

Arabic Text Segmentation

By
Dr. Salah M. Rahal
King Saud University-KSA

Outline

- **Introduction.**
- **Arabic Language**
 - **Arabic Language Features.**
 - **Challenges for Arabic OCR.**
- **OCR System Stages.**
- **Text Segmentation.**
- **Databases,**
- **Conclusion.**

Introduction

O C R (Optical Character Recognition)

OCR is the recognition of text by a computer, i.e. It is the process of using computer system to translate images of text (printed or handwritten) into machine-editable text.

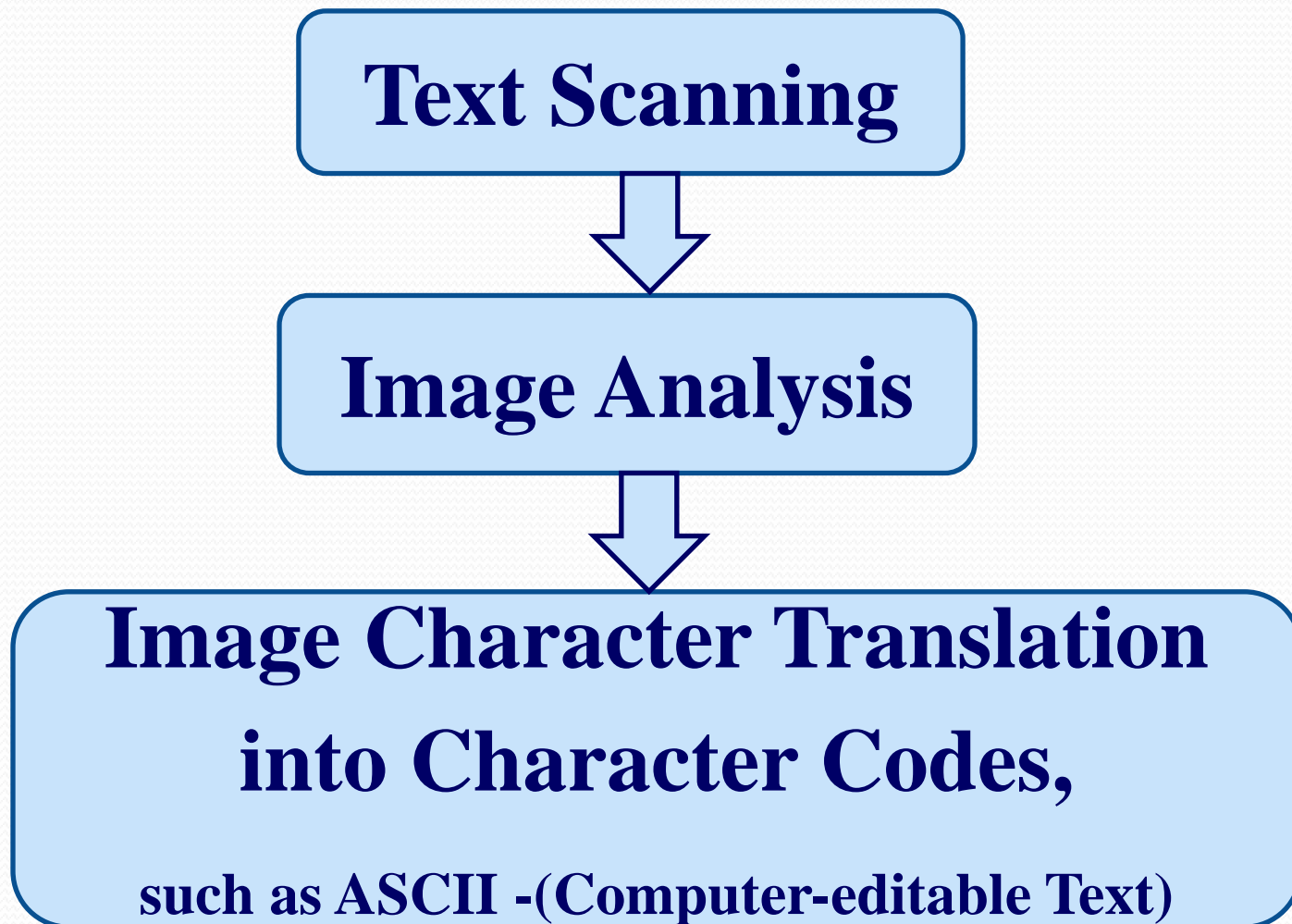


Translation of the character image into digital characters.

OCR Goal:

Simulation of the human ability to read both machine-printed and handwritten texts.

OCR involves:



OCR is an important front end for different systems such as Electronic Document Management (EDM) systems.

Excellent OCR now exists for Latin based languages, but there are few systems that read Arabic.



Arabic Language

Arabic Language

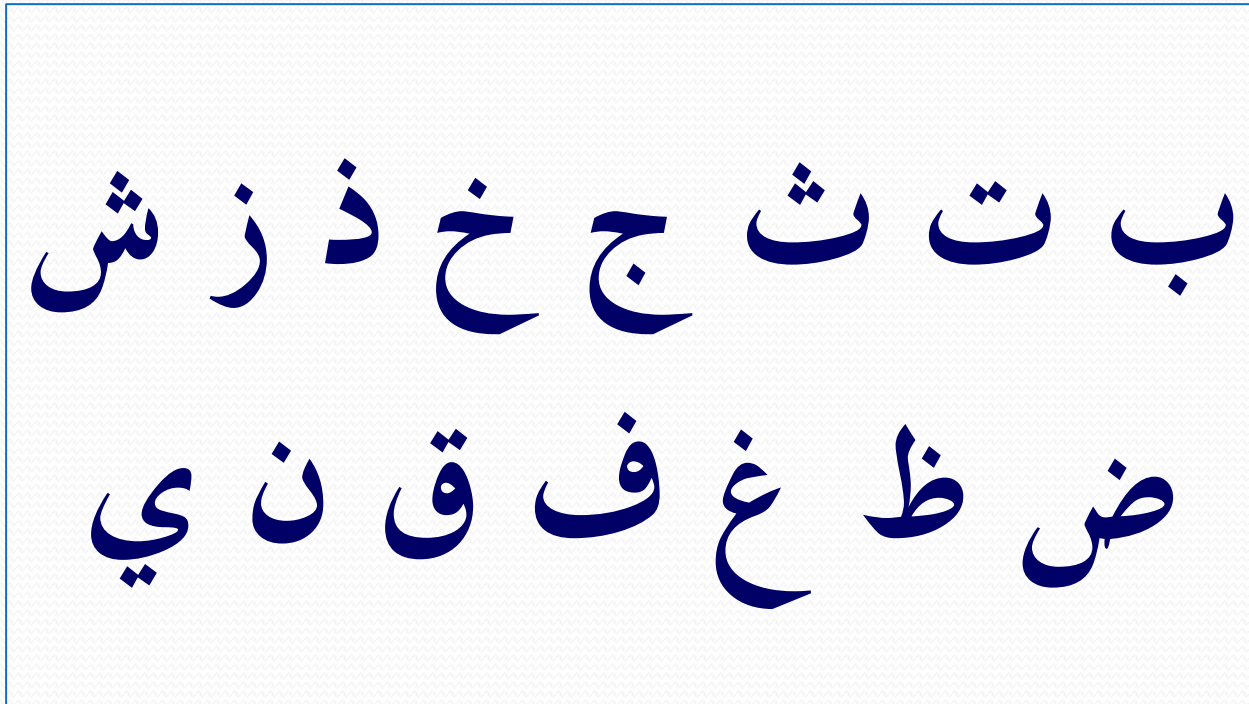
- **Arabic language is a rich language. It contains a large number of words*.**
 - **More than 420 million speakers.**
 - **Official language of Arabic Countries.**
 - **One of the six official languages of the United Nations** (along with Chinese, English, French, Russian and Spanish).
 - **More than 1.5 milliard Muslims need Arabic language.**
 - **Other languages use Arabic alphabet, for example Pashto, Persian, Sindhi, and Urdu.**
- * No standard reference list containing all Arabic words.**

Arabic Language Features:

➤ Arabic language has **28** basic characters:

ر	ذ	د	خ	ح	ج	ث	ت	ب	ا
Reh	Thal	Dal	Khah	Hah	Jeem	Theh	Teh	Beh	Alef
ف	غ	ع	ظ	ط	ض	ص	ش	س	ز
Feh	Ghain	Ain	Thah	Tah	Dad	Sad	Sheen	Seen	Zain
	ي	و	ه	ن	م	ل	ك	ق	
	Yeh	Waw	Heh	Noon	Meem	Lam	Kaf	Qaf	

➤ **15** of Arabic alphabet have **dot(s)**:



Characters with dot(s).

Arabic Language Features

- **Dot(s) can exist in the form of one, two, or three dots.**
- **Dot(s) can be written either above or below the character.**

One dot	ب ج خ ذ ز ض ظ غ ف ن
Two dots	ث ق ي
Three dots	ش

- In addition to the basic 28 characters, there are **supplementary characters**:
 - Hamza (ء) in the middle & in the end:
 - in the middle **تهنئة**
 - in the end **سماء** ، **يرجى**
 - Hamza combined with other letters:



Arabic Language Features

- **Madda (~):**

آ آفاق

- **Alef maksoura (ى)**

ى على

- **Teh marbuta (ة)**

ة الجزيرة الكورية

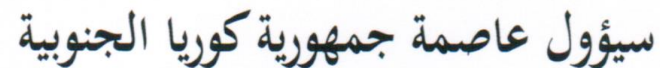
- **Lam Alef (لا) : It consists of two letters (ل ا)**

Other Arabic Features:

- **Arabic text** (both handwritten & printed) is written from right to left.
- **Arabic script is cursive** (printed & handwritten).
- **Arabic characters are connected from the baseline of the word.**



سيؤول عاصمة جمهورية كوريا الجنوبية



سيؤول عاصمة جمهورية كوريا الجنوبية

Arabic Language Features

- **Arabic contains only one case characters (no upper and lower case).**
- **The digits used in the Arabic are called **Arabic-Indic digits** (originally invented in India & adapted by the Arabic language).**

٩	٨	٧	٦	٥	٤	٣	٢	١	٠
9	8	7	6	5	4	3	2	1	0

Arabic Language Features

- **19 joining groups – Same body.**
Difference is number of dots (or hamza).
(example : ب ت ث).

No	Schematic Name	Joining Group	Group Characters
1	Alef	ا	أ إ آ
2	Beh	ب	ب ت ث
3	Hah	ح	ج ح خ
4	Dal	د	د ذ
5	Reh	ر	ر ز
6	Seen	س	س ش
7	Sad	ص	ص ض

Arabic Language Features

8	Tah	ط	ظ
9	Ain	ع	غ
10	Feh	ف	ف
11	Qaf	ق	ق
12	Kaf	ك	ك
13	Lam	ل	ل
14	Meem	م	م
15	Noon	ن	ن
16	Heh	ه	ه
17	Waw	و	ؤ
18	Yeh	ي	ى
19	Teh Marbuta	ة	ة

In addition to 28 letters, Arabic text includes:

- ❖ punctuation marks,**
- ❖ spaces and,**
- ❖ special symbols.**
- ❖ Mathematical symbols.**

Punctuation Marks, Such as:

!	؟	"	‘	.	%
EXCLAMATION MARK	question mark	QUOTATIO N MARK	COMMA	DECIMAL POINT / FULL STOP	PERCENT SIGN
#	\$	()	*	/
NUMBER SIGN	DOLLAR SIGN	OPENING PARENTHE SIS	CLOSING PARENT HESIS	ASTERISK	SLASH

Texts in Chinese, Japanese, and Korean were generally left unpunctuated until the modern era, when they adopted Western punctuation marks:

http://en.wikipedia.org/wiki/Punctuation#Conventional_styles_of_English_punctuation

Some punctuation marks in Arabic look different from the English counterparts:

	Comma	Question mark
Arabic	،	؟
English	,	?

Challenges for Arabic OCR

- Arabic characters are **cursive** and not separated as is the case with Latin script.
- **Shapes:** Characters change **shape** depending on their position in the word
 - ➔ much of the distinction between isolated characters is lost when they appear in the middle of a word.

Challenges of Arabic OCR

A character can have up to four shapes according to its location in the word: start, middle, end, and isolated. Examples:

No of Shapes	Character Shapes
1	و ر د
2	ا س س ا م م ض ض م م ي ي
3	ه ه ه ق ق ق
4	ع ع ع ع غ غ غ غ

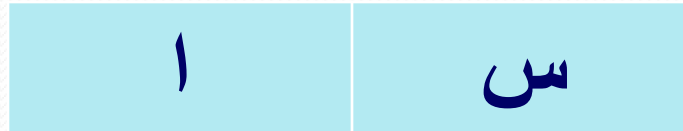
Example – (ع) :

على ، العربية ، مع ، قطاع

- **Dots:** Different characters with same body. Distinction only by the number and location of dots.

ذ	د	ج	ح	خ	ب	ث	ث
ض	ص	س	ش	س	ر	ز	ر
يا	نا	ع	غ	ع	ظ	ظ	ظ

- Characters of the same font have **different sizes:**



- Arabic writing contains **many fonts** and **writing styles:**



Challenges of Arabic OCR

- **6 Arabic characters are not connectable with the succeeding character.**

ا د ذ ر ز و

They are joined from the right side only.

ا د ذ ر ز و

In the joining type defined by the Unicode Standard all the Arabic letters are Dual Joining, except these letters which are joined from the right side only.

Challenges of Arabic OCR

- if one of these characters exists in a word, it divides that word into sub-words

كوريا الجنوبية ذات اقتصاد مزدهر

كوريا الجنوبية ذات اقتصاد مزدهر

- Sometimes Arabic writers neglect to include whitespaces between words when the word ends with one of these letters.

- Repeated characters are sometimes used:



- There are two ending letters which sometimes indicate the same meaning (ة ه).
- There is often misuse of the letter Alef (ا) in its different shapes (أ إ).

- **The letter (و) can be either a sub-word or individual word. The meaning of the word is "and". It is often misused. It should have whitespace after it, but mostly it is neglected.**

Challenges of Arabic OCR

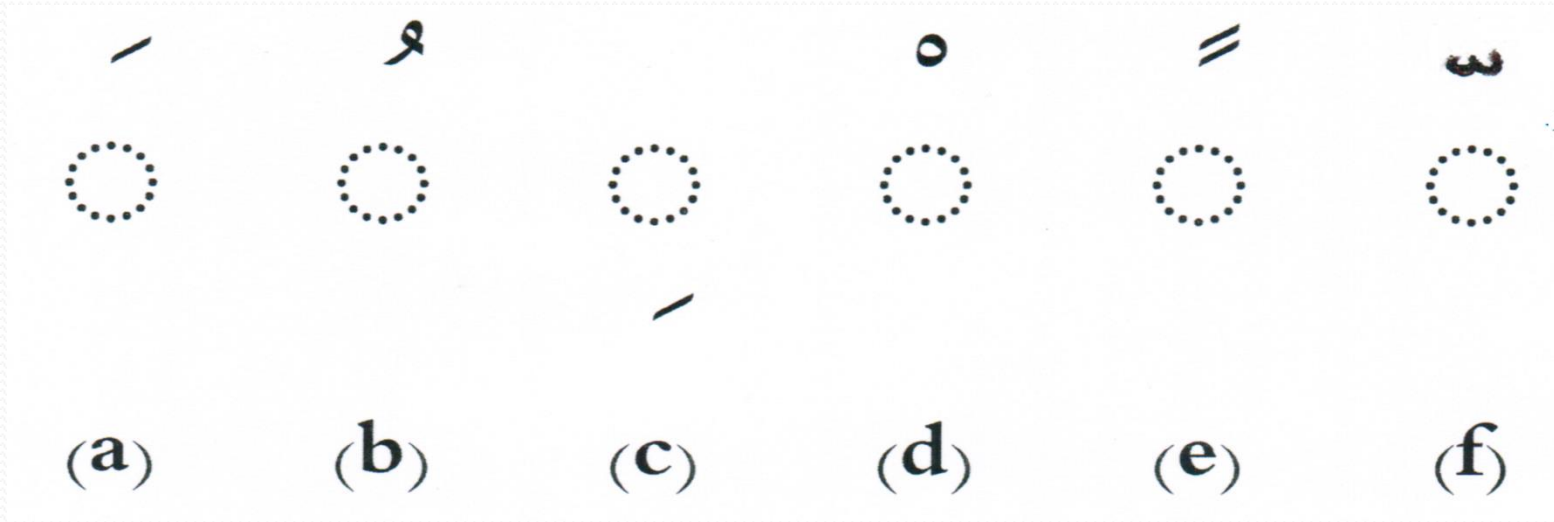
- Arabic language contains a similarity between the following letters and digits:

Digit		Letter	
One	١	ا	Alef
Five	٥	هـ	Heh

Same between “full stop (.)” and “Arabic number Zero(٠)”.

Challenges of Arabic OCR

➤ Diacritical marks:



Diacritical marks: (a) fat-ha, (b) dumma, (c) kesra, (d) sukun, (e) nunation, and (f) shadda.

Challenges of Arabic OCR

Diacritical marks (called Harakat) are used above and below the letters to help in pronouncing the words and in indicating their meaning.

عِلْمٌ	عَلَّمَ	عَلَمٌ	عِلْمِ
It is known	He taught	Flag	Science

Notes on Arabic text (both handwritten & printed):

- **Small No of characters having the same shape in any position.**
- **The position of the character may differ relating to the line: on the line (سـ), under the line (ر) , up the line (ا).**
- **Width & length of characters differ from one character to another.**

- Certain compounds of characters form “ligatures”.

المجتمع، ملائمة، يمتنعون، الجامعة

المجتمع، ملائمة، يمتنعون، الجامعة

- The connecting letter known as Tatweel, or Kashida is used to adjust the left alignments; this letter has no meaning in the language.

- **Arabic handwritten text segmentation is still considered to be a major challenge in document image analysis due to the different styles of handwriting and the connectivity of the Arabic letters.**



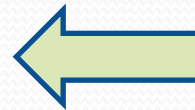
O C R

System Stages

Arabic Text Recognition System



Text Acquisition



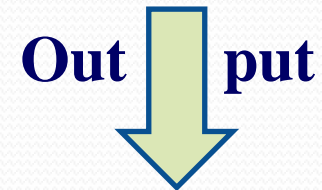
Text



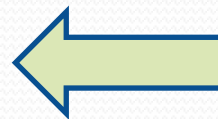
**Document
Pre-processing**



Segmentation



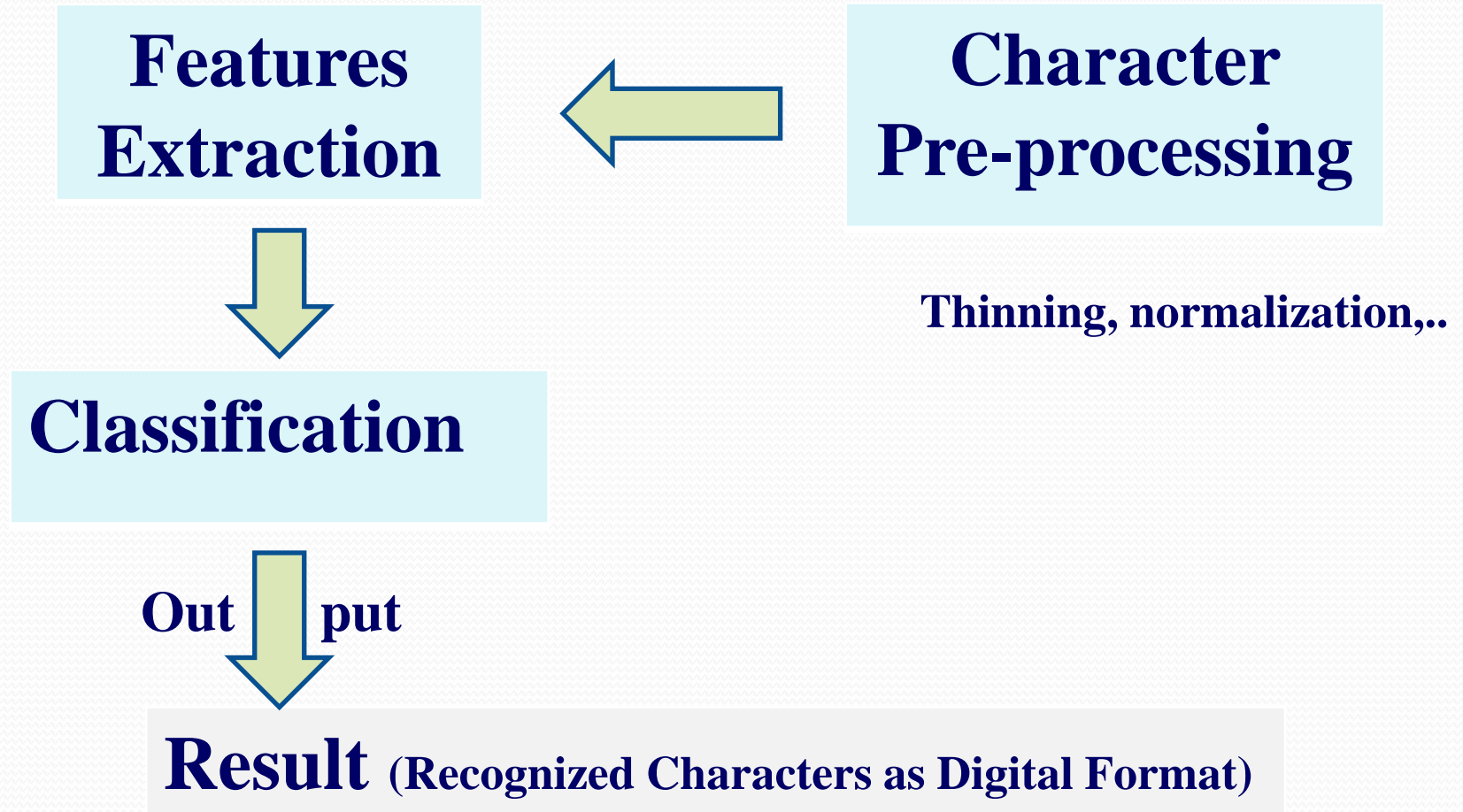
Recognition Step



isolated character
(As images)

Arabic Text Recognition System

Recognition Step



Text Segmentation

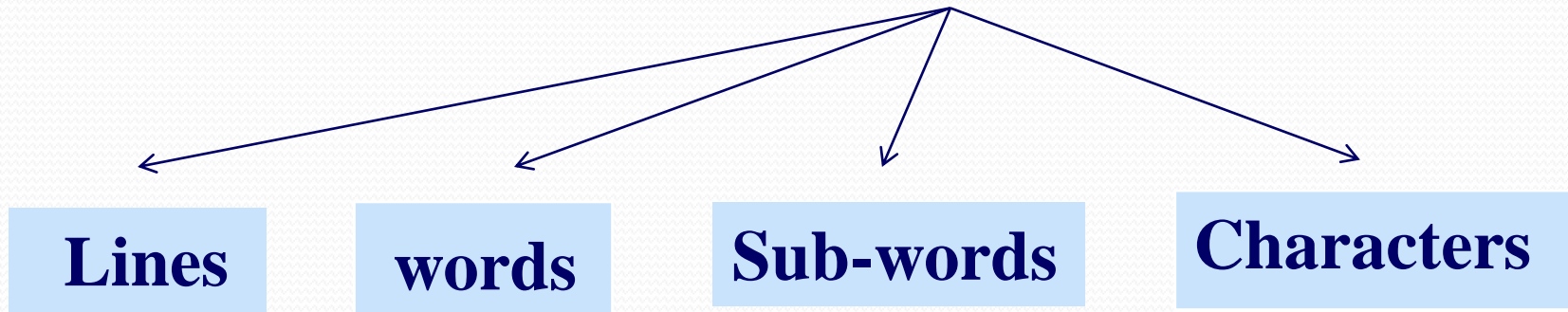
Arabic character segmentations face many technical difficulties. The most challenging problem is the **cursive** characteristic of Arabic text (printed or handwritten). Letters within a word are joined to one another by a baseline and words are separated by spaces. Most of the characters are formed by curves and loops.

Loops are usually drawn in clockwise direction.

While the segmentation is relatively simple in printed Roman texts, it is still an open question in Arabic.

Sub-components of Segmentation

The first critical step in the development of text recognition system is the segmentation of the text. This divides the text into its sub-components



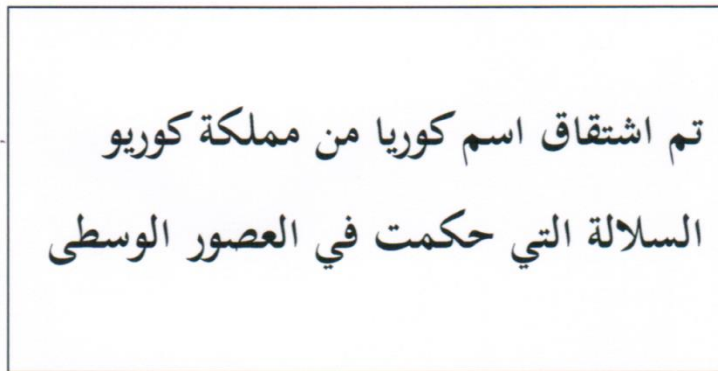
It is an important stage: The reached result in each step directly affects the recognition rate.

Text Segmentation Steps:

1- Lines Segmentation.

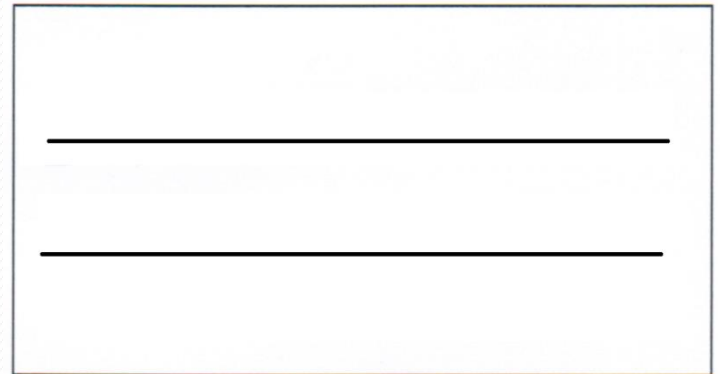
Aim: Segmentation of an image document into horizontal lines using horizontal projection.

Input: image document.



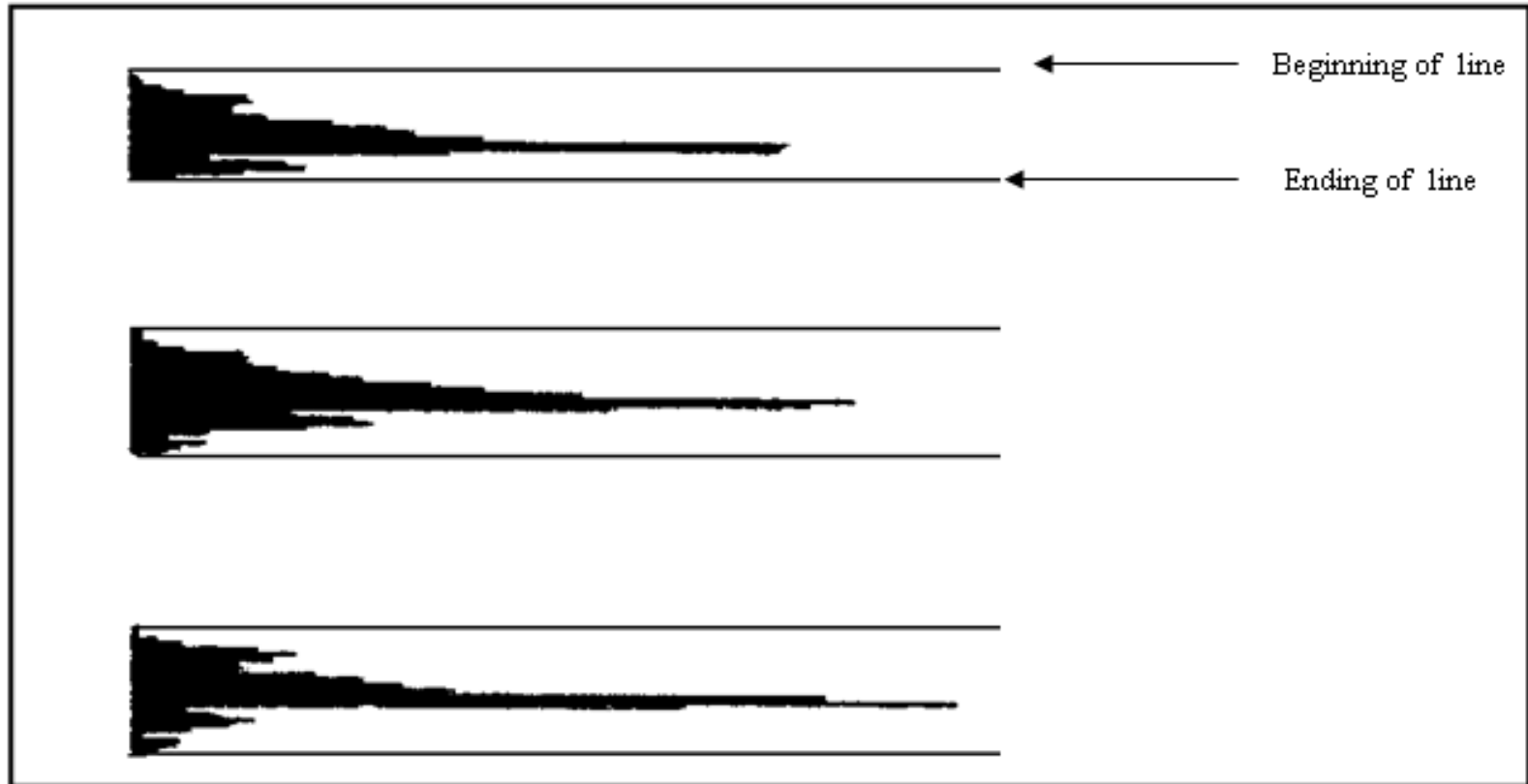
Arabic handwritten document.

Output: line images.



Segmentation of the image document into its horizontal lines.

Segmentation

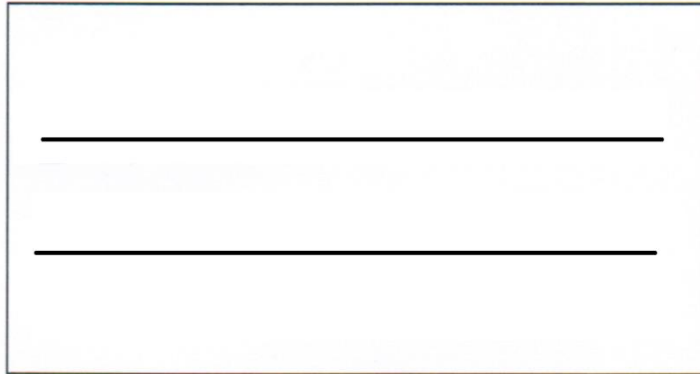


Horizontal projection of lines.

2- Word Segmentation

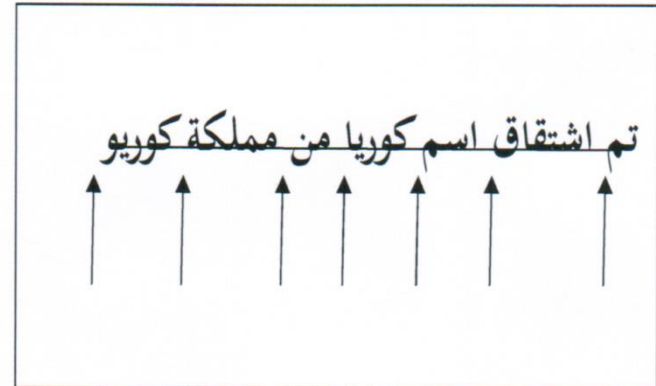
Aim: Segmentation of a line into words/sub-words using vertical projection.

Input: Line image.



Line image.

Output: Word images.



Segmentation of a line into its words.

Segmentation



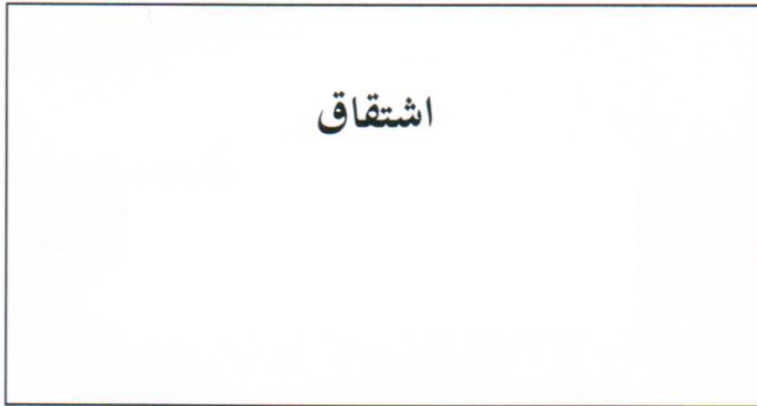
Vertical projection of a line.

3- Sub-words Segmentation

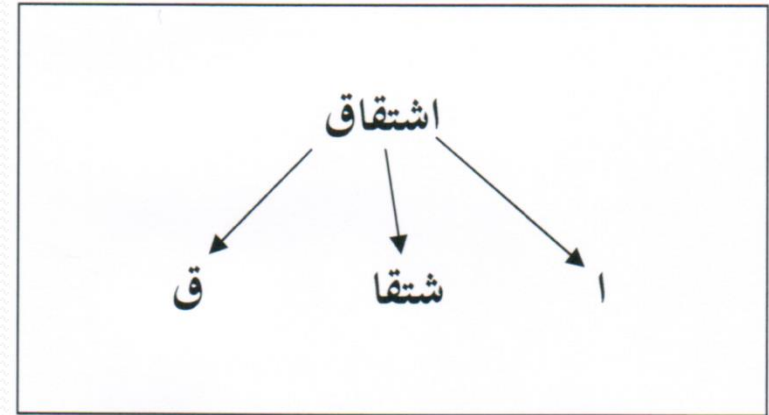
Aim: Segmentation of a word into its sub-words.

Input: Word image.

Output: Sub-words images.

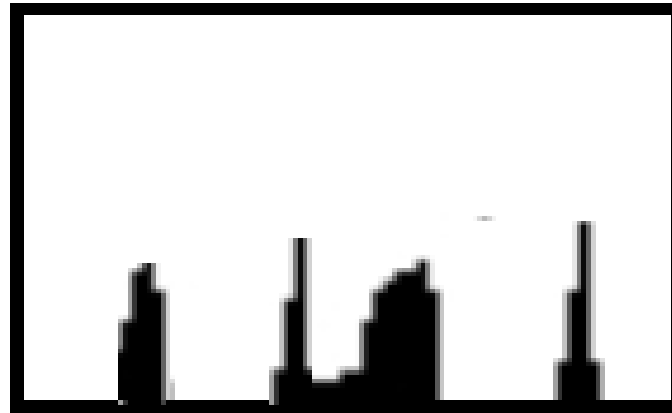


Word image.



Segmentation into its sub-words.

Words/sub-words Segmentation

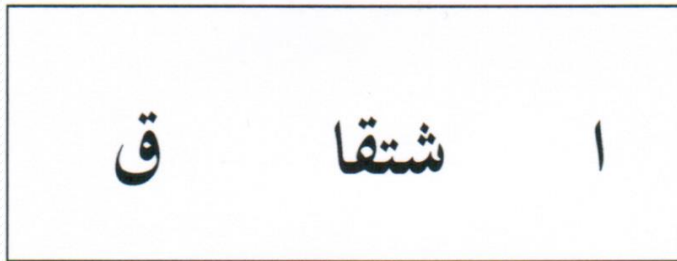


Vertical projection of a line.

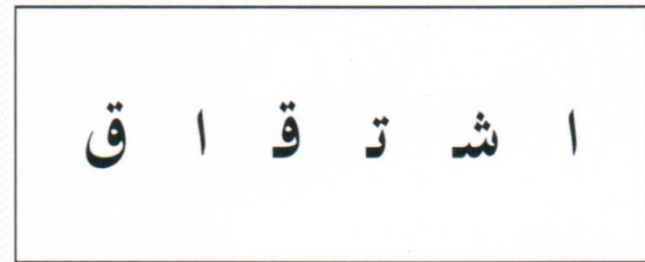
4- Character Segmentation

Aim: Segmentation of a connected part into its isolated characters.

Input: Connected part images. Output: Character images.



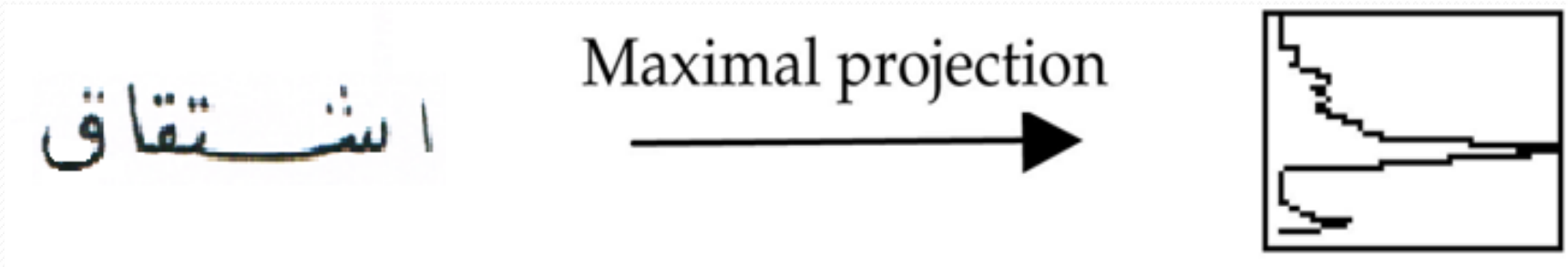
connected parts.



Segmentation of a connected part into its characters.

- **The “baseline” is the line at the height at which letters are connected and it is analogous to the line on which an English word sits.**
- **Letters are wholly above the baseline except for descenders and some markings.**

Baseline



Horizontal projection of a word used to detect the baseline.

Databases

Some databases for printed words were cited:

- **Database of 6 million Arabic words selected from different sources.**
- **The Linguistic Data Consortium (LDC) at the University of Pennsylvania produced “Arabic Gigaword” that contains more than 1 Giga Arabic words (5-th Edition, 2011).**

Handwritten databases:

- **A database of 100 different writers which contains Arabic text and words. It contains the most common Arabic words that are used in writing checks and some handwritten pages.**
- **A database of 26,400 names (town/village) completed by 411 writers.**

It was created by the Institute for Communications Technology (IFN), Technical University Braunschweig, Germany and Ecole Nationale d'Ingénieur de Tunis (ENIT).

This database has been used recently in a number of other research projects.

Conclusion

The Arabic Language characteristics (cursiveness, different sizes for the same letter, dot(s),..) and the meaning change imposed especially by diacritical marks make the high segmentation rate and the high recognition rate a challenging questions in the development of high reliable Arabic OCR.

A good database should be performed to achieve the mentioned above purpose.

Tank You