N° d'ordre :
N° de série :

**PEOPLE'S DEMOCRATIC REPUBLIC OF ALGERIA**
**Ministry of Higher Education and Scientific Research**

**ECHAHID HAMMA LAKHDAR UNIVERSITY - EL OUED**
**FACULTY OF EXACT SCIENCES**
**Computer Science department**

**End of Study Memory**
**Presented for the Diploma of**

# ACADEMIC MASTER

Domain : **Mathematics and Computer Science**
spinneret: **Computer Science**
Speciality : **Artificial Intelligence and Distributed Systems**


Presented by :

- **Atia Djaber**
- **Touati Tliba Mohammed**


## Theme

# Proposition an Algorithm for Segmentation an Ambiguous Text of Arabic

Supported in: - - 2019 In front of jury:

| | | | |
|---|---|---|---|
| M. | Kholladi Nedjoua Houda | MCA | President |
| M. | Othmani Samir | MAA | Reporter |
| Dr. | ZAIZ Faouzi | MAA | Superviser |

**Année Universitaire:** 2018/2019

# Thanks

We want to thank from the bottom of our hearts, before all, ALLAH
the all-powerful of my possessions gives the strength, the will and the courage to carry out this work.
We would like to thank warmly and respectively all those who contributed from near and far to the realization of this project of end of study and we thank our supervisor Mr. ZAIZ Faouzi, master at the University Echahid hamma lakhdar El-Oued , to have proposed and supervises this subject
We express our deep gratitude to him for sharing his knowledge, but also his methods of work, and especially his scientific rigor.
We thank the jury members for agreeing to judge our work.
Our gratitude goes also to all those who collaborated in our formation especially the teachers of the department of Computer Science, Echahide University Hamma Lakhdar El-Oued Also to our colleagues of the promotion 2018-2019.
We also thank all those who have participated from near or far to develop this work.

# Dedication

To our families, our brothers

, our sisters

our friends and all the people who support us

# الملخص

شهدت انظمة التعرف على الكتابة (OCR: Optic Character Recognition) تطورا كبيرا في العديد من المجالات مثل: التعرف على محتوى الصكوك البنكية, تحويل الملفات المطبوعة الى نصوص قابلة للتعديل بالآلة. اعطت العديد من البحوث المنجزة في مجال التعرف على الخط نتائج جيدة جدا خاصة اللغات اللاتينية,ولكن بالرغم من ذلك مزالت تعاني من نقائص كثيرة في مجال التعرف على الخط العربي وخاصة المكتوب بخط اليد بسبب طريقة الكتابة والإلتصاق والتقاطع بين الأحرف. قمنا بتقديم دراسة عامة حول انظمة التعرف على الخط العربي, ثم تطرقنا الى مرحلة مهمة وهي تقسيم الكلمات ,حيث تطرقنا الى مختلف المشاكل التي تواجه تقسيم الخط العربي بعدها شرحنا مختلف الطرق المستعملة للتقسيم ,ثم خطوات التقسيم. في هذا العمل , قمنا بطرح خوارزمتين تعتمد على البحث عن نقاط الاهمية لتجزئة الكلمة ,حيث الخوارزمية الاولى تعمل على استخراج مشكل التقاطع بين الحروف والخوارزمية الثانية تعمل على التقسيم الافقي للكلمات المكتوبة بشكل عمودي. قمنا بتجريب العمل المطروح على عينة من الصور المكتوبة بخط اليد العربي ,وكانت النتائج جيدة ومشجعة جدا.



**الكلمات المفتاحية: التعرف على الكتابة , كتابة اليد , التجزئة , PAW .**

**ABSTRACT**

OCR: Optic Character Recognition has developed in many areas such as: recognition of the contents of bank instruments, conversion of printed files into adaptable texts. Many of the research done in the field of handwriting recognition has produced very good results especially in Latin languages. However, there are many shortcomings in the field of handwriting recognition, especially in handwriting, because of the way of writing, sticking and intersecting letters. We presented a general study on the systems of recognition of the Arabic calligraphy, and then we addressed the important stage of the division of words, where we addressed the various problems facing the division of Arabic calligraphy then explained the different methods used for division, and then division. In this work, we have introduced two algorithms based on the search for the points of importance of word segmentation, where the first algorithm works to extract the problem of the intersection of letters and the second algorithm works on the horizontal division of words written vertically. We tried the work on a sample of pictures in Arabic handwriting, and the results were very good and encouraging

**Mots Clés : OCR, segmentation, caractères arabes, pseudo mot**

**RÉSUMÉ**

OCR: La reconnaissance optique de caractères s'est développée dans de plisseur domaines tels que: la reconnaissance du contenu des instruments bancaires, la conversion de fichiers imprimés en textes adaptables. De nombreuses recherches dans le domaine de la reconnaissance de l'écriture manuscrite ont donné de très bons résultats, notamment dans les langues latines. Cependant, il existe de nombreuses lacunes dans le domaine de la reconnaissance de l'écriture manuscrite, en particulier de l'écriture manuscrite, en raison de la manière d'écrire, de coller et de croiser les caractères. Nous avons présenté une étude générale sur les systèmes de reconnaissance de la calligraphie arabe, puis nous avons abordé l'importante étape de la division des mots : nous avons abordé les divers problèmes de la division de la calligraphie arabe, puis expliqué les différentes méthodes utilisées pour la division, puis la division. Dans ce travail, nous avons introduit deux algorithmes basés sur la recherche des points d'importance de la segmentation des mots, le premier algorithme permettant d'extraire le problème de l'intersection des lettres et le deuxième l'algorithme traitant de la division horizontale des mots écrits verticalement. Nous avons essayé le travail sur un échantillon de photos en écriture arabe, et les résultats ont été très bons et encourageants.

**Keywords: OCR, segmentation, Arabic Characters, PAW**

# TABLE OF CONTENT

**II. SEGMENTATION**



**III. SEGMENTATION AND FEATURES EXTRACTION OF WRITING**

**IV. TEST AND RESULT**

# LIST OF FIGURES

## I.OPTICAL CHARACTER RECOGNITION



## II.SEGMENTATION

## III. SEGMENTATION AND FEATURES EXTRACTION OF WRITING

## IV. TEST AND RESULT

# LIST OF TABLES

## I.OPTICAL CHARACTER RECOGNITION



## III. SEGMENTATION AND FEATURES EXTRACTION OF WRITING



## IV. TEST AND RESULT

# ABBREVIATIONS

**- OCR : Optical Character Recognition.**

**- PAW : Peace of Arabic Word.**

**- GIF Graphics Interchange Format**

**- BMP Bitmap**

**- JPEG Joint Photographic Experts Group**

**- AOCR Arabic Optical Character Recognition.**

**- PS Pen Size.**

**- HDP Horizontal Division Points**

**- VFP Vertical Fusion Points**

**- HFP Horizontal Fusion Points**

**- VDP Vertical Division Points**

**-PVN Point of Horizontal Nature**

**-PHN Point of Vertical Nature**

# GENERAL INTRODUCTION

The research on the recognition of Arabic characters exposes a field that is rapidly expanding and indefinitely evoked by such an important place in the last two decades. Today recognition of Arabic characters is a concern whose relevance is undisputed by the community of researchers who have devoted their efforts to reducing constraints and expanding the kingdom of recognition of Arabic characters.

Writing for communication has always been a primary concern of man. The written word has been, and will remain, one of the great foundations of civilizations and the world par excellence of conservation and transmission of knowledge. Despite the advances of other means of communication such as audio visual, many are the applications whose existence begins on paper, more particularly in the office, in desktop publishing (to facilitate the composition from a selection of several documents), in the post office (reading addresses and automatic sorting), in banks (processing of checks, invoices). However, in spite of technological progress, the keyboard is still an obligatory means of communication with the computer.

we say that the Recognition is on-line if the data is acquired dynamically during writing. Often, a graphics tablet and an electronic pen are used by a user. On the other hand, recognition is off-line when the source image is the result of a scanner or an image database.

The purpose of this memory is to propose a system for the segmentation of an Arabic handwriting text. In offline writing

In the first chapter we will introduce and present a state of the art in the field of document recognition, then we will to look at problems of the OCR, also we will present the different approaches, methods and techniques in this field.

In the second chapter, we will present the most important step that is the segmentation of handwritten and present the approaches used in the segmentation of handwritten text especially the segmentation of the word in characters as well as the strategy that is used.

in the chapter three, we will explain our contribution to solving many of the problems facing Arabic handwriting, where we have developed many new rules to help segment words correctly.
In the fourth chapter, we will explain the program interface and the steps that are used to acquire the image, then the filter of the image obtained. After that we will explain also the steps of converting the images which requires to upload a series of images contains parts of the word processing, then we will present the results obtained by our system.
We finish this work with a conclusion on the results obtained by the method used,

# Chapitre I

# OPTICAL CHARACTER RECOGNITION

# INTRODUCTION

The field of recognition and segmentation of handwritten in an automated way is the one of the most important research topics in artificial intelligence field. There is a lot of research on the writing of Latin, Chinese, Japanese…...etc. unlike the Arabic has a little chance from these research.

Optical Character Recognition (OCR) is a process of detecting and extracting text of an image file, an image embedded in an electronic document or a scan of a document. It used in several field such as in postal domain for the recognition of the code of the postal address and the automatic reading in checks in bank domain also in the administrative domain for the electronic management of document flows or indexing documents and searching for information in digital libraries field.... etc.

In this chapter we will introduce and present a state of the art in the field of document recognition, then we will to look at problems of the OCR, also we will present the different approaches, methods and techniques in this field.

## 1. DEFINITION

Optical Character Recognition (OCR) is the process of extracting character's symbols from images that contain text [23].

First, OCR system segment these images into a set of lines, words and characters. After that, extracts from each basic character form a characteristic vector which is unique for describing the given form. After that, performs the recognition of the character.

The result is a textual transcription of the image in which we can perform the automatic processing: searching for statements, words, manipulating text. Generally, the OCR concerns the processing of a digital document [22, 24].

## 2.THE ASPECTS OF OCR

In the field of recognition of handwritten Arabic, we can distinguish two types of recognition systems [2]:

- On-line and off-line recognition
- Global or analytical recognition

### 2.1 On-line and Off-line recognition

In this section, we will present the difference between on-line and off-line recognition. Also, we will see the main advantages and disadvantages of each one.

#### 2.1.1 On-line writing (or dynamic [11])

On-line recognition is to perform parallel or real-time treatments with writing, and It is reserved for handwriting. In this case, it is necessary to use equipment such as a graphics tablet or an electronic pen. It has the advantage of having additional information such as movements and pressure. In addition, it allows to correct and modify the writing interactively [25, 26].

#### 2.1.2 Offline writing (or static [11])

On the other hand, Offline recognition uses a scanner or camera to acquire a binary or grayscale image of an existing document. In addition, the off-line case is one that corresponds to the classic reading task performed by humans. In this case, all temporal information about the order of the points of the plot is lost. In addition, one must also consider the problem of variability of the thickness and the form of the lines of writing especially the cursive writing [11].

Without prejudging here, the difficulty of one case compared to the other, we just can see that in the online recognition case, the results are often better for similar conditions of experimentation (vocabulary size, number of writers, etc.). This comes from temporal information that provides valuable insights into the dynamics, velocity and morphology of writing [22].

| Comparison Criterion | Online | Offline |
|---|---|---|
| Acquisition Tools | electronic pen and graphic tablet computer | scanner or camera |
| Picture Noise | low | existence of a significant noise |
| Information Available | • the position<br>• the direction of the movement<br>• the end points<br>• the starting points<br>• order of the features | absence of contextual information |

***Table 1.1:*** *Brief comparison between online and offline approaches [16]*

In order to make a system of character recognition more robust a lot of segmentation method have been developed, despite what they have been reached from development, but the situation is still far from reaching their ambitions. according to the segmentation process there are two recognition approaches have been applied [21].

- Global approach.
- Analytical approach.

## 2.2 Global and Analytical recognition

In this section, we will present the difference between global or analytical recognition. Also, we will see the main advantages and disadvantages of each one.

### 2.2.1 Global approach

In this case, the word is considered as a single entity and it is described independently of the characters that constitute it [21]. This approach has the advantage of keeping the character in its surrounding context, which allows a more effective modeling of the variations of the writing. However, this method is penalizing by the memory size, the calculation time and the complexity of the processing, which increases linearly with the size of the lexicon considered, thus a limitation on vocabulary [1].

### 2.2.2 Analytical approach

In this approach the word is segmented into characters or morphological fragments less than character called graphemes. The recognition of the word consists in recognizing the segmented entities then tending towards a recognition of the word, which is a delicate task that can result in different types of errors **[21]**, a recognition process according to this approach is based on an alternation between two phases:

- The segmentation phase [1,26].
- Segment identification phase [1,26].

Two solutions are then possible:

- Explicit segmentation (external).
- Implicit (internal) segmentation.

Moreover, analytic methods, as opposed to global methods, have the advantage of being able to generalize to the recognition of a vocabulary without limits, because the number of characters is naturally finite. In addition, the extraction of primitives is easier on a character than on a string [16, 21].

## 3 RECOGNITION PROCESS

From various forms of writing to decision-making phase there are several steps are essential to make a recognition process, the figure 1.1 illustrate the process of writing recognition

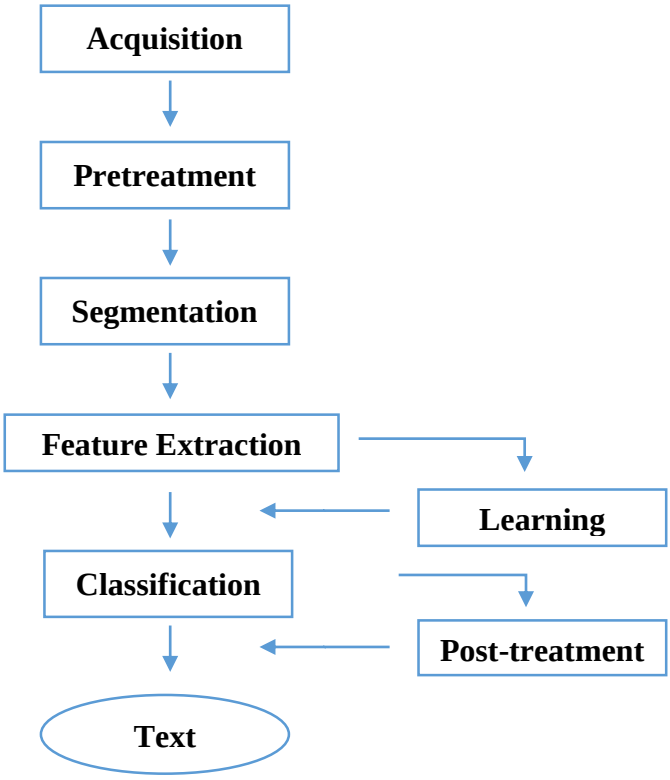



***Figure 1.1:*** *General model for OCR systems [22]*

First, a preprocessing step is performed on the acquired image, this step allows to reduce the noise in the image as much as possible for getting the best acquisition of writing, and this noise can adversely effect on the writing in the image. Secondly, it's the segmentation step, in which we try to segment the writing into basic characters or shapes. After that, comes the third step to extracts the characteristics from these basic characters or shapes. The benefit of this third step is to reduce the amount of information and focus on the characteristics that we need in the recognition; the next step is the classification where the system makes decisions by assigning a label to each given character or shape. In this step the primitive's elements that extracted in the previous step are used for determine the segment of the text according to previously established rules, the model which obtained on learning phase we will use it in this phase to determine the character in which class it belongs [22].

The last step is the post-treatment, which can improve the recognition rate by filtering the decisions of the previous step [22].

## 4. OCR PROBLEMS

In this section, we will present the different issues that faces the recognition system of handwriting.

There are a lot of problems which hinder the recognition process, among these problems, we might mention the following [11]:

- Spatial layout of the text
- Number of writers
- Size of the vocabulary

### 4.1 Spatial layout of the text

The classification of Tappert [11] indicates that the presentation of the text can undergo two types of constraints:

1. External which leading to a writing: pre-cased, zoned, guided or general, as following: [1, 11]
    - pre-cased: When the writer must to write inside predefined boxes (ex: slips)
    - zoned: where the writing must be done in well delimited zones
    - guided: in the guided case, the baseline is the border must the writer respect them
    - general: Related to writing in free sites
2. Internal the writer usually leads to discrete, grouped, cursive or mixed writing, it is obvious that the discrete writing is the easiest for processing because of the presence of the space between the letters unlike of cursive writing which is difficult to process because of the unknown boundaries between the letters [1, 11].

### 4.2 Number of writers

The complexity of a recognition system is related also to the number of the writers which can be divided into three types: single-writer, multiple-writer, omni-writer [11,2].

The simplest case is when there is a single-writer. In this case, only one writer can use the system after learning his writing, in multiple writers, the system can recognize the writings of a limited group of potential writers, either after adaptation to the writing of each one or without adaptation, while in omni-writers, the system must be able to generalize its learning to any type of writing. in the cases of the single-writer or multiple-writer it exists an *individualized initial learning* phase, so the system is able to adapt without forgetting the particularities of each individual's writing. If the system is provided with a *permanent learning* mechanism, it will be able to adapt to the next changes of the writings over time. in the case of omni-writer system, the learning can only be very general, which leads to a very high degree of complexity for these systems [26, 2].

### 4.3 Size of vocabulary

We make the difference between limited vocabulary applications (more than 100 words) and those with very extensive vocabularies (less than 100 words). in the first case the complexity is less definite, Because the fewer vocabulary the less memory congestion and in this situation we encourage the use of direct and rapid recognition methods, by systematically scanning of all the words in the lexicon. In the second case, tens of thousands of words forming a dictionary which cause the problems of memory congestion and the problem of the access time to each word. In this situation, the only effective methods that can be considered are the Searching Tree approaches with successive refinement [11, 2].

The Figure 1.2 can summarize the degrees of generality and complexity of a handwriting recognition system, where the origin of the axes corresponds to the simplest and most constrained system at the same time while on the other hand, as much we move away from the origin of the axes as much growing the generality and the complexity [22,2].

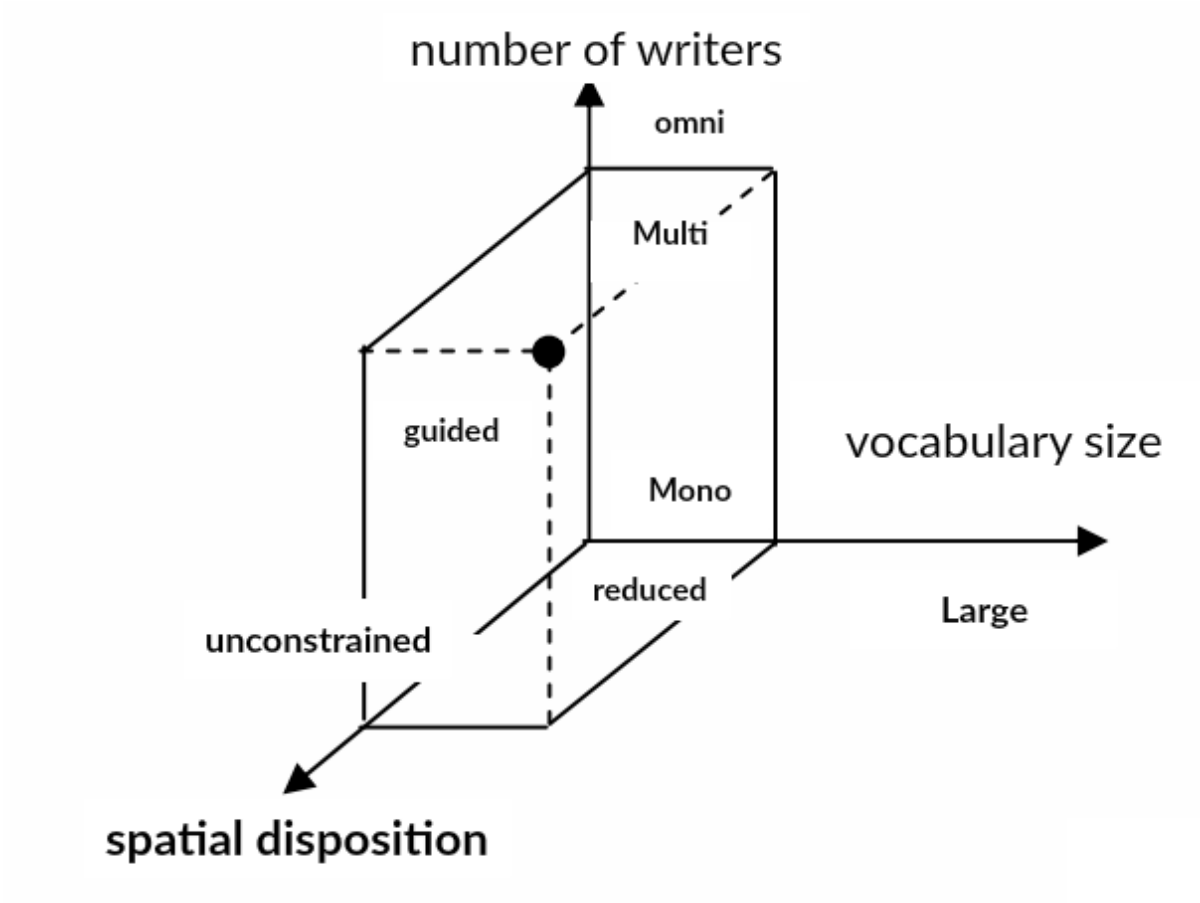



**Figure 1.2:** Graphs of complexity of OCR systems [11].

There are other types of criteria which can influence on the complexity of OCR systems. which are relative with the intrinsic variations of writing, in the context of cursive writing [22], Among these variations, we can mention:

- The physical condition of the paper or the thing which the text was written in.
- The variations specific of the writers

### 4.4 The physical condition of the paper or the thing which the text was written in

- The position of the sheet, of the forearm of the writer [22].
- the angle of the pen relative to the direction of writing can affect the direction of the baseline, which generates fluctuations or inclination of writing [22].

### 4.5 The variations specific of the writers

Variations specific to the writer (that depend on his physical and mental state) reflect the personal style in terms of speed (in the case of fast writing, a lot of letters can merge into a different global forms), continuity and regularity. All these elements influence on the shape of the letters (leaning, curly, rounded, linear, etc.) and of course the shape of the ligatures, sometimes compromising the identification of the boundaries between letters. [22]

Figure 1.3([22]) shows a classification of different types of writing according to the nature of writing and the input mode and the application considered.

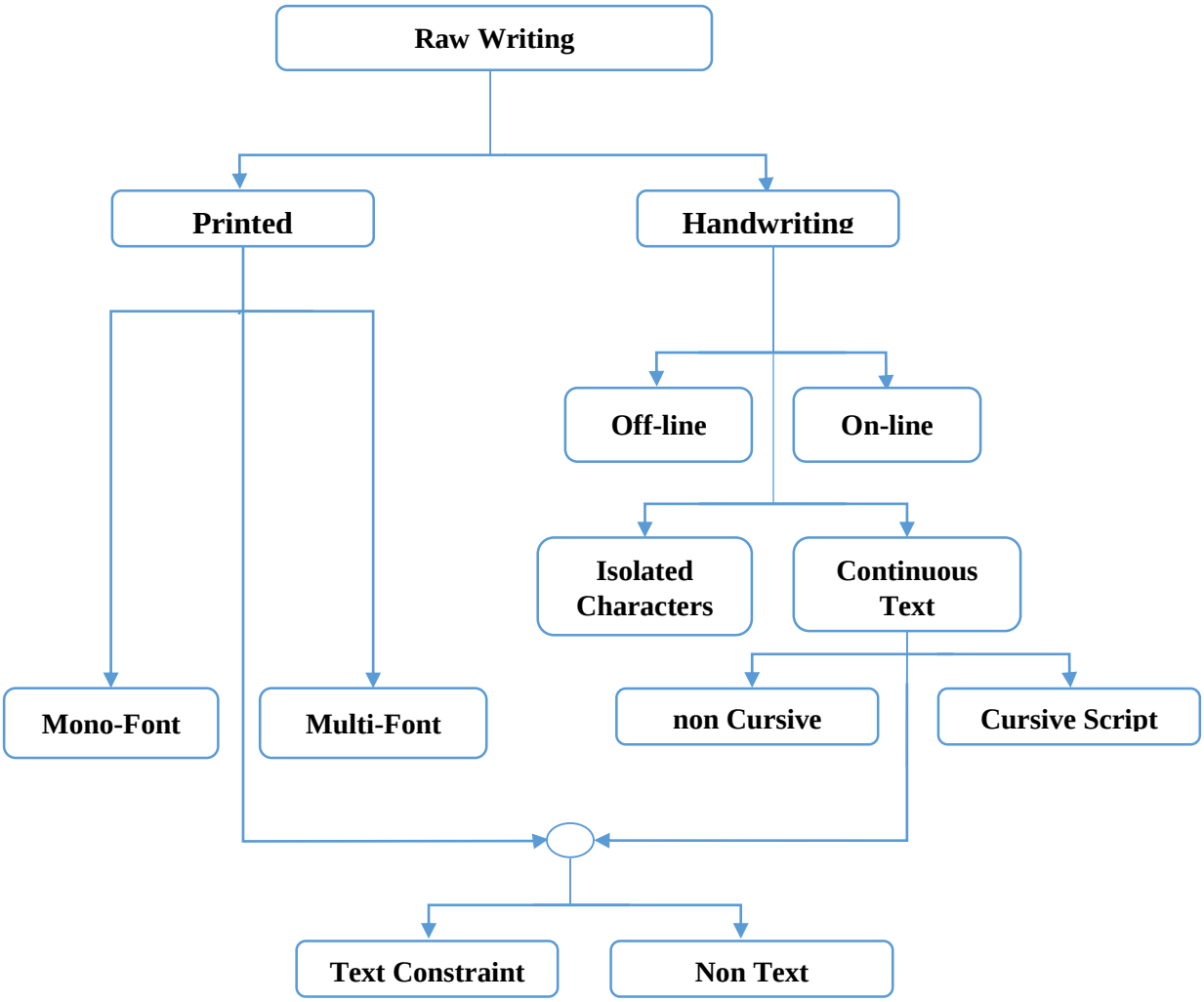



**Figure 1.3:** Dichotomy of different writings [22]**.**

## 5. CHARACTERISTICS OF ARABIC WRITING

Arabic is written by more than 250 million people [28]. Although spoken Arabic varies somewhat from one country to another, the writing system is a standard version used by all the Arab people for their communication. The automation of Arabic character recognition is considered by more than 27 nations who use Arabic characters in their writing [2, 29]

Arabic writing was developed from a type of Aramaic., The Aramaic language has fewer consonants than Arabic, so new letters were created by adding points to already existing letters. the Arabic letters may use other small marks called diacritics to indicate short vowels. Unlike several other languages like the Latin, Chinese, and Japanese scripts that are widely examined, [2, 20] Handwritten Arabic text recognition remains a challenge as there is limited work on it [28]. Arabic script has the following characteristics:

- By nature, the Arabic script is cursive. [2, 29]
- The Arabic text is written from right to left. [2, 29]
- With 28 base letters. [2, 29]
- More than half of the Arabic letters are composed of a main body and secondary components like the letter Jim (جـ) has a point below its main body, and the letter Kha (خـ) has a point above its main body, the letter Yeh (يـ) has two point below its main body. [2, 29]
- The type and position of the secondary component are very important features of Arabic letters. For example, recognizing of the two points below the main body is enough to recognize the letter Yeh (يـ) because it is the only letter with two dots below its main body. In addition, some letters can only be distinguished by the secondary components, for example the letters Fa (فـ) and Qaf (قـ) are different only in the number of points above the main body, and Ta (تـ) median and Yeh (يـ) median are differ only in the position of the two points. [2, 29]
- among the Arabic letters 16 which have points
- The points can be one, two or three, as they can be above or below the baseline

- On the other hand, the width of the Arabic letters differs from one letter to another; in handwritten scripts. In addition, usually there are differences in the form of letters from one writer to another,
- There is a small number of letters that have the same shape whatever the position like: (و) and (ر). [2, 29]
- Most letters of Arabic words are linked together depending on their position
    - In the start of the word.
    - In the middle of the word.
    - In the end of the word.

  The table below (Table 1.2) show to us changing of the shape of the character according to its position in the word. [2]
- During writing an Arabic letters the pen may need more than one direction. the Figure 1.4 illustrate the variation of the directions during the writing. [21]

***Figure 1.4:*** *Directions rolled during writing [21]*

| character | position | | | |
|---|---|---|---|---|
| | initial | median | final | isolated |
| Alif | ـا | ـا | ا | ا |
| Beh | بـ | ـبـ | ـب | ب |
| Teh | تـ | ـتـ | ـت | ت |
| Theh | ثـ | ـثـ | ـث | ث |
| Jim | جـ | ـجـ | ـج | ج |
| Ha | حـ | ـحـ | ـح | ح |
| Kha | خـ | ـخـ | ـخ | خ |
| Del | د | ـد | ـد | د |
| Thel | ذ | ـذ | ـذ | ذ |
| Ra | ر | ـر | ـر | ر |
| Zey | ز | ـز | ـز | ز |
| Sin | سـ | ـسـ | ـس | س |
| Chin | شـ | ـشـ | ـش | ش |
| Sad | صـ | ـصـ | ـص | ص |
| Dhad | ضـ | ـضـ | ـض | ض |
| Tad | طـ | ـطـ | ـط | ط |
| Dha | ظـ | ـظـ | ـظ | ظ |
| Ayn | عـ | ـعـ | ـع | ع |
| Ghayn | غـ | ـغـ | ـغ | غ |
| Fa | فـ | ـفـ | ـف | ف |
| Qaf | قـ | ـقـ | ـق | ق |
| Kaf | كـ | ـكـ | ـك | ك |
| Lam | لـ | ـلـ | ـل | ل |
| Mim | مـ | ـمـ | ـم | م |
| Noun | نـ | ـنـ | ـن | ن |
| He | هـ | ـهـ | ـه | هـ |
| Waw | ـو | ـو | ـو | و |
| Ya | ـي | ـيـ | ـي | ي |

***Table 1.2:*** *The different forms of characters according to their position in the word [2].*

The Arabic words has three different types either with isolated characters "وادي" such as or with linked characters such as "محمد" or with pseudo-words such as "نصائح" has two pseudo-words the first one "نصا" and the second "ئح".

- some Arabic letters have alhamzeh (zigzag shape) "ء" and almad "~" as shown in the next tables (table 1.3).

**(a)**

| character | position | | | |
|---|---|---|---|---|
| | initial | median | final | isolated |
| Alif + med | | | | آ |
| Alif + hamza | | | ـأ | أ |
| Teh | | | ـإ | إ |
| Waw + hamza | | | ـؤ | ؤ |
| Ya + hmza | | ـئـ | ـئ | ئ |

**(b)**

| character | position | | | |
|---|---|---|---|---|
| | initial | median | final | isolated |
| Ta | | | ـة | ة |
| Lamalif | | | | لا |

**(c)**

| character | position | | | |
|---|---|---|---|---|
| | initial | median | final | isolated |
| Lamalif+ med | | | | |
| Lamalif | | | | |
| +hamza | | | | |

***Table 1.3:*** *(a) The additional characters, (b) and (c) Hamza and Med and the positions they occupy with Alif, Waw and Ya [2]***.**

- In some fonts there are many Characters may have written in a combined way, these combinations or ligatures either vertical ligatures which are optional or horizontal ligatures which are obligatory [1], the Vertical ligatures are used for aesthetic and for artistic quality of the document.

| | | | |
|---|---|---|---|
| {ـجـ ,ـحـ ,ـخـ }قـ | {ـجـ ,ـحـ ,ـخـ }فـ | {ـجـ ,ـحـ ,ـخـ }لـ | {ـجـ ,ـحـ ,ـخـ }مـ |

***Table 1.4:*** *Characteristics that can be ligated vertically [30]*

ل م ج ة :disjointed letters.

**لمجة:**obligation ligatures.

**لمجة** : aesthetic ligature between the first two letters.

**لمجة**: aesthetic ligature between the first three letters.

***Figure 1.5:*** *Horizontal and vertical ligatures. [1]***.**

## 6. BRIEF OVERVIEW OF AOCR

Optical Character Recognition or OCR is great invention in the present era and very helpful in our daily use like in the apps of translate for English and a lot of other language. The research of OCR is still under development for other languages such as Arabic language, but cannot generalize the algorithm of the Latins languages on the Arabic language, for this case there is an extension of OCR is AOCR (Arabic Optical Character Recognition) [16].

The recognition of Arabic writing (AOCR: Arabic OCR) dates back to the 70's, since then, many solutions have been proposed. They are as varied as those used in Latin [20]. From the first works in the recognition of Arabic writing was taken into account the two modes static and dynamic. The whole research was all focus was in the field of handwriting and print writing [21].

Despite the huge development and what the evolution in AOCR Arabic language recognition technologies reached, it remains insufficient and almost unsuitable for artistic Arabic writing because it is very difficult to identify words and pseudo-words and the boundaries between characters with the existence of overlaps between characters and characters in vertical ligature and intersections between segments of letters ... etc. [4]

## CONCLUSION

In this chapter we have presented some general concepts related to optical character recognition, and the different aspects of the OCR as well as the organization general recognition system, then we have briefly mentioned the different stages involved in designing a character recognition system.

In addition, we presented the main problems encountered by the OCR, and their impact on the complexity. Then we see the main morphological and typographic properties of Arabic writing.

# Chapter II

# SEGMENTATION

## INTRODUCTION

In the previous chapter, we presented the OCR (Optical character recognition), then we presented the different approaches, methods and techniques in the OCR field.

The purpose of this chapter, to present the most important step that is the segmentation of handwritten and present the approaches used in the segmentation of handwritten text especially the segmentation of the word in characters as well as the strategy that is used.

## 1. PROBLEMS POSED BY WRITING IN CURSIVE

The image that contain a word which written in cursive is represented by a two-dimensional signal where no scheduling information is present, in this word there exists a sequence of letters whose logical order of interpretation is from right to left by convention [2].

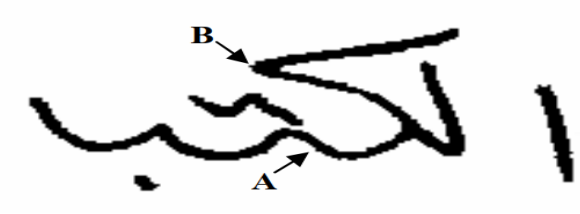



***Figure 2.1:*** *cursive word [3].*

As shown in the figure 2.1 we cannot easily claim that the point B is whether or not it is in a letter prior to point A.

The only way to restore the consistent order of the path is to segment the word into letters. The problem in this kind of writing is that the difficulty of determining precisely the location of the beginning and the end of a letter is so difficult [3].

## 2. SEGMENTATION OF DOCUMENT

In offline recognition of handwriting, the form of the data which being processed is a scanned or photographed image. The segmentation is a very important step in the handwriting recognition process that consist to segment the text image (which is existing in image) into

graphemes, letters or sub-words which is so useful for next processing. The segmentation has a great influence on the recognition rate, which means that the more the segmentation is good and precise the more recognition rate is good. On the other hand, if the segmentation was bidding it will lead to a bed recognition rate [4].

Generally, there are four levels of segmentation, as following:

- segmentation of the page,
- segmentation of text into lines,
- segmentation of lines into words,
- segmentation of words into characters.

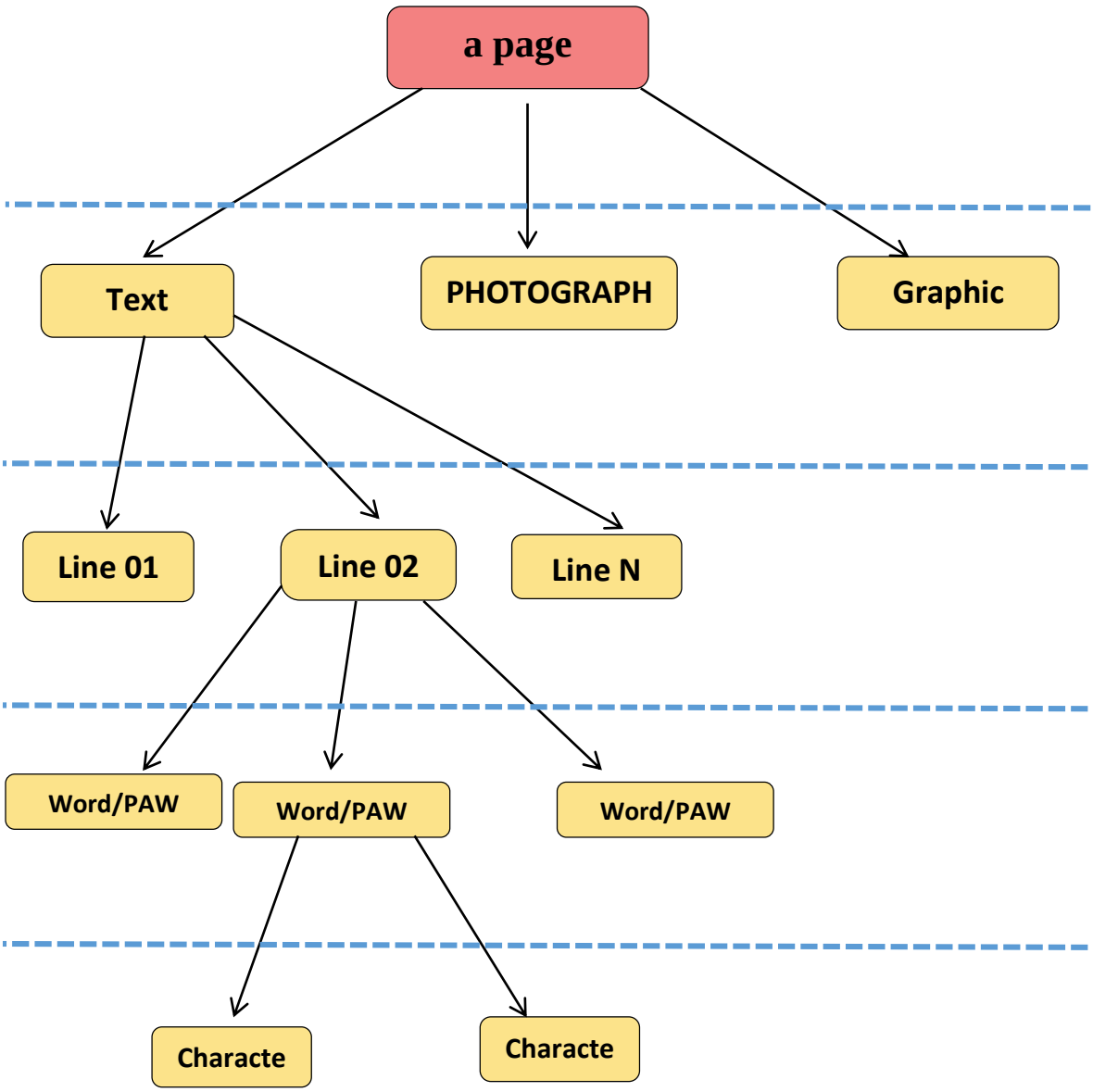



**Figure 2.2:** Illustration of the different levels in the segmentation process [7].

The first level allows the identification of different information areas according to their physical appearance, the functionality of the second level is to separate the lines of the text, and in the third level the word will be isolated from the sentence as the writer write, the last level allows to determine the location of the characters in each isolated word [5].

Each segmented character is validated instantly by recognition. The parts of the word that are not recognized are considered to be cursive letters and it sent to specialized module in the recognition of this type of fonts [6].

### 2.1 Segmentation of the page

This step allows to determine in every page, the information areas according to their physical appearance, It is generally associated with logical labeling, which consists of determining the nature of the media represented in each zone (text, graphic, photograph, etc.). [7], This classification allows to orient the recognition towards systems specialized in the analysis of each type of media [8].

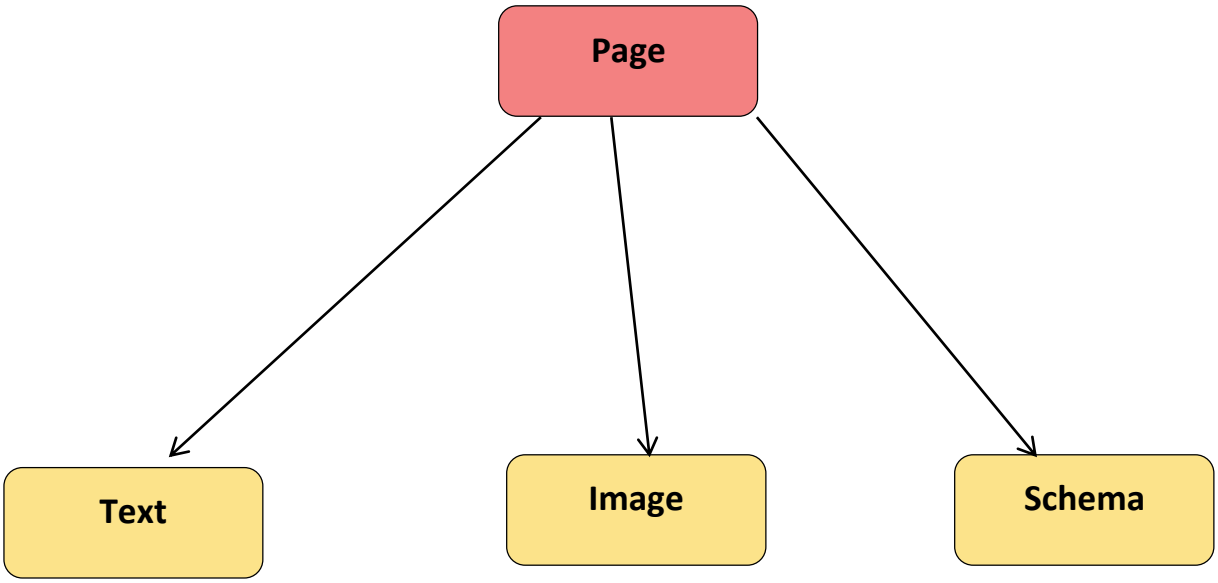



**Figure 2.3:** Detection of different areas of a document page [7].

## 2.2 Segmentation of text into lines

In this step the system separates the different lines of the text for extracting the words. Then the characters which consisting the words. Most of the studies proposed in this field are based on decomposition of the image into connected components [9].

However, others use techniques that Largely based on the histograms of horizontal projections [9], and Some authors choose specialized methods for line segmentation of Arabic handwriting [2]

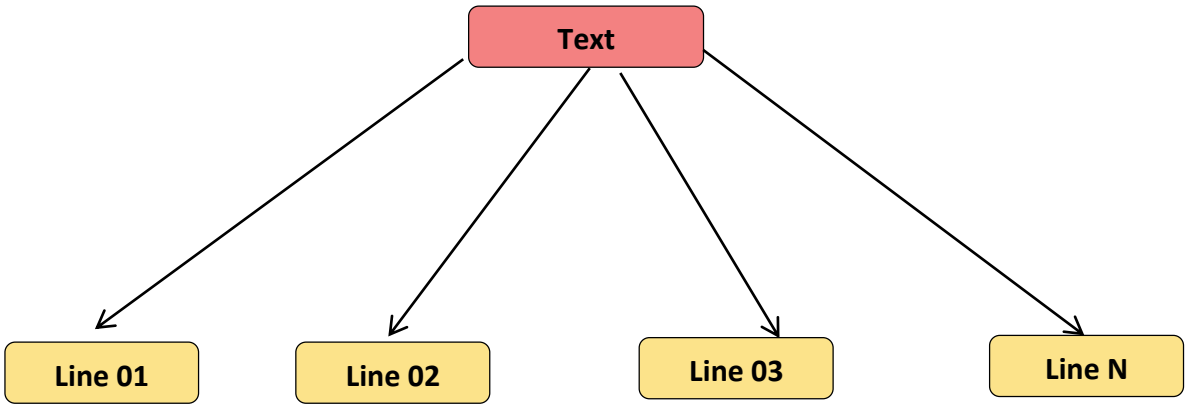



**Figure 2.4:** Segmentation of text into lines [7].

## 2.3 Segmentation of lines into words

This step separates the different lines of text to extract the words and then the letters that make up the words. Most of the proposed studies in this area rely on the analysis of the image into connected components [9]. On the other hand, others use techniques based largely on the histograms of the horizontal projections [10]

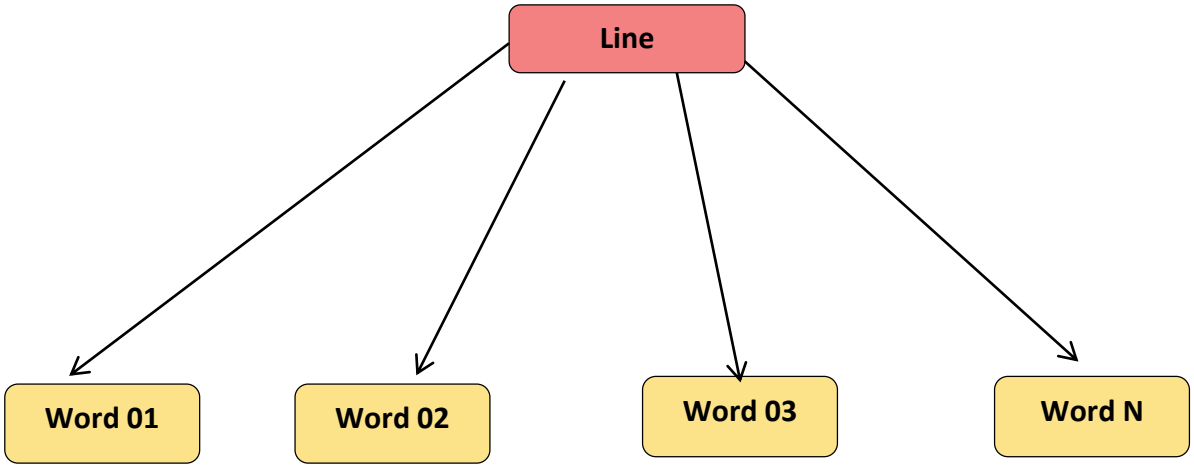



***Figure 2.5:*** *Segmentation of Line into Words. [7].*

### 3.4 segmentation of words into characters

This segmentation segment words into characters (or graphemes) which is the most delicate step in the whole process of a handwriting recognition system. An error in the character segmentation will produce errors in the recognition [4], The goal of this segmentation is to decide whether an isolated motif of an image (character or other identifiable entity of the word) is correct or not [8].

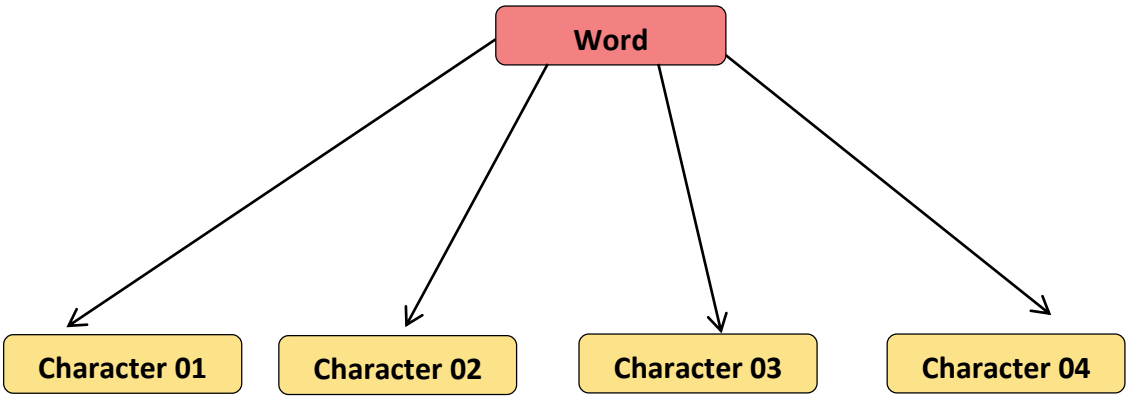



***Figure 2.6:*** *segmentation of words into characters [7].*

## 3. SEGMENTATION STRATEGIES

Tappet and some of author use the terms interne and extern segmentation, depending on whether the segmentation is done separately or simultaneously with the recognition. Dann and Wong and other authors talk about straight segmentation and segmentation-based recognition, to express the same previous meaning [1].

According to what Casey and Lecolinet said that the classification of the methods follows the use or un-use of the recognition during the segmentation phase is not a good classification [1, 2]. Because we can for example use a spelling checker as a post-processor and in this case it can suggest to substitute a letter output by the classifier by two letters, and this is actually a use of a segmentation of the sub-image [11].

According to the point view of Casey and Lecolinet to distinguish between these methods it is better to be according to how segmentation and classification interact throughout the process, like in the previous example the segmentation in two event the first one before the classification and the second one after the classification [2].

After testing the methods, we can classify them into three basic segmentation strategies, adding to the them the hybrid methods based on the last three basic strategies [9,2]:

- **Explicit analytical approach:** Where the segments are identified based on the properties of letter similarity. This approach uses a technique of cutting the image into significant components which is called "*dissection"*.

- **Implicit analytical approach:** Where the system search for components which correspond on the alphabet in the image.

- **Global approach:** In this approach the system tries to recognize the whole word instead of character by character.

In addition to these approaches the hybrids approaches where each of them has different ratios of these three basic approaches can be also used.

### 3.1 Explicit analytical approach

The way that this classical approach works, is to segment and recognize the characteristic elements close to letters (which is called graphemes) [12]. In this case there is an explicit pre-segmentation which is often based on perceptive criteria like the location of ligatures (Which is a symbol resulting from the combination of two or more symbols or two or more letters to form a uniform form that expressing the both symbols) between the letters. Then, the graphemes are recognized individually. The identification of the word is completed due to the contribution of contextual information [3].

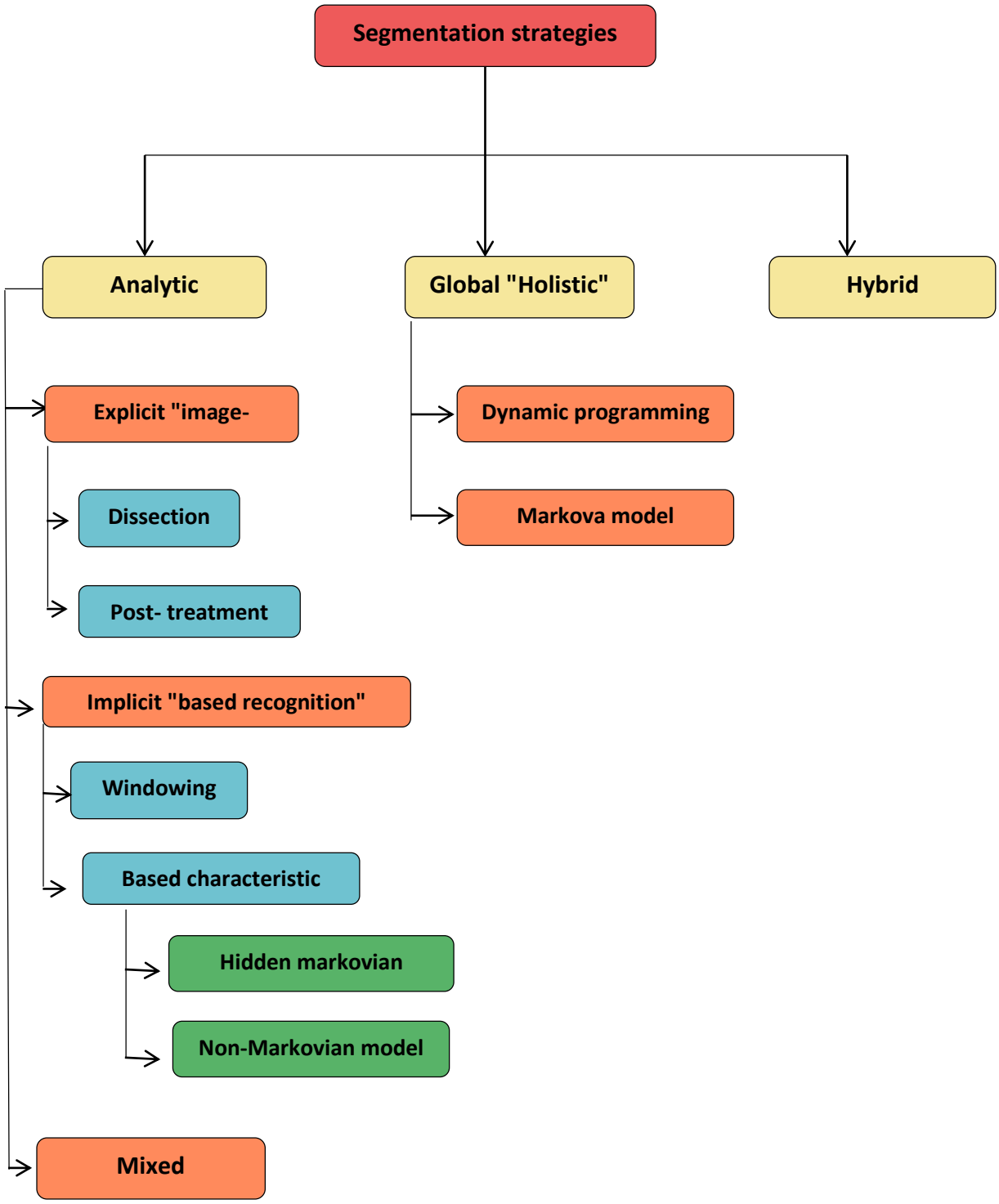



***Figure 2.7:*** *Hierarchy of segmentation methods according to R.G.Casey [3].*

### 3.2 Implicit analytical approach

This approach tries to avoid some of pitfalls of explicit segmentation. In this case, there is no pre-segmentation of word, the segmentation is done during recognition. The systems search in the image for grouping of graphemes or components that correspond to its letter classes. [13] Classically, it can do it in two ways [76]:

- Windowing, the idea is to use a sliding window which has variable width to find sequences of potential segmentation points that will be confirmed or not by character recognition
- Search for primitives, for getting the best recognition as possible as the system can, the system looks for the combination of primitive, this way performs the recognition by using different techniques including the models of Hidden Markov and Neural Network ... etc.

### 3.3 Mixed analytical approach

Actually, the most of current methods are mixture of the two previous strategies [14]. The pre-segmentation is the first step that performed for cutting the image, then we chose the best recognition combination, without forgetting the possible combination 1, 2…. n graphemes [3].

### 3.4 Global approach

In the global approach, the general form of the word is enough to recognize it, this approach supposed to be robust to noise or imperfections of the signal resulting that there is no segmentation (neither explicit nor implicit), usually the recognition depends on a lexicon (a list of word patterns) [2].

The use of this method remain limited because of it requires a model for each word in the lexicon and each model it must to be learned. This method can be used in applications where the lexicon is restricted as it is the case in the automatic processing of checks.
Here too, where different techniques are used such as dynamic programming and hidden Markov models [3].

### 3.4 Hybrid approach

In this approach a lot of recognition strategies is combined resulting in one approach to exploit the points of strength and weakness of complementary approaches (ascending and descending strategies). This approach has a pre-segmentation from which are derived assumptions that will be validated later [3]. Also we can add information about the style of writing - as proposed by Crettez - can help to improve performance by choosing a primitive Extractor specific to each writing style [3]

## 4. SEGMENTATION OF CURSIVE WRITING

Images size is very important in computer science field, and the big images size makes big problem in processing, converting single image to a sequences of pieces (graphemes) raises big size problem, so the size of the graphemes should not be too small to be statistically significant and should not be too big so as not to exceed the size of a letter, where the graphemes are extracted from performing the segmentation operation on image. It is really important that a given grapheme is a subset of a single letter. This condition is necessary to construct a word model as the concatenation of letter models [15].

### 4.1 Segmentation from the skeleton

After generating a skeleton from a given text image, we try to identify certain patterns, to deduce candidate’s points for segmentation. The detection of these patterns introduces calculations of curvatures and angles, which are compared with thresholds adjusted to obtain the desired result [16].

Xavier Dupré confirm that this approach is wrong in about 10% of cases. configurations that are difficult to segment are those for which the letters are often entangled, such as the "tt", or the high-bound letters ('b', 'o', 'v', 'w') with their successor [15].

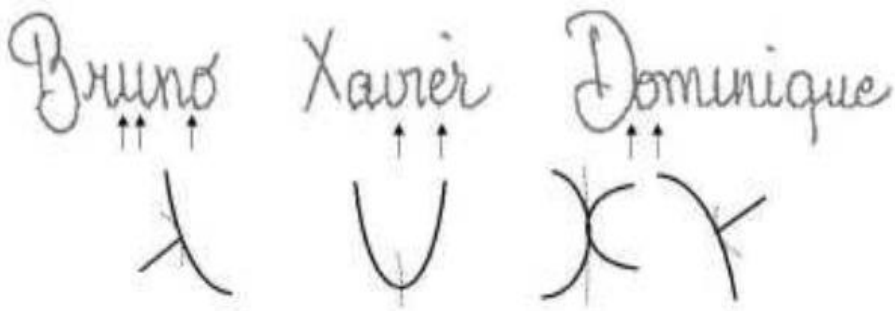



***Figure 2.8:*** *Segmentation based on the skeleton [15].*

### 4.2 Segmentation from the contour

The segmentation from the contour consists in determining the best candidate points of segmentation between graphemes, based on the local extrema of the contour, which are associated according to a proximity criterion (see Figure 2.7) [15].

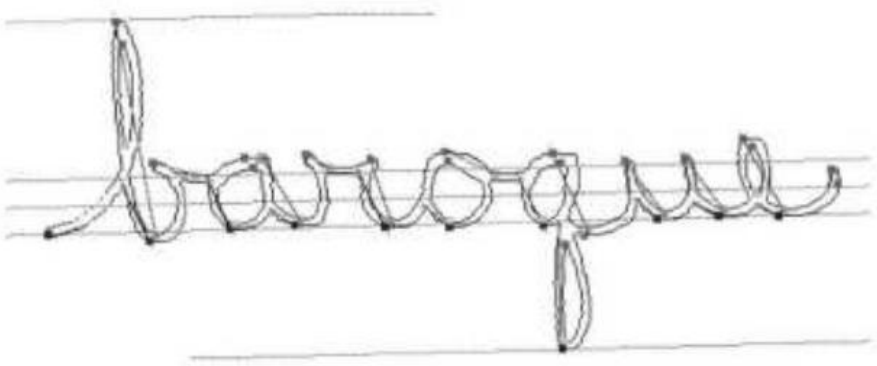


***Figure 2.9:*** *Extrema of the upper and lower contour are associated, and connected by a rope [15].*

Grapheme segmentation from the contour requires many adjustments before finding the decision criteria. This development by trial and error is the common point of many image processing related to handwriting recognition. it is easy to adjust when the quality of the writing is good, these pretreatments can have quite erratic behavior when the writing is of poor quality [15].

### 4.3 Segmentation from histograms

B. Yanikoglu and P. Sandon propose a method of segmentation using histograms [3] It consists of calculating projection histograms in several directions close to the vertical [9]. The chosen lines are those that intercept the least black pixels, with a regular spacing constraint in the image (see Figure 2.10) [10].

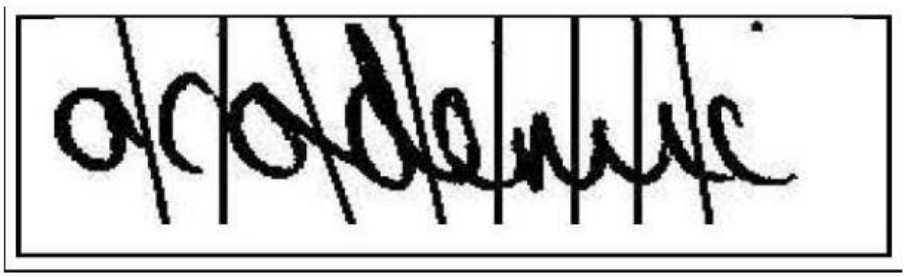


***Figure 2.10:*** *Segmentation from projection histograms in several directions [15].*

### 4.4 Segmentation based on reservoirs

X. Dupré extends to cursive writing the reservoir based on technique initially applied to the segmentation of linked digits. He points out that decision rules are more difficult to implement in the case of letters, because they are of varying sizes. [13].

### 4.5 Segmentation based on slippery windows

The principle is to use a sliding window of variable width by cutting the image into vertical bands. This cutting can be regular or not, possibly with partial recovery of the successive bands (see Figure 2.9). This makes it possible to find sequences of potential segmentation points that will be confirmed or not by the recognition of characters. [9].

By varying the size of the window and its position, we obtain several sequences of segmentation points that will be analyzed by the recognition system. The analysis of the contents

of the window can be done directly on the pixels of the image or can be done on the grouping of low level primitives. This method requires two steps [3]:

1- generation of segmentation hypothesis (sequences of segmentation points obtained by windowing)
2- choice of the best hypothesis of the stage of recognition (validation).

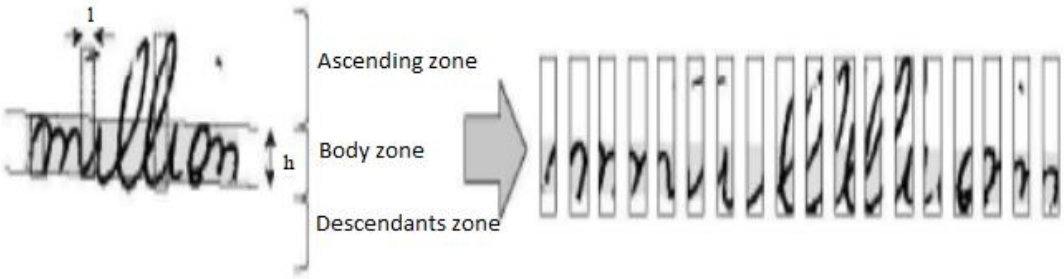



***Figure 2.11:*** *Segmentation based on sliding window: cutting the word into vertical strips [15].*

This technique has the advantage of being simple, robust to noise, and independent of connectivity. However, the width of the window of observation is not easy to determine a priori and it is necessary to manage the conflicts between the various hypotheses envisaged [17]. In addition, the generated sequence of images contains a lot of noise (recovery of two successive letters). This is also true in the case of vertically overlapped letters, but which do not necessarily touch each other like:

- a bar in the letter 't' with the next letter
- descendants such as 'و' and 'ر' in Arabic [15].

## 5.WORD COMPOSITION PROCESS

Composition is the reverse process of segmentation, during which the word is constructed using the different labels (labels) of characters (shapes) obtained after the classification phase in addition to a dictionary containing the models of the words [18].

Each label is compared with the dictionary Word label as long as the sequence of these labels we repeat the operation on the next label otherwise the word considered known and the label being processed is taken for construction a new Word (see Figure 2.10) [19].

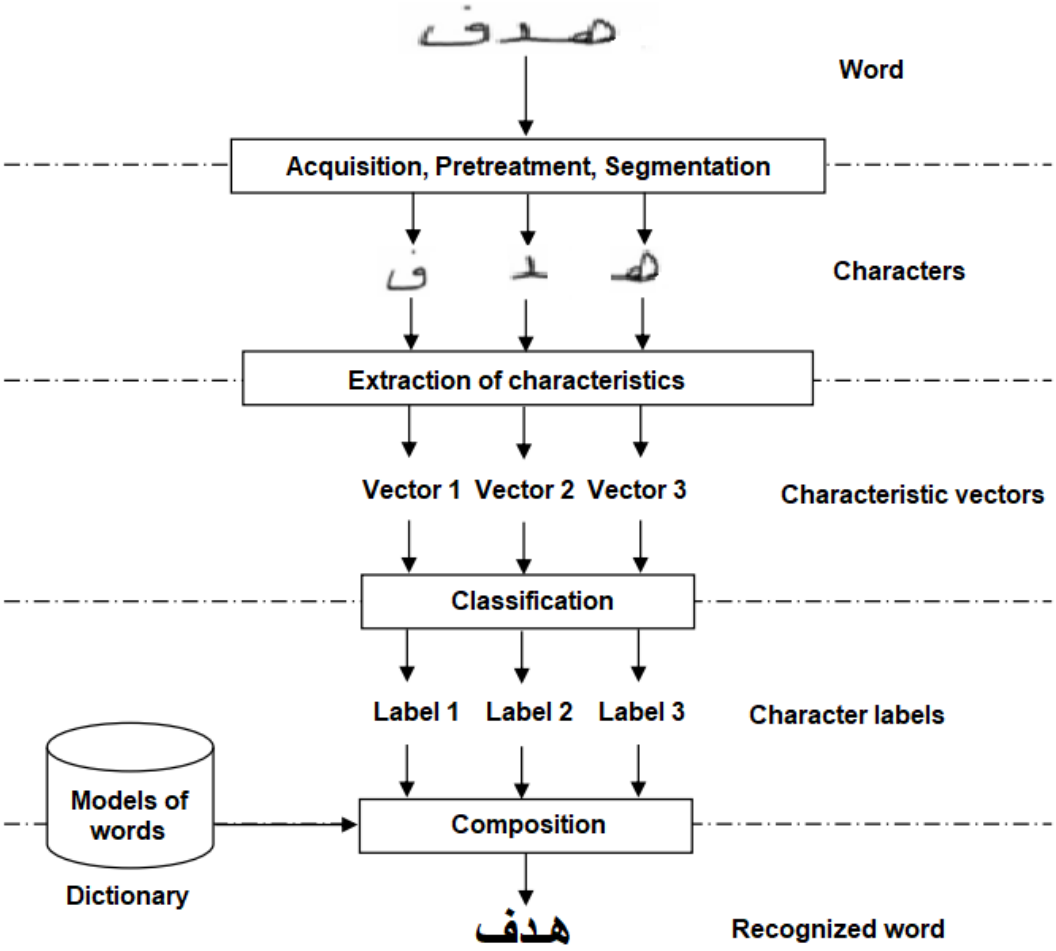



***Figure 2.12:*** *Composition process [7].*

## CONCLUSION

Segmentation is important to properly process the mass of multimedia data passed on throughout the day around the globe.

In this chapter we have presented the different methods used in text segmentation**,** these methods have been developed a lot in recent times but they remain insufficient because the Arabic words manuscripts, sub-words, and the boundaries between the characters are difficult to recognize with the existence of cuts and overlaps in the segments of characters that which makes the segmentation of Arabic documents very difficult.

# Chapter III

# SEGMENTATION AND FEATURE EXTRACTION OF WRITING

## INTRODUCTION

Arabic writing recognition techniques have evolved considerably as a result of many researches, but many problems still exist in the recognition of handwritten Arabic writing, because of the difficulty of recognition of words and semi-words and the overlap of letters and words written vertically and the problem of adhesion between letters, and the problem of irregular spaces between letters.

In this section, we will explain our contribution to solving many of the problems facing Arabic handwriting, and the way algorithms work to solve the problem of intersection of letters and words written vertically, where we have developed many new rules to help divide words correctly.

## 1.THE PROPOSED METHOD OF THE SEGMENTATION OF ARABIC HANDWRITING

In this part we will present the different steps of our proposed segmentation method.

**1.1The flowchart of the proposed method**:

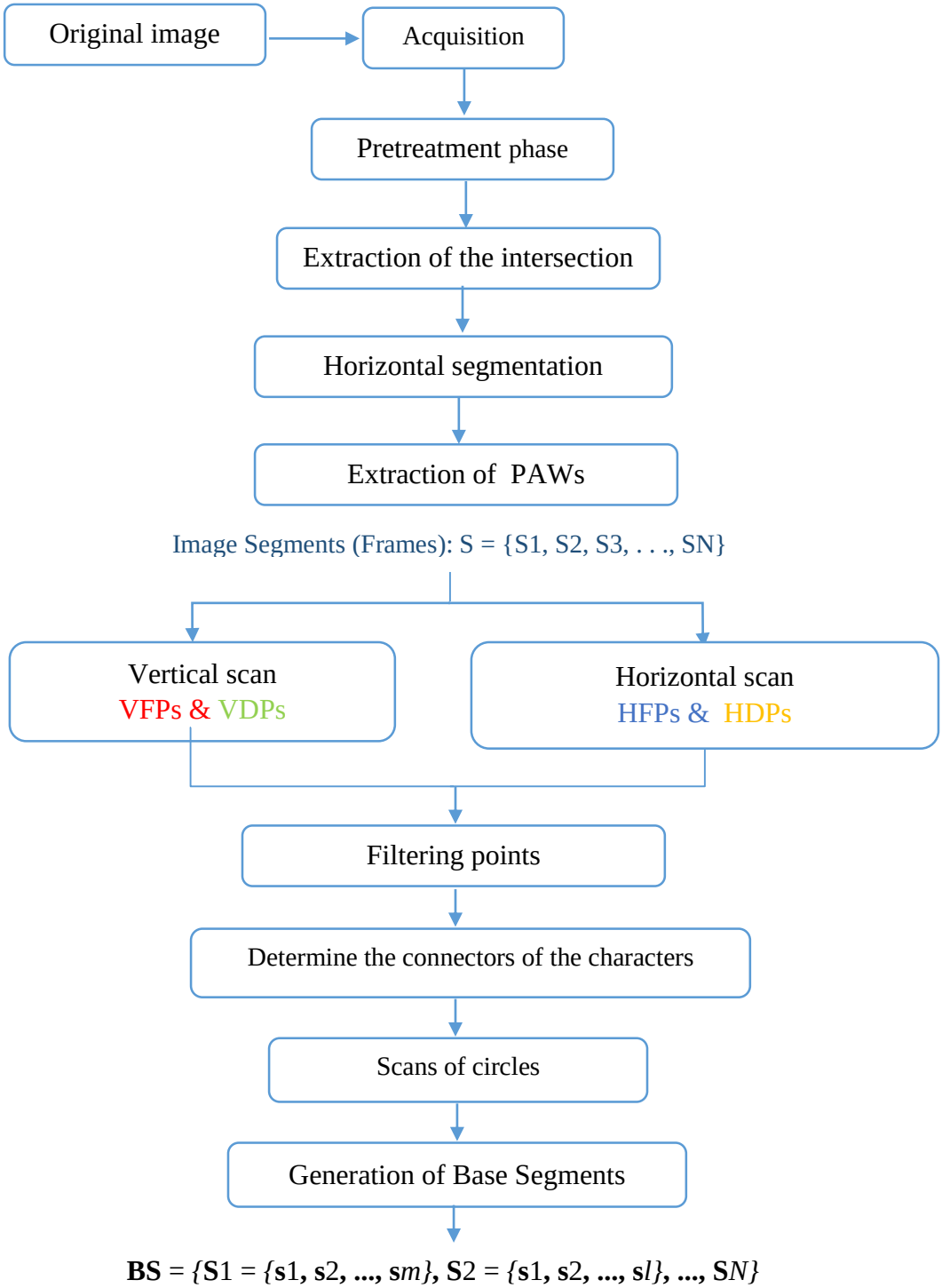



**Figure 3.1:** Organigram illustrates The general principle of the proposed method

## 1.2 The steps and the principle of the proposed method

### *1.2.1 Acquisition Step:*

This step consists to acquire or loading a paper image from physical sensors (Scanner, Camera……etc. or from a "digital image" database) and converting this paper (digital image) into a known image format such as jpeg, gif, bmp…etc. Which can be in color or Grayscale or transform it into a matrix of pixels with a minimum of possible degradation. It's very important step and an influencer step (because can influence on the next step).

### *1.2.2 Pretreatment Step:*

Pretreatment step consist to improving the quality of the information that acquired form the obtained image from acquisition phase (which is a raw image, a noisy image) for prepare the image for the next step (the segmentation). During the pretreatment phase we are going to perform filters for reducing the noise which is superimposed on the data and for trying to keep only the significant information of the shape represented, the superimposed noise may be due to the acquisition conditions such as lighting, incorrect document layout, incorrect document placement or the quality of the original document. This step divided into two main steps, namely:

a) Step of Binarization
b) Step of Filtering and Noise Removal

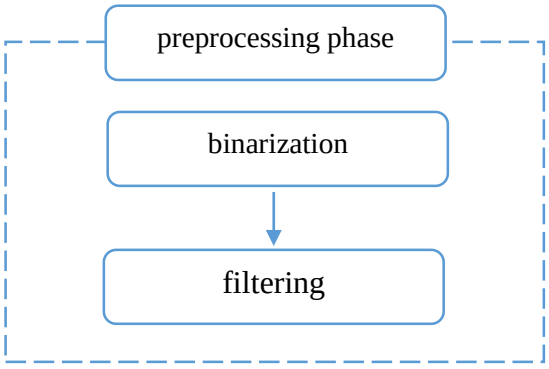



***Figure 3.2:*** Steps of preprocessing phase.

#### *a) Binarization*

The idea of binarization is to transform the raw image which can be grayscale or colored image that acquired from the previous phase into image presented of two main color black and white. In this

case we choose the Global Thresholding method which is fast method, the way that this method works is to take every pixel and compare it with the threshold and decide to take the value either white or black according to weather is higher or lower. This transformation uses the RGB/YUV conversion formulas, as follows:

$$Y = 299 \times R + 0.587 \times G + 0.114 \times B$$

$$U = 0.492 \times (B - Y)$$

$$V = 0.877 \times (R - Y)$$

***b) Step of Filtering and Noise Removal***

After the binarization step the obtained image may have a lot of stain (which mean noisy image), to get rid of these stains (noise) or to clean the image from this stains we use the filtering, which uses pixel masks to eliminate the spots of the image to better represent the points of the interest. The process of the system is to eliminate every pixel that respect this two condition:

1) Black pixel surrounded with white pixels
2) A white pixel between two black pixels either horizontally or vertically.

### *1.2.3 Step of Extract and separate the intersection between characters*

Intersection of letters is one of the special features of the Arabic handwritten manuscript to increase the beauty of writing, the horizontal expansion of the letters causes the intersection of the letters due to overlapping with the vertical expansion of the other letters. This intersection causes problems in dividing the word for character recognition, so we suggested a solution to this problem by finding the intersection and extracting from the word, the following steps illustrate the mechanism of work to find and extract the intersection

**a) Vertical projection:** The vertical projection finds possible points for the intersection of characters by finding the points that achieve the following property:

The search for the intersection starts from the bottom corner on the left. As shown in the following picture:

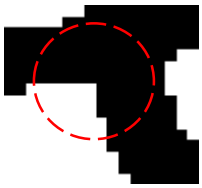


***Figure 3.3:*** The first angle to find the intersection

In order to reduce the potential points of the intersection of letters, we calculate the length of the vertical extension of the letter where the calculation of the length is divided into two parts:

**Part 01:** The length of the vertical extension of the letter above the intersection.
**Part 02:** The length of the vertical extension of the letter below the intersection.

The image below shows the result of the character calculation where the yellow color indicates the character length

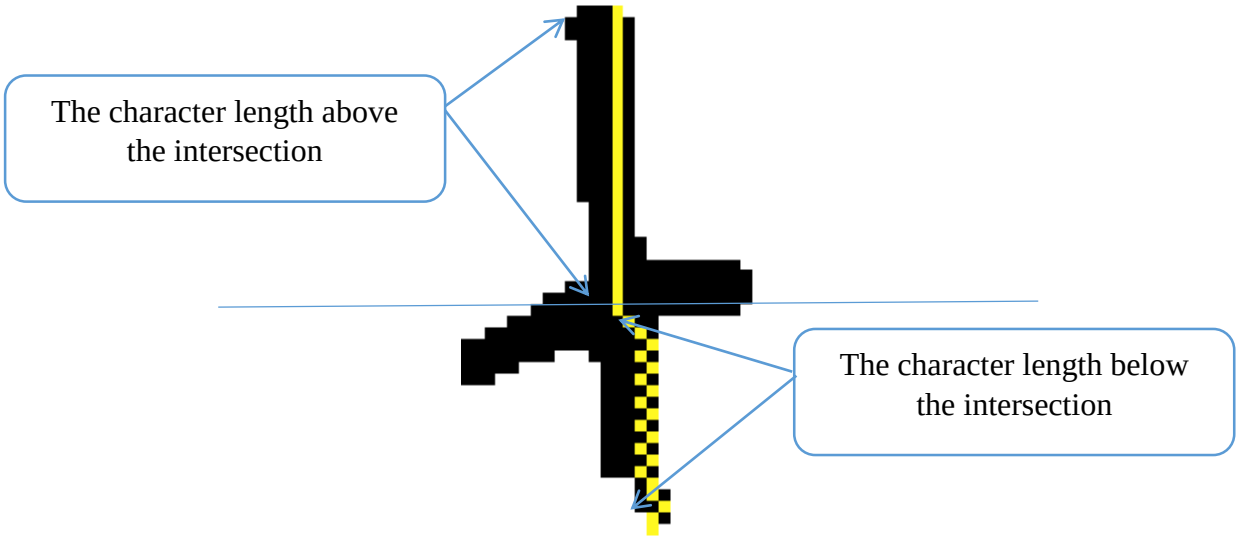



***Figure 3.4:*** Image illustrating character length.

The point (pixels) that meet the conditions mentioned above, we change the color of this point to Magenta.

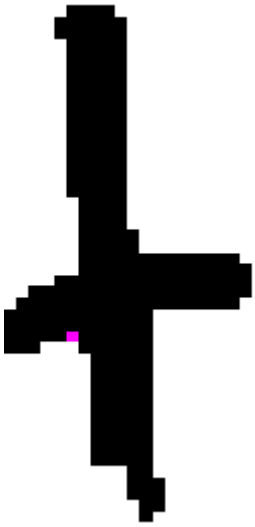


***Figure 3.5:*** An image showing a possible point of an intersection state.

we use the following code to calculate the character length:

**Note:** The following code shows how to calculate the length of the bottom part of the intersection only.we use the code in reverse to calculate the length of the top part of the intersection.

**Where :**
**Img:** source image .
**getRGB (x, y):** function makes it possible to obtain the color components of the pixel in position (x,y).

```
int MiddleY = Y;  int MiddleXBottom = X;
boolean BalanceBottom = false;
int BorderSize = 8;      int Max = 0 ,  Min = 0;
while (img.getRGB(MiddleXBottom, MiddleY + CompteurBottom) != -1) {
  MoveBottomLeft = 0;
  MoveBottomRight = 0;
  Max = 0;
  Min = 0;
            ///////          MoveBottomRight         /////////
   while (img.getRGB(MiddleXBottom + MoveBottomRight, MiddleY + CompteurBottom) != -1) {
               MoveBottomRight = MoveBottomRight + 1;
            }
           ///////          MoveBottomLeft         /////////
    while (img.getRGB(MiddleXBottom - MoveBottomLeft, MiddleY + CompteurBottom) != -1) {
               MoveBottomLeft = MoveBottomLeft + 1;
            }
            CompteurTopRight = CompteurTopRight + 1;
            if (img.getRGB(MiddleXBottom, MiddleY + CompteurBottom) == -1) {
               BalanceBottom = true;
            }
        if (BalanceBottom && (MiddleX -  (MiddleX - MoveBottomLeft)) < 8) {
               Min = MiddleXBottom - MoveBottomLeft;
               Max = MiddleXBottom + MoveBottomRight;
               MiddleXBottom = (int) Math.round((Max + Min) / 2);
            }
```

```
if (MiddleXBottom + 1 < img.getWidth() && MiddleY + CompteurBottom<   img.getHeight()){
   if (img.getRGB(MiddleXBottom + MoveBottomRight, MiddleY + CompteurBottom) == -1) {
                if (img.getRGB(MiddleXBottom + 1, MiddleY + CompteurBottom) != -1) {
                   MiddleXBottom = MiddleXBottom + 1;
                } else {
                   if (img.getRGB(MiddleXBottom - 1, MiddleY + CompteurBottom) != -1) {
                      MiddleXBottom = MiddleXBottom - 1;
                   }
                }
            }
        }
}
```

From this magenta point, we search the other possible angles that form the intersection between the two letters. If there are other angles, we place a point at each corner of the intersection of the two letters. we assign the magenta color to the angles at the top on the right and bottom on the left ,and the cyan color of the angles at the top on the left and bottom on the right. the image below shows the points at the corners of the intersection and its color:

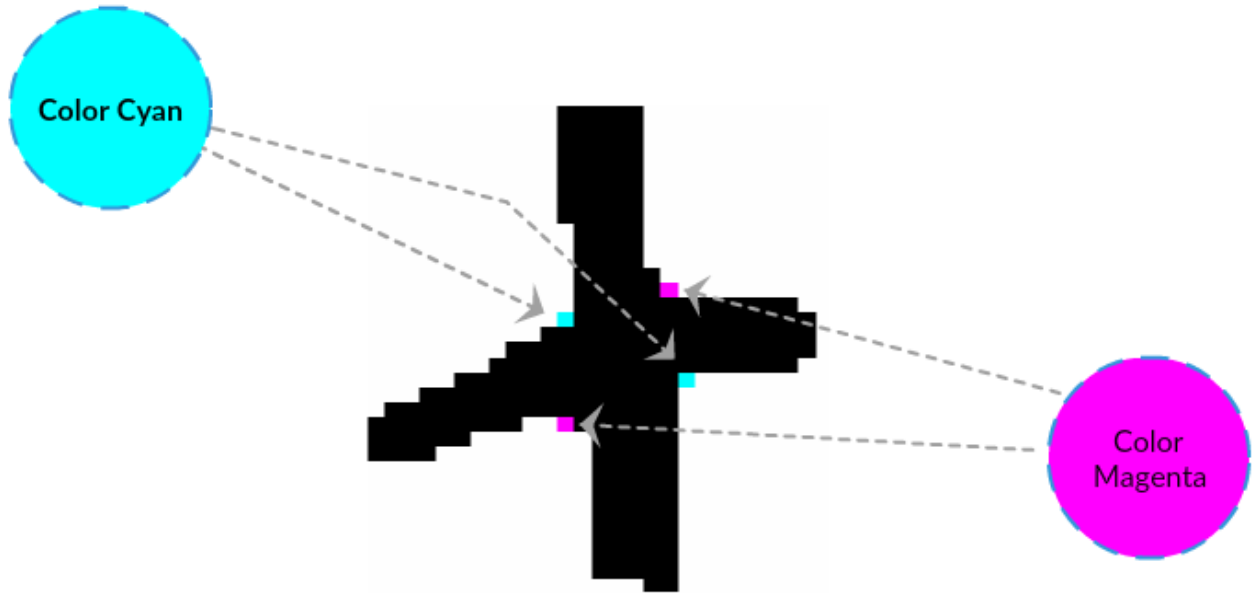



**Figure 3.6:** Picture showing the four points that define the intersection.

After identifying the points at the corners of the intersection we move on to the next step, which is to connect the magenta points with the cyan points, by drawing a line connecting the four points. the cyan lines represent the horizontal border of the letter, while the magenta lines represent the vertical border of the character.Using this code.

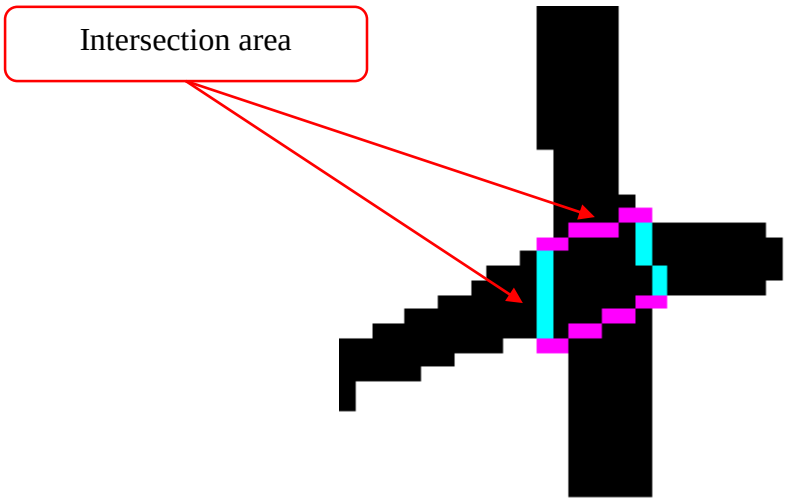



**Figure 3.7:** Image illustrating the intersection area of the two letters.

In order to draw lines between two points we use the following code:

```
File f = new File(PATH);

BufferedImage img= ImageIO.read(f);

Graphics2D graphics2D = img.createGraphics();

graphics2D.drawLine( PointCyanX ,PointCyanY ,  PointMagentaX, PointMagentaY);
```

We select all points with the vertical extension of the letter and put it in a matrix and then turn it into a single image. to get the following picture

***Figure 3.8:****Extract the vertical extension of the intersection of the two letters.*

After extracting the vertical character, the intersection becomes like this image

**Figure 3.9:** Image after extracting the vertical extension of the

In order to retrieve the missing horizontal character points, we fill the white space inside the intersection area and change the blue points to black, and then magenta points to white.

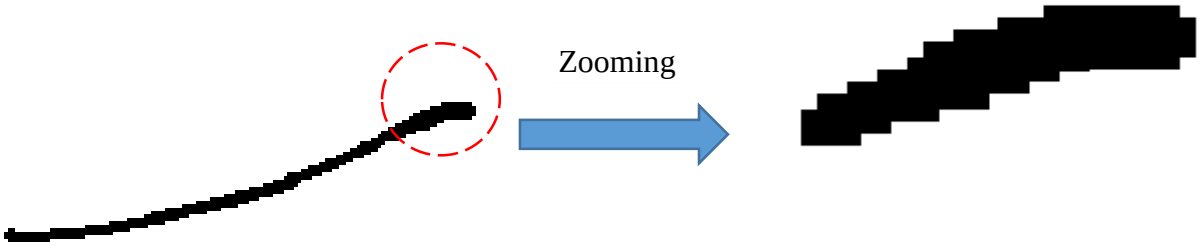



***Figure 3.10:*** Image of the vertical extension of the intersection of the two letters after retrieval of missing pixels.

After extracting the intersection, we get two pictures. an image of the word without the intersection and the image of the vertical extension of the letter, as shown in the table below:

| The word before extracting the intersection | |
|---|---|
|  | |
| **The word after extracting the intersection** | |
|  |  |

**Table 3.1:** Word before and after intersection extraction (الزارات).

Another example of a word that contains an intersection:

| The word before extracting the intersection | |
|---|---|
|  | |
| The word after extracting the intersection | |
|  |  |

**Table 3.2:**Word before and after intersection extraction (الشريفات).

### *1.2.4 Step of Horizontal segmentation:*

Arabic handwriting is characterized by vertical writing, making it difficult to segment ate the word to recognize letters. Because of this, we proposed another algorithm to solve this problem by using the horizontal character separation technique, depending on the many rules to help determine vertical writing. we use the pink point to indicate the possibility of writing vertically. The following image shows the pink point:

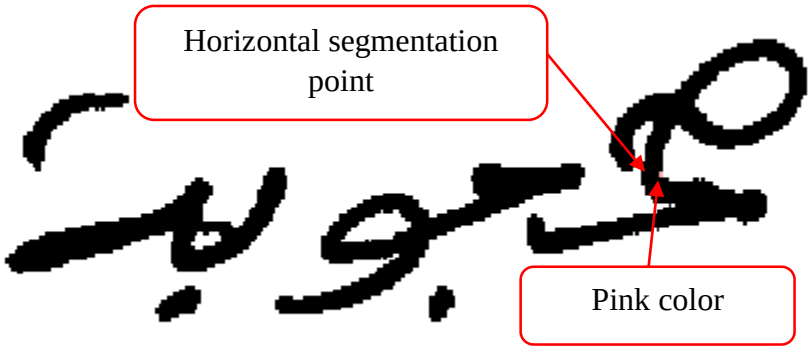



**Figure 3.11:**Picture showing vertical writing of a word...

The following images illustrate some cases of vertical writing :

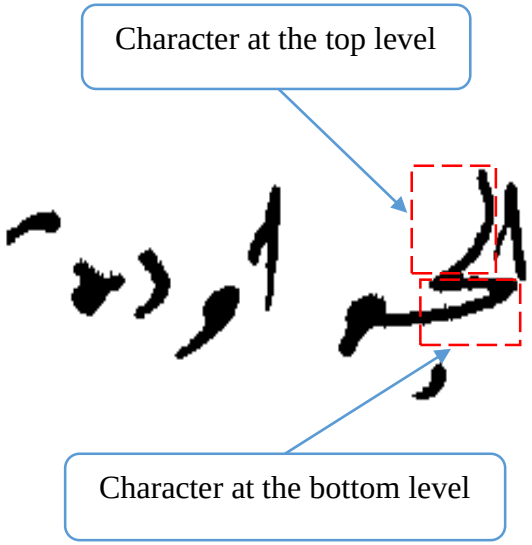



**Figure 3.12:** *Examples illustrating vertical writing of words (ل& ح).*

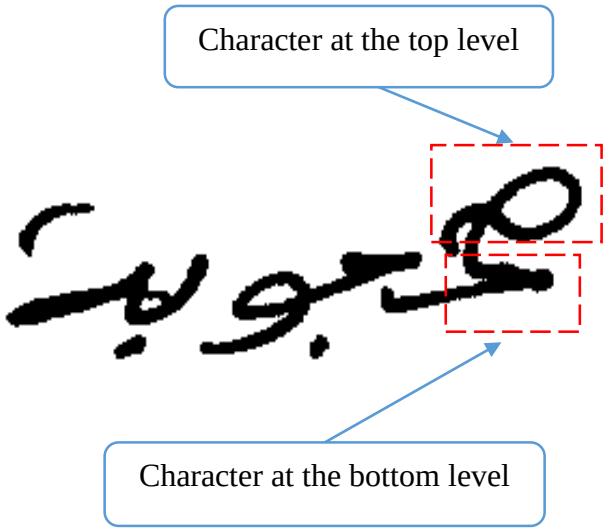



**Figure 3.13**: Examples illustrating vertical writing of words (م & ح).

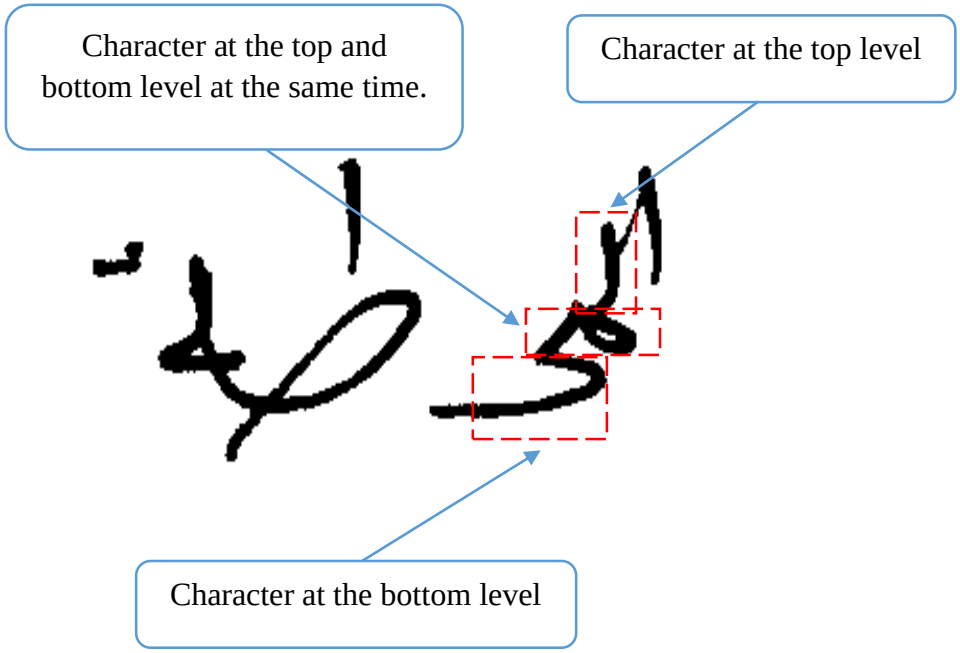



**Figure 3.14:** Examples illustrating vertical writing of words (ل & م & ح).

### 1.2.4.1 Detect writing vertically for word

The word vertical is defined using a control point, where we calculate the length of the vertical extension of the letter at the top level and make sure that there is an angle between the letter at the top

level and the letter at the bottom level belong to the following sectors (2-1-8), and then descend vertically to check the color change feature.

#### 1.2.4.1.1 Calculate the length of the letter extension at the top.

The process of calculating the length of the letter at the top level helps us to verify the existence of the case of vertical writing using a method similar to the method mentioned above.

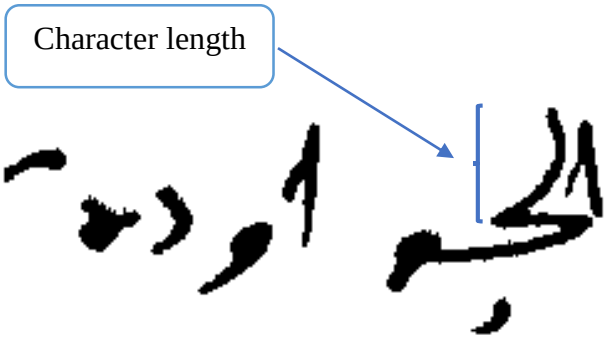



**Figure 3.15:**An image showing the character length at the top

#### 1.2.4.1.2 Feature change colors:

The color change feature allows us to know the words written vertically by using the vertical projection where we first check the black color (the color of the writing), second check for a white space (background color), and third check for the black color again. the following image shows the color change feature.

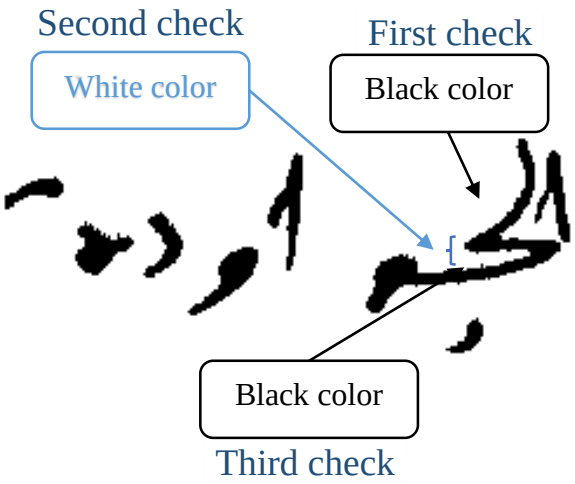


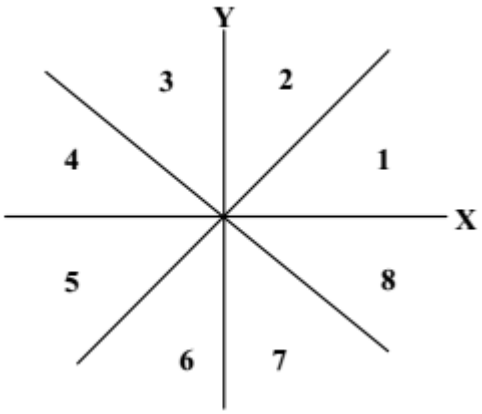

**Figure 3.16:** Illustration of color change feature.

We check the color change feature using the following code:

```
int m = 1;
int N_Color =Color.BLACK.getRGB();
int Comptuer = 0;
int NewX = 0;
h = 1;
 while (Sep_img.getRGB(PointX- h, PointY) == Color.BLACK.getRGB()) {
              NewX= PointX - h;
               h = h + 1;
               }
               if (Sep_img.getRGB(PointX, PointXY+ m) == N_Color && Comptuer == 0) {
                  Comptuer = Comptuer + 1;
                  N_Color = Color.WHITE.getRGB();
              }
             if (Sep_img.getRGB(PointX, PointXY+ m) == N_Color && Comptuer == 1) {
                 N_Color = Color.BLACK.getRGB();
                  Comptuer = Comptuer + 1;
             }
            if (Sep_img.getRGB(PointX, PointY+ m) == N_Color && Comptuer == 2) {
                   N_Color = Color.WHITE.getRGB();
                   Comptuer = Comptuer + 1;
            }
           if (Comptuer == 3) {
                    Is_Ha = true;
            }
 m = m + 1;
 h = 1;
```

**1.2.4.1.3 Sectors of angles :**

Sectors help to get a description of the angle between the letters above and the letter at the bottom. The following image shows the sectors mentioned above




***Figu**re* **3.17***: Illustration of the sectors.*

***1.2.5 Step of Extraction the Different Connected Components:***

To handle this situation and obtain the different connected components, the algorithm we propose will start scanning from the write to the left in order to have the first black point (black pixel), then we use recursive function with the help of the eight neighboring points to handle the rest of the PAW (Piece of Arabic Word), at the end of this step we will generate and extract the different connected components and we will handle the text as a set of segments.

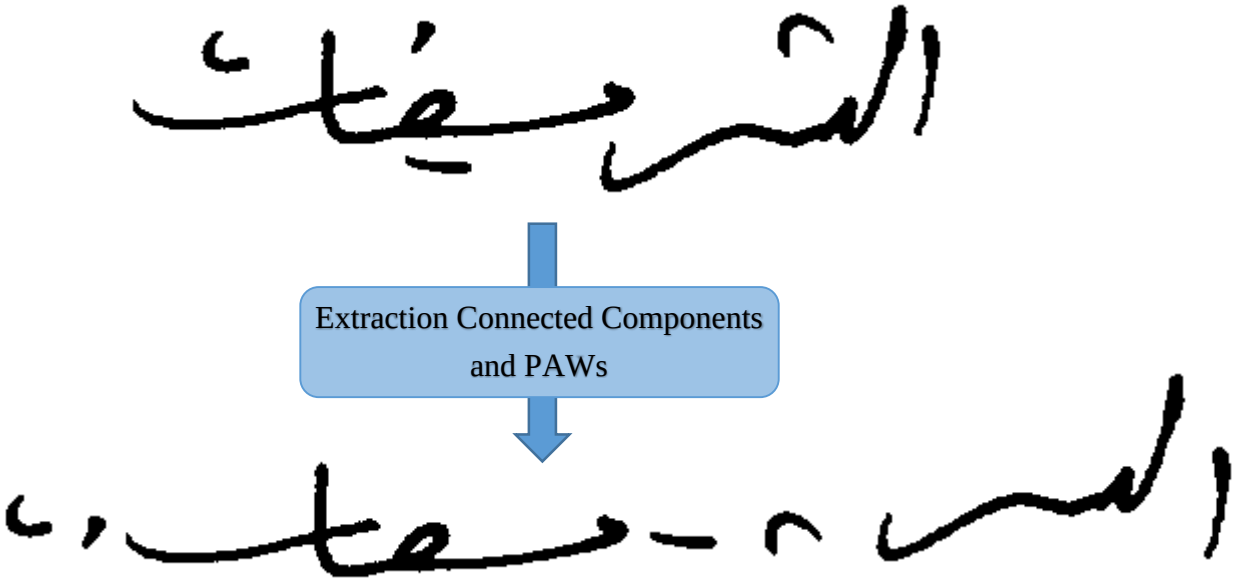



**Figure 3.18:**Example of extracting the different connected. components of a text

this obtained connected components may be more than one letters (PAWs) interconnected by the intersection of the character extensions that compose them, or may be one point or two connected diacritical points of the letters like ‘ش, ‘ث, ‘ت, ‘ب’…. etc. After we obtains the connected components how the system will distinct between them?

To handle this problem, we divide into this cases:

- Connected component that are either PAW or several of PAWs that are interconnected by the intersection of the character extensions that compose them.
- Isolated PAW which mean letter.
- The isolated small pieces are the diacritical points or diacritical marks

Or we distinct between the different connected components by the size and thickness of each of them

### *1.2.6 Step of Determining* Different points of interest:

After we extract the segments in the previous step we are going to determine the points of interest that connected component have shush as Point of Vertical Nature which inform us to detect the connector of the character and the Point of Horizontal Nature which help in identifying the limits

of the character and the appearance of their form, to determine PVNs and PHNs we use a vertical and a horizontal scan and predefined masks.

To determine each of PVN and PHN we use two vertical scanning masks and two horizontal scanning masks respectively, the first two masks has two points are: Vertical Fusion Points (VFPs) which is in red, Vertical Division Points (VDPs) which is in green, (see the figure 3.4) and second two masks has also two points are: Horizontal Fusion Points (HFPs) which is in blue, Horizontal Division Points (HDPs) which is in yellow (see the figure 3.5)

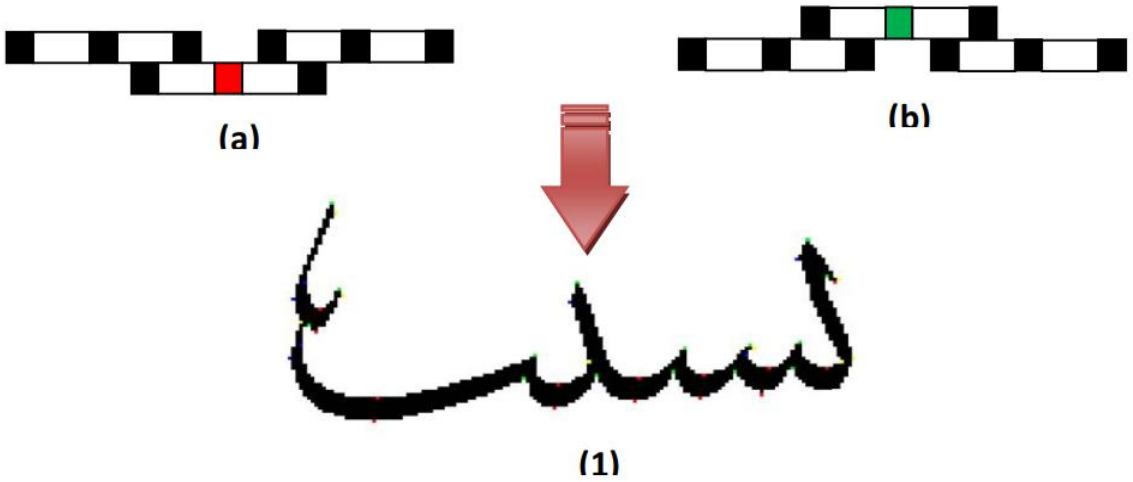



***Figure 3.19:*** *Vertical scan masks ((a) fusion mask (b) division mask) [4].*

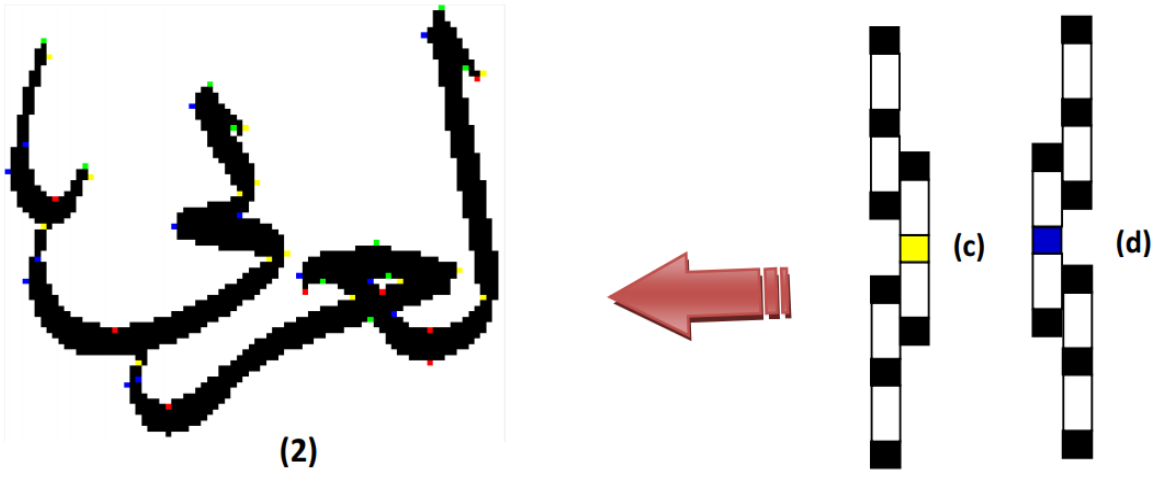



***Figure 3.20:*** *Horizontal scan masks ((c) fusion mask (d) division mask) [4].*

### *1.2.7 Step of Filtering and Eliminating Incorrect Points*

The generated point may be ambiguous and difficult in the segmentation process because of contains an additional point (or incorrect), which requires to eliminate them by a filtering step.

Eliminate the points near each other when T & L < pen_size, such that T is the thickness and L is the length of the lines of the segments.

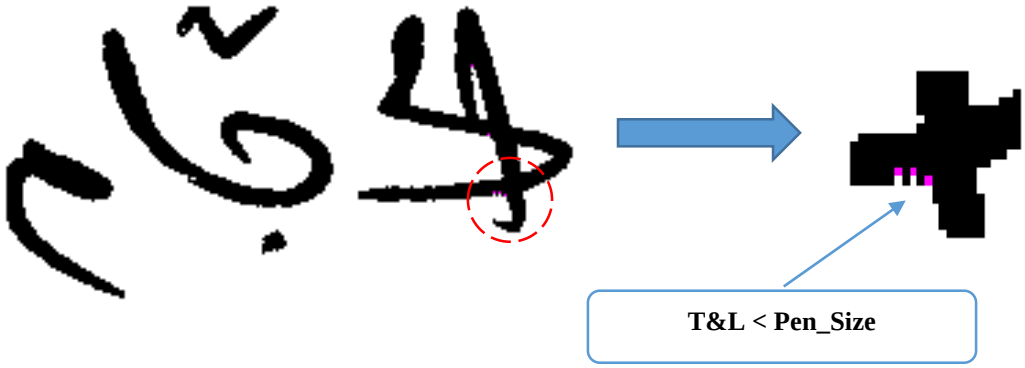



**Figure 3.21:** illustration of points filtering.

### *1.2.8 Step of Discovering circles of characters:*

In this step which is should be done before the segmentation step, it takes every segment that has no simple structure and search for the location of the circle of the character (if it exists) such as the letter 'ق','و','ط', 'ف','ـة' which compose the connected components, try to determine the descendant points that are result of overlapping the character that have descendant extension like 'و', 'ر' with the other characters belongs to the same word or to other pseudo-word, and also try to determine the location of the intersect between the characters that are compose the pseudo-word.

The system distinct the circle by four colored point that are consecutive on its perimeter each one then has its own meaning and fixed order as follows:

- a HDP point which is in yellow it is in the right side of the circle
- a VFP point which is in red at the bottom of the circle
- a HFP point which is in blue on the left side of the circle
- a VDP point which is in green at the top of the circle.

As shown in the following figures, those points that are form the circle can have obtained from the fusion and division function.

***Figure 3.22***: *illustration the determination the circles of the characters.*

## CONCLUSION:

In this chapter, we presented the general structure of the proposed system for the division of handwritten calligraphy, and then we explained all the stages. We also provided our contribution to the development and improvement of the system by providing two algorithms, the first algorithm to extract the intersection of letters, and the second algorithm to divide the writing vertically.

# Chapter IV

# TEST AND RESULTS

# INTRODUCTION

In the previous chapter, we explained the steps used to divide the handwritten Arabic words.e explained the two algorithms in which the first algorithm is based on finding and extracting the intersection. The second algorithm depends on the horizontal segmentation of vertically written words.

In this chapter, we will explain the program interface and the steps used to acquire the image, and then filter the image obtained. Then the stage of converting the image where we get a series of images contains parts of the word processing. Finally, we will display the results obtained by our system.

## 1. Programming Language

In this work, we used the Java programming language, which is one of the most powerful programming languages that contain many powerful libraries that are easy to use, including libraries dealing with images, such as image acquisition, processing and creating new images.

Why java ؟

Java programming language has many advantages, including the following:

Java is easy to learn.

Java is object-oriented.

The portability of software.

Java is platform-independent.

## 2. interface

When we run the program we see this main screen, which contains a pre-loaded image, as shown in the image below.

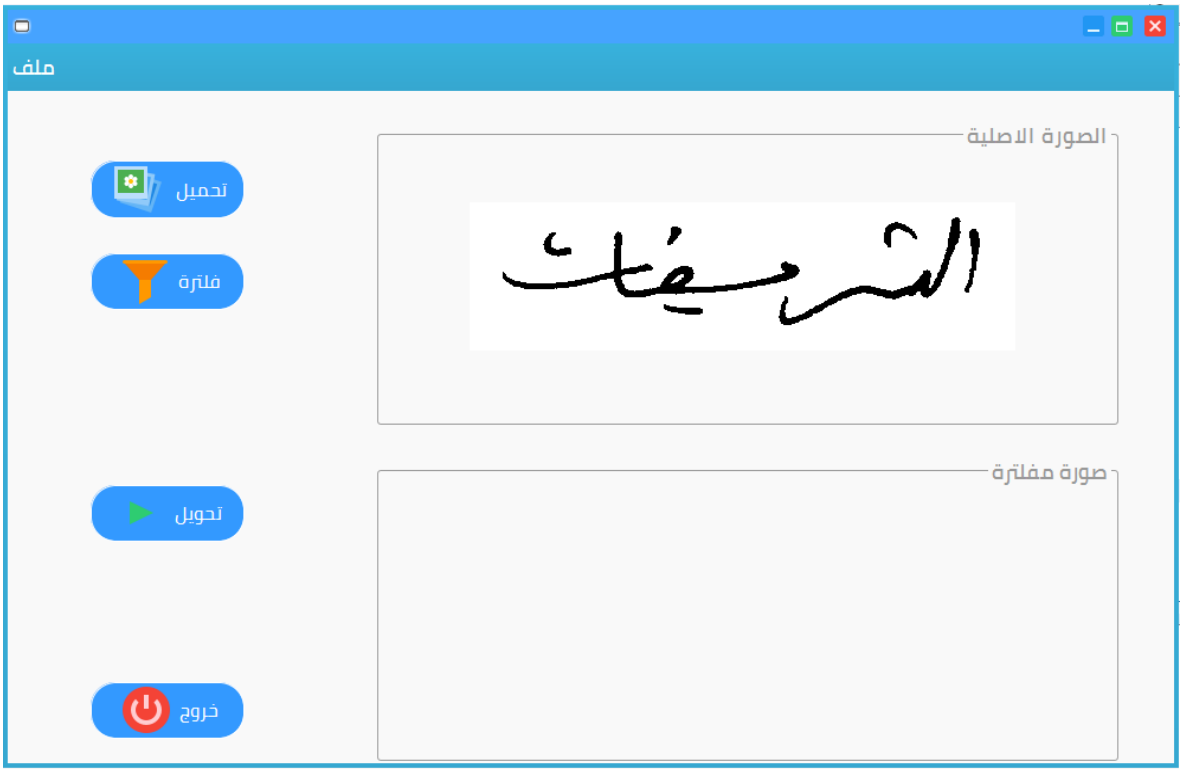



**Figure 4.1:** The main interface of the program.

### 2.1. Load image

Through تحميل button or from the main menu ملف / فتح, we can browse inside the computer files to select the image we want to work on. All files containing the following extensions appear (jpeg, gif, bmp. . . etc).

As shown in the following picture:

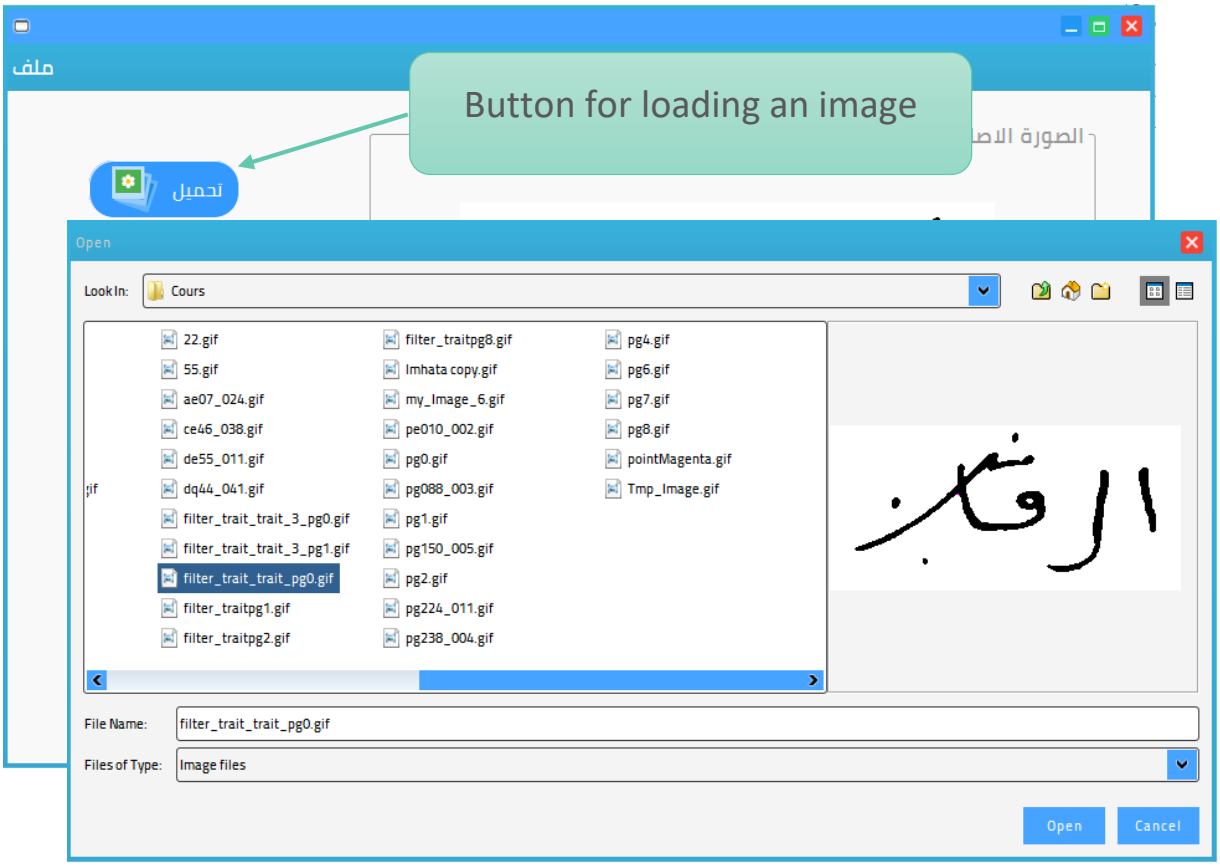



**Figure 4.2:** Example to upload an image**.**

### 2.2 filtering and removing noise

This process is very important for improving image quality and facilitate their analysis. Filtering allows good representation of control points by eliminating noise points.

The following image shows the filtering process:

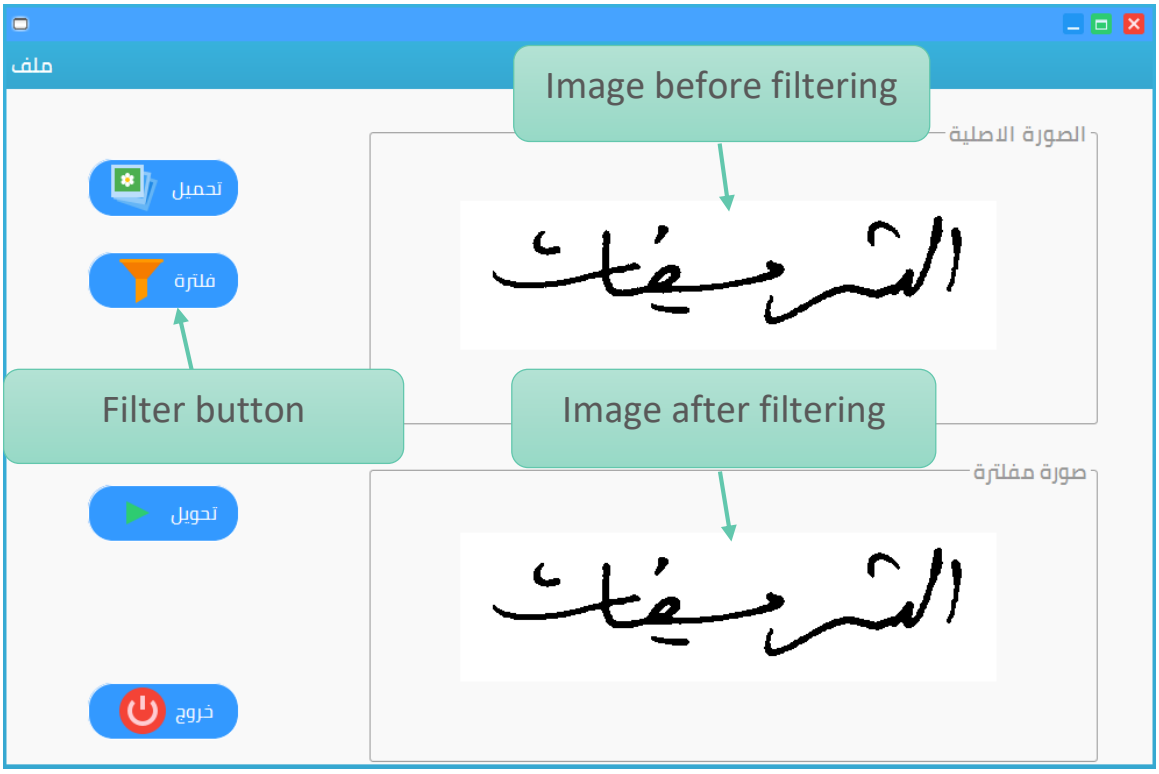



**Figure 4.3:** Filter image.

**2.3 Convert image:**

we click on the button تحويل to convert the image to a series of images containing the points and letters that form the word.


**Figure 4.4:** Image conversion stage

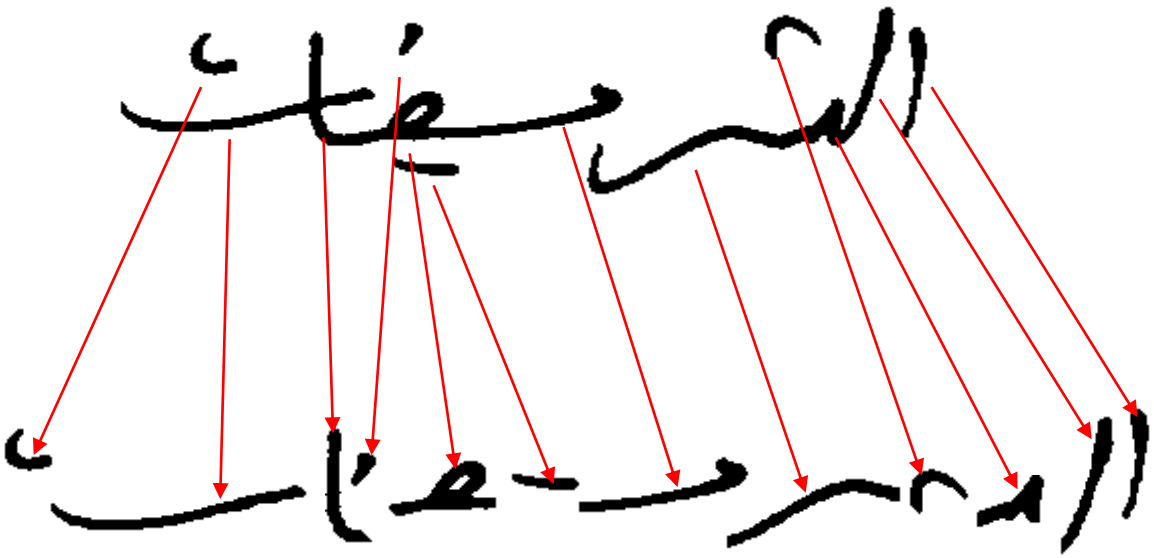


**Figure 4.5:** Segmentation of a word that contains an intersection between letters.

Another example of a word that contains a vertical writing:

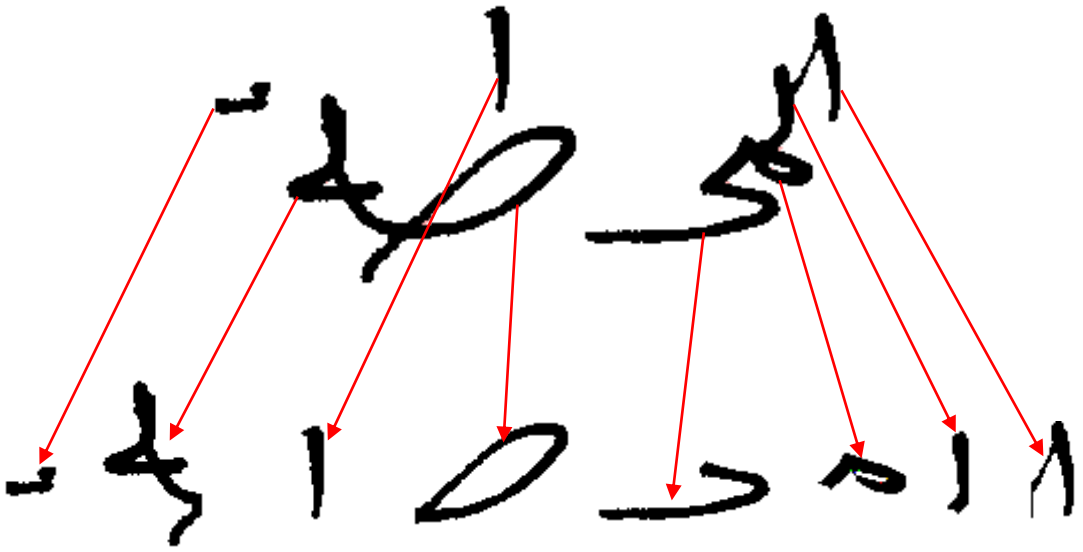


**Figure 4.6:** Segmentation of a word that contains a vertical writing.

### 3. Phase of tests & evaluation of the results:

After the presentation of the architecture of our system as well as the results obtained in each step until the end, we used two algorithms to improve the segmentation of words.we get all the necessary basic segments of the text handwritten processed. The results obtained from this method were very good. however, as we apply our proposed method on 55 images (as a basis for testing) contains handwritten texts of varying degrees of difficulty.

The estimated success rate of this segmentation technique evaluated at 92.73% of the images were segmented successfully despite the difficulties and problems related to nature and style of writing. And 9.27% failures in the results resulting from the existence of elved complications in the writing method and unlimited difficulties in the distinction between letters and words and vertical writing of words.

The 55 images tested are divided into three categories, the first category contains simple handwritten texts, the second category contains medium-difficulty texts ,the third category contains very difficult texts, which contain intersections between letters and vertical writing ,the results obtained for each category according to their level of complexity in the following table:

| The categories | Level of complexity | Nbr images | The success rate |
|---|---|---|---|
| Set of images 01 | Simple Texts | 25 | 100.00 % |
| Set of images 02 | Medium texts | 18 | 88.88% |
| Set of images 03 | Difficult texts | 12 | 83.33% |
| Total number of images | | 55 | 90.73% |

**Table 4.1 :The** success rates of our proposed system according to the complexity level

**Conclusion:**

In this chapter we explained the basics of the program for the segmentation of Arabic handwriting, using two algorithms where the first algorithm depends on the horizontal separation of letters and the second algorithm to extract the intersection of letters.

In conclusion, our contribution has achieved good results for the segmentation of Arabic handwriting in the field of OCR. despite the many problems they face because of the nature, quality and irregularity of writing

# GENERAL CONCLUSION

Despite the efforts and intensive work done in the field of visual recognition of writing, there is no reliable OCR system. But as the authors try to improve scores for better results In the case of our study, we presented a segmentation method that proved its performance in terms of recognition rate.

However, the major problems influencing research in AOCR are the lack of standardization of calligraphy of Arabic characters is the lack of in-depth studies on the classification of fonts from the point of view of calligraphy and body, and also the absence of tools such as as dictionaries, databases and statistics relating to Arabic writing.

Solving these problems would be of considerable benefit, both in terms of the simplification of the task of the AOCR, and the validation and portability of the products produced.

By trimming this work, we hope to have covered much of the field of research in segmentation of Arabic characters, and to be able to contribute to the evolution of the researches, although the efforts of our days intensify in this field and every day new articles are published, dealing with the subject.

**Perspectives**

The work that we have done during this memoir is a step and a first contribution to the structural segmentation of the Arabic handwriting, however we think it can be improved, and extended by the following points:

- Add a dynamic filtering technique instead of a fixed thresholding method used (adaptive thresholding),
- Take into consideration the management of the cut writing in order to arrive at a more effective recognition model (especially for cursive writing).

- The results obtained depend on the readability of the writing, for more recognition rates better representing the effectiveness of the system other tests are necessary.

[1] S. Haitaamar :" segmentation de texte en caractère pour le reconnaissance optique de l'écriture arabe". University of EL-HADJ LAKHDHAR Batna, July 2007.

[2] M. ZAIZ Faouzi:" Les Supports Vecteurs Machines (SVM) pour la reconnaissance des caractères manuscrits arabes". University of Mohamed Khider – BISKRA, July 2010.

[3] M. Côté :" Utilisation d'un modèle d'accès lexical et de concepts perceptifs pour la reconnaissance d'images de mots cursifs ". PhD thesis submitted to superior national school of telecommuncations ,France, June 1997

[4] S.Bachir , H.Abdelkader :" Proposition d'une méthode de segmentation de l'écriture arabe manuscrite artistique". University of ECHAHID HAMMA LAKHDAR El-Oued June 2018.

[5] H. Oulhadj, J. Lemoine, E. Petit, H. Wehbi :" Combinaison d'algorithmes pour la reconnaissance des chiffres et des lettres batons dans un environnement multiscripteur d'ecriture courante mixte ". Laboratory of Study and Research in Instrumentation, Signals and Systems, University of Paris XII, France, May 1999.

[6] N. Cristianini, J. Shawe-Taylor :"An Introduction to Support Vector Machines and Other Kernel-based Learning Methods". Cambridge University Press, 2000.

[7] B.G.Imane :" Proposition d'un modèle de classificateur logique, application à la reconnaissance du texte arabe imprimé ". University of ECHAHID HAMMA LAKHDAR El-Oued June 2014.

[8] A. Soudi, et al :" Arabic Computational Morphology knowledge – based and Empirical Methods ". Springer, Volume 38, 2007.

[9] J. Park :" Hierarchical character recognition and it's use in handwritten word/phrase recognition". Phd thesis, New York University, November 1999.

[10] A.Bennasri, A.Zahour, B. Taconet :" Extraction des lignes d'un texte manuscrit arabe".Vision Interface '99, Trois-Rivières, Canada, 19-21 May 1999.

[11] A. Belaïd :" Reconnaissance automatique de l'écriture et du document ". LORIA-CNRS , scientist Campus B.P. 239, 54506 Vandoeuvre-Lès-nancy, France.

[12] A. Boukharouba , A. Bennia :" Reconnaissance de Caractères Imprimés Omni-fonte ". 3rd International Conference: Sciences of Electronic, Technologies of Information and Telecommunications, Tunisia, March 27-31, 2005.

[13] F. B. Samoud, S. S. Maddouri, K. Hamrouni :" Segmentation de chèques bancaires arabes ". 3rd International Conference: Sciences of Electronic, Technologies of Information and Telecommunications, Tunisia, March 27-31, 2005

[14] L. Robadey :" 2(CREM): Une méthode de reconnaissance structurelle de documents complexes basée sur des patterns bidimensionnels ". PhD thesis submitted to the Faculty of Sciences of the University of Friborg, Switzerland, 2001

[15] F. Menasri :" Segmentation d'image Application aux documents anciens " Thesis Doctor of the University Paris Descartes in Computer Science, France, June 2008.

[16] Al-Rashaideh H:" Preprocessing phase for Arabic Word Handwritten Recognition ". Institute of Computer Science and Automation, Tom 6, N0 1,2006, cmp.11-19, Russia, February 26, 2006.

[17] S. Chevalier et al : "Étude de primitives spectrales pour la reconnaissance de caractères manuscrits dans le cadre d'une approche markovienne 2D". DGA/Parisian Expertise Center, France, November 2005.

[18] S. Quiniou :" Intégration de connaissances linguistiques pour la reconnaissance de textes manuscrits en ligne PhD thesis of INSA de Rennes Ph.D., IMADOC - IRISA, MATISSE, France, December 17, 2007.

[19] S. Carbonnel E. Anquetil :" Modélisation et intégration de connaissances lexicales pour le post-traitement de l'écriture manuscrite en-ligne". IRISA, INSA, France, Mars 2009.

[20] N. Ben Amara et al :" Utilisation des modèles markoviens en reconnaissance de l'écriture arabe : Etat de l'art" National School of Engineers of Monastir - 5019 Monastir - TUNISIA, LORIA-CNRS, Tunisia, April 2001.

[21] Shubair A et al.:" Off-line Arabic handwritten word segmentation using rotational invariant segments features ". The international Arab journal of information technology, Vol. 5, No. 2, April 2008.

[22] Mehennaoui Zahra.:" reconnaissance de l'ecriture arabe manuscrite a base des machines a vecteurs de supports ". university of badji-mokhtar– annaba –2006.

[23] H. Schahrazed "SEGMENTATION DE TEXTES EN CARACTERES POUR LA RECONNAISSANCE OPTIQUE DE L'ECRITURE ARABE", end of study memory, 08, July, 2007.

[24] Baka Abdeladim and Fillali Hicham:" traitement et reconnaissance des caracteres ". University of M'hamed Bougara Boumerdes June 2016.

[25] T. Paquet, L. Heutte, Y. Lecourtier :"Problématique de la Reconnaissance de l'Écriture". ASTI'2001 des Sons, des Images et des Documents à leur Interprétation, France, 2001.

[26] P. Smrž et al :" Off-line Recognition of Cursive Handwritten Czech text". Masaryk University, February 1998.

[27] G.Lorette, Y.Lecourtier, "Reconnaissance et Interprétation de Texts Manuscripts Hors-line: Un Problème d'Analyse de Scène", Bigre Num 80-CNED National Symposium on Writing and Document, Nancy, CNED, July 1992

[28] Jawad H. et al.:" Component-based Segmentation of Words from Handwritten Arabic Text ". International Journal of Computer Systems Science and Engineering 5:1 2009.

[29] A.Gheith, N. Anssari:" Novel Moment Features Extraction for Recognizing Handwritten Arabic Letters ". Journal of Computer Science 5 (3): 226-232, 2009.

[30] T.Abderrazeq, B.Touhami:" Utilisation des caractéristiques statistiques et les Moments de Zernike pour la Reconnaissance des Lettres Arabes Manuscrits ". University of Ahmed Draia – Adrar, June 2018.