4	1	7	21	61	1887	11	2	3	٠	7	2000	94.4
5	14	2	2	1	3	1960	2	7	8	1	2000	98
6	٠	6	12	5	1	6	1937	4	1	28	2000	96.9
7	٠	5	17	1	٠	2	3	1970	2	٠	2000	98.5
8	٠	8	٠	٠	3	1	6	٠	1965	17	2000	98.3
9	1	22	3	1	1	2	8	٠	6	1956	2000	97.8
Sum	1952	2048	2077	1984	1925	2026	1967	2010	1987	2024		97.13

تحلیل گشتاور زرنیک

- یادگیری خوب داده های آموزش در صورت استفاده از هر دو مولفهٔ حقیقی و موهومی.
- قدرت تعمیم آن خوب نیست.
- روی ارقام ۳ و ۴ خطای زیادی دارد. (۹۴.۳٪ و ۹۴.۴٪)
- روی سایر ارقام بازشناسی خوبی دارد.
  - برای ترکیب با ویژگی مکان مشخصه مناسب به نظر می رسد
    - نتیجهٔ ترکیب: ۹۸.۴۵٪
    - ۱.۳۲٪ بهبود
- سرعت استخراج ویژگی بسیار کند
  - ۶۶ بار کندتر از مکان مشخصه

Classification

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

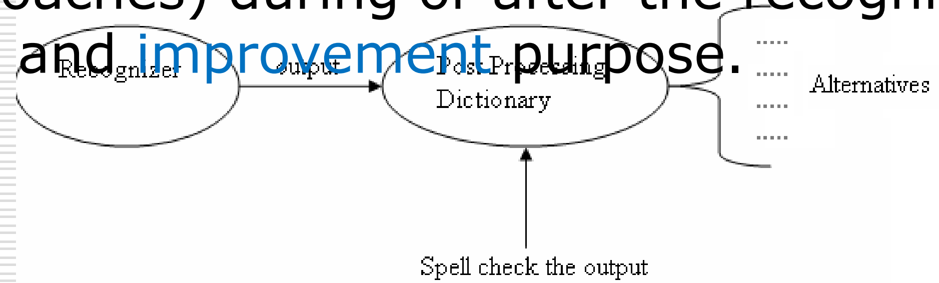
There is no such thing as the “best classifier”. The use of classifier depends on many factors, such as available training set, number of free parameters etc.

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 the context information is the utilization of a dictionary for correcting the minor mistakes.

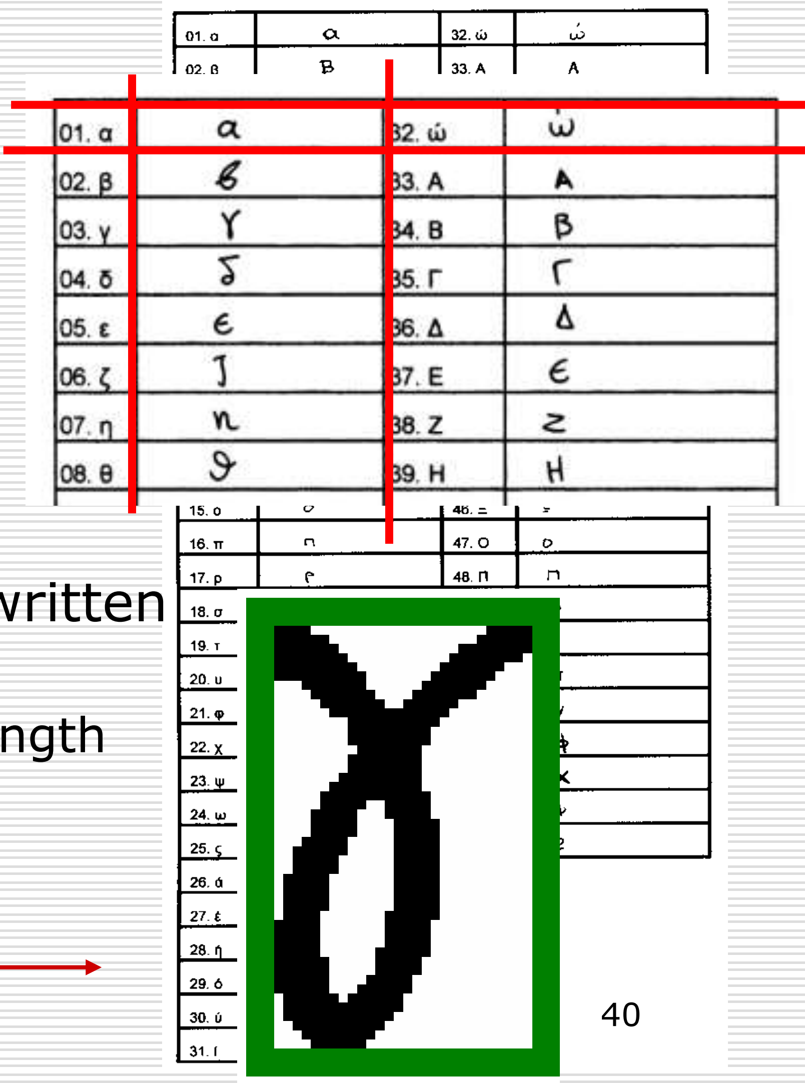
□ In addition to the use of a dictionary, a well-developed lexicon and a set of orthographic rules (lexicon-driven matching approaches) during or after the recognition stage for verification and improvement purpose.



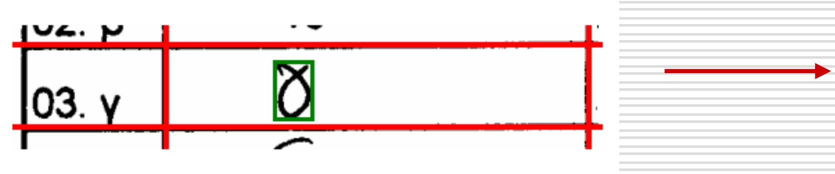
□ Drawback : Unrecoverable OCR decisions.

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 “ά”, “έ”, “ή”, “ί”, “ύ”, “ό”, “ώ”

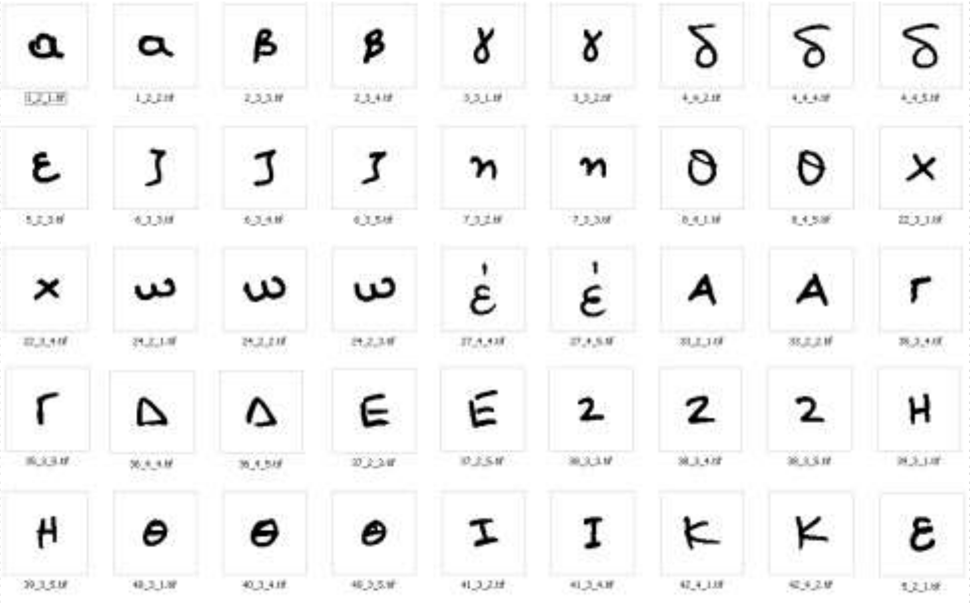


- □ The steps led to the Greek handwritten character database are:
  - Line detection using Run Length Smoothing Algorithm (RLSA)
  - Character extraction



CIL- Greek Handwritten Character Database

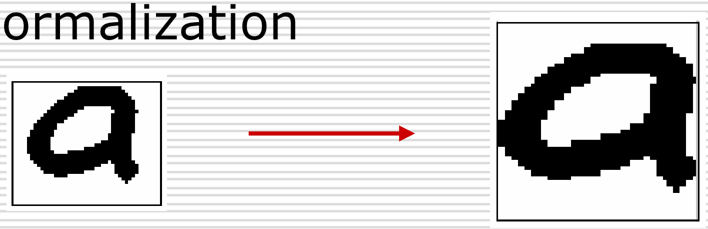
- □ CIL Database:
  - 125 Greek writers
  - 5 forms per writer
  - 625 variations of each character led to an overall of 35,000 isolated and labeled Greek handwritten characters



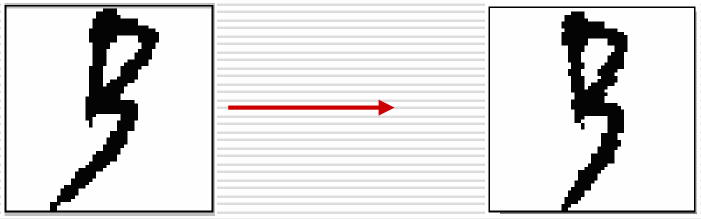
Proposed OCR Methodology

□ Pre-processing :

- Image size normalization

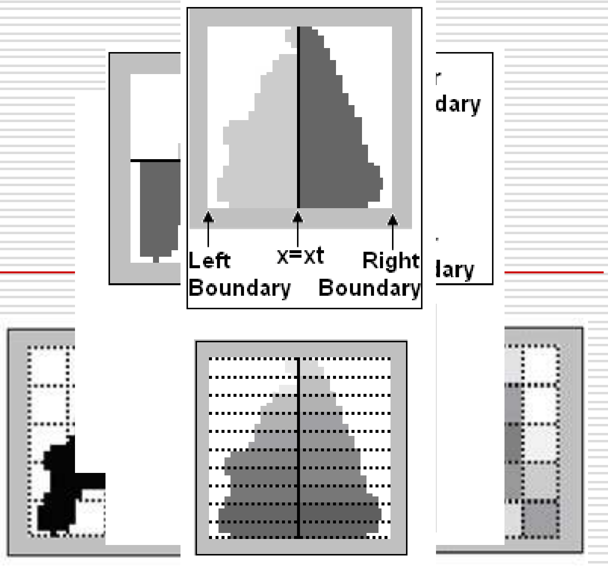


- Slope correction



□ 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

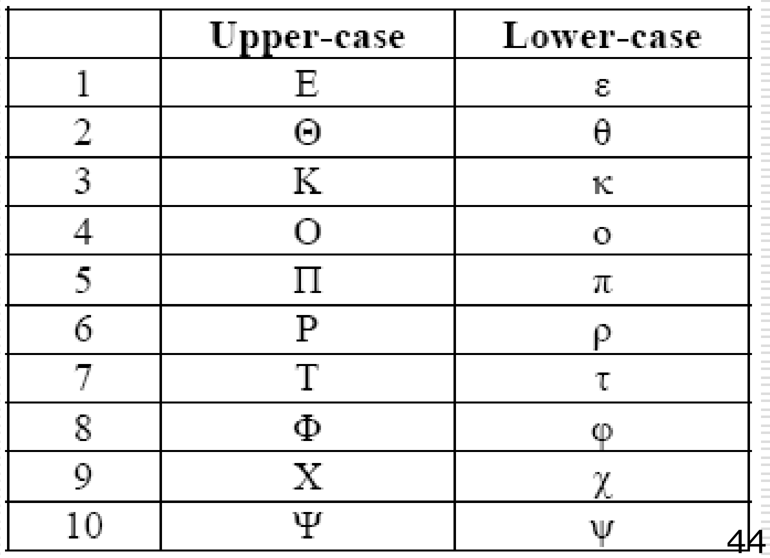
  ✓Upper/ lower profiles are computed by considering for each image column, the distance between the horizontal line $y = y_t$ and the closest pixel to the upper/lower boundary of the character image. This ends up in two zones depending on y_t . Then both zones are divided into vertical blocks. For all 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

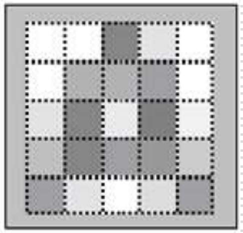
- 10 pairs of classes were merged, due to size normalization step, resulting to a database of 28,750 characters.



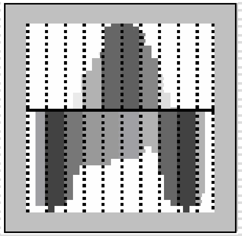
	Upper-case	Lower-case
1	Ε	ε
2	Θ	θ
3	Κ	κ
4	Ο	ο
5	Π	π
6	Ρ	ρ
7	Τ	τ
8	Φ	φ
9	Χ	χ
10	Ψ	ψ

Experimental Results

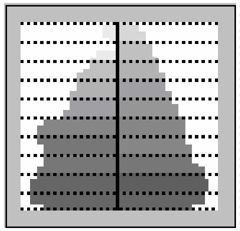
- □ 1/5 of each class was used for testing and 4/5 for training
- □ Character images normalized to a 60x60 matrix
- □ Features
  - Based on Zones
    - ✓ 5 horizontal and 5 vertical zones =>25 features



  - Based on Upper and Lower profiles
    - ✓ 10 vertical zones => 20 features



  - Based on Left and Right profiles
    - ✓ 10 horizontal zones => 20 features

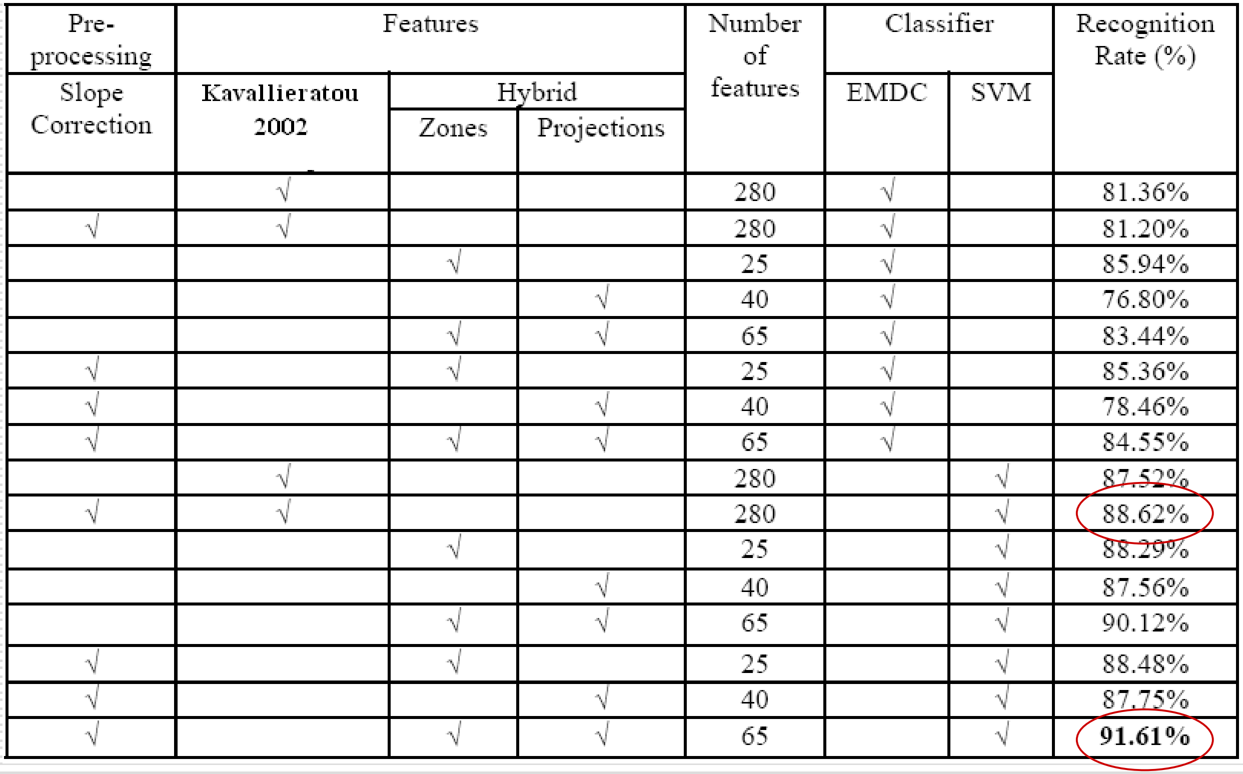


  - Total Number of features

    25 + 20 + 20 = 65

Experimental Results

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

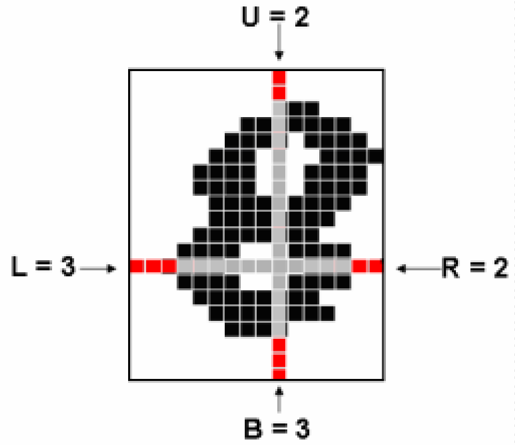


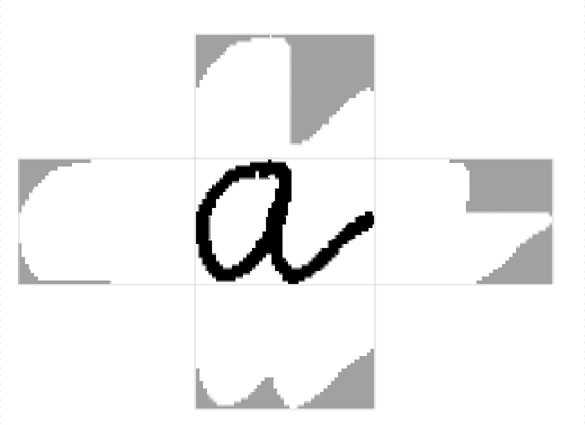
Pre-processing	Features			Number of features	Classifier		Recognition Rate (%)
Slope Correction	**Kavallieratou 2002**	Hybrid			EMDC	SVM	
		Zones	Projections				
	√			280	√		81.36%
√	√			280	√		81.20%
		√		25	√		85.94%
			√	40	√		76.80%
		√	√	65	√		83.44%
√		√		25	√		85.36%
√			√	40	√		78.46%
√		√	√	65	√		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

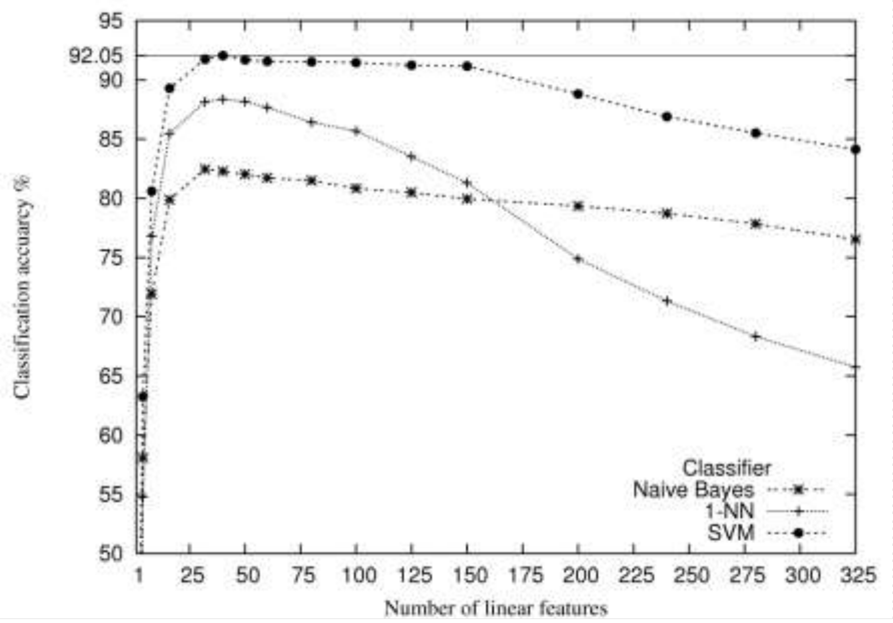




Experimental Results

- 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

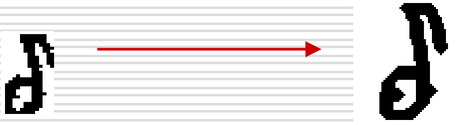


- Recognition Rate = 92.05%
- Number of features = 40

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

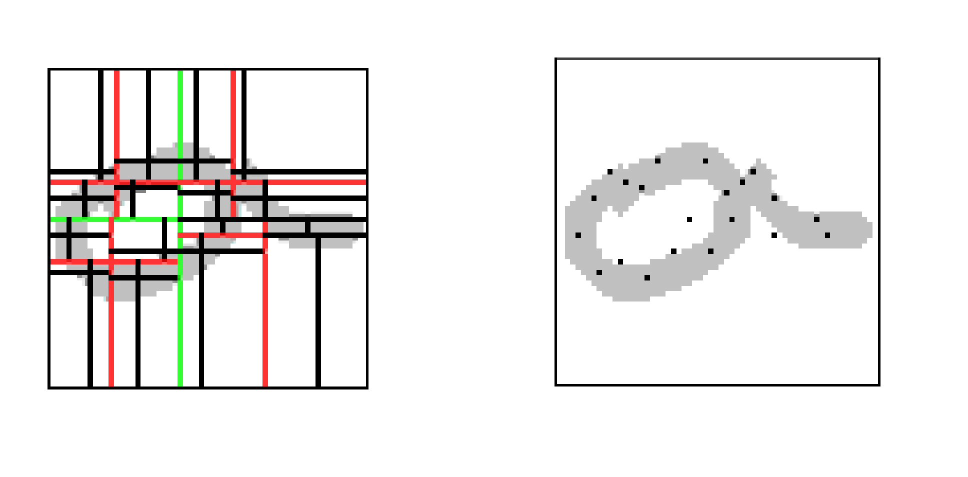
- G. Vamvakas, B. Gatos, I. Pratikakis, N. Stamatopoulos, A. Roniotis and S.J. Perantonis, "**Hybrid Off-Line OCR for Isolated Handwritten Greek Characters**", *The Fourth IASTED International Conference on Signal Processing, Pattern Recognition, and Applications* (SPPRA 2007), ISBN: 978-0-88986-646-1,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

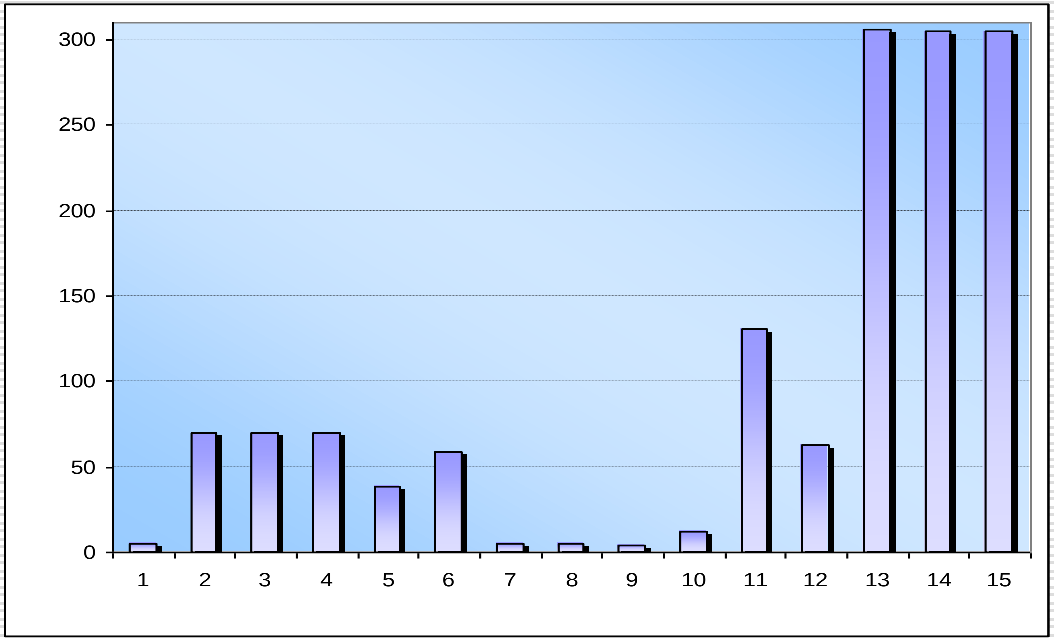
- "**An Efficient Feature Extraction and Dimensionality Reduction Scheme for Isolated Greek Handwritten Character Recognition",** 9th International Conference on Document Analysis and Recognition (ICDAR 2007), Curitiba, Brazil, September 2007. Waiting...

Future Work

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



Time efficiency



1. Loci
2. DCT 81
3. DCT 144
4. DCT 225
5. Kirsch
6. Improved Kirsch
7. Contour Profile
8. Improved Contour Profile
9. Projection
10. Zoning

۱۱. گرادیان

۱۲. گرادیان بهبود یافته

۱۳. اندازهٔ گشتاور زرنیک

۱۴. مولفهٔ حقیقی گشتاور زرنیک

۱۵. مولفهٔ حقیقی و موهومی گشتاور زرنیک