



UNIVERSITE ECHAHID HAMMA LAKHDAR - EL OUED
FACULTÉ DES SCIENCES EXACTES
Département D'Informatique



Mémoire de Fin D'étude
Présenté pour l'obtention du Diplôme de

MASTER ACADEMIQUE

Domaine : **Mathématique et Informatique**
Filière : **Informatique**
Spécialité : **Systèmes Distribués et Intelligence Artificielle**

Présenté par :

- **SAIHI Bachir**
- **HEZLA Abdelkader**

Thème

**Proposition d'une méthode de segmentation de
l'écriture arabe manuscrite artistique.**

Soutenue le 04-06- 2018 Devant le jury:

M.	NAGOUDI El Moatez Bellah	MCD	Président
M.	LEJDEL Brahim	MCA	Rapporteur
M.	ZAIZ Faouzi	MCD	Encadreur

Année Universitaire: 2017-2018

Remerciement

*Nous tenons a remercier du fond du cœur, avant tous, ALLAH
le tout puissant de mes avoir donne la force, la volonte et le courage de realiser
ce travail.*

*Nous tenons a remercier chaleureusement et respectivement tout ceux qu'ont
contribue de pres et loin a la realisation de ce projet de fin d'etude et nous
tenons remercier notre encadrant*

*Mr. ZAIZ Faouzi, maitre a l'Universite Echahid hamma lakhdar
El-Oued, d'avoir propose et encadre ce sujet.*

*Nous lui exprimons notre profonde gratitude pour nous avoir fait profiter de ses
connaissances, mais aussi de ses methodes de travail, et surtout de sa rigueur
scientifique.*

Nous remercions les membres de jury pour avoir accepte de juger notre travail.

*Notre reconnaissance va aussi a tous ceux qui ont collabore a notre formation
en particulier les enseignants du departement d'Informatique, Universitaire
Echahide Hamma Lakhdar El-Oued Aussi a nos collegues de la promotion
2017-2018.*

*On remercie egalement tous ceux qui ont participe de pres ou de loin a
elaborer ce travail.*

Dédicace

*A nos familles, nos frères, nos sœurs
nos amis et tous les gens qui nous supportons*

AEK & Bachir

Résumé

Jusqu'aujourd'hui, les systèmes de reconnaissances de l'écriture gardent une grande importance telle que elles sont utilisés dans des différents domaines tels que l'authentification de chèques bancaires, l'échange à distance des fichiers informatiques pour les télécommunications et l'authentification et l'identification des manuscrits. Ils sont capables de transformer l'image d'un papier contenant une écriture manuscrite ou imprimée en fichier numérique contenant le même texte et manipulable par machine.

Malgré, les différents et les nombreuses recherches réalisées durant ces dernières années afin d'améliorer les systèmes de reconnaissance du texte manuscrit, il reste beaucoup de lacune dans les systèmes non latines comme la langue arabe. L'écriture arabe artistique présente un grand challenge vu qu'elle est écrite en centrant beaucoup plus sur le côté esthétique et la beauté d'apparence sans avoir respecté les formes de base des différents caractères. Pour cela, elle se caractérise par une irrégularité de formes, des chevauchements et des intersections entre les segments de lignes constituant les caractères.

Dans ce travail, nous proposons une méthode de segmentation basée multi-scans (N-Scans) à deux phases pour la segmentation de l'écriture arabe manuscrite artistique. Elle consiste à chercher des points d'intérêt. Par la suite, ces points vont permettre de surmonter le problème d'irrégularité, de chevauchements et d'intersections dans l'écriture artistique.

L'approche est testée sur un ensemble d'images d'écriture artistique. Elle donne des résultats encourageants et ouvre des perspectives dans le domaine de la reconnaissance du manuscrit arabe en général.

Mots Clés: OCR, Manuscrit Arabe, PAW, Segmentation.

Abstract

Till now, writing recognition systems keep a very big importance due to the necessity of using them in different fields like bank check authentication, the remote exchange of computer files for telecommunications and authentication and identification of manuscripts. They are able to transform the image of handwritten or printed text paper into numeric file that contains the same text and useable by machine.

In spite of, different and numerous researches done through these last years in the aim of enhancing handwriting recognition systems, still a lot of lacks in on Latin systems like Arabic language. Artistic Arabic writing present a big challenge because its written with concentrating much more on esthetic view and the beauty of appearance without respecting the basic forms of characters. For that, it is characterized by irregularity of forms, overlaps and intersections in characters lines segments.

In this work, we propose a segmentation method based on multi-scans (N-Scans) with two phases for segmenting of Arabic artistic handwriting. First, it searches a set of points of interest. Then, those points are used to handle irregularity, overlapping and intersections within artistic handwriting.

The approach is tested on a set of artistic handwriting images. It gives a very encouraging results and opens many perspectives in the field of handwriting recognition.

Key words: OCR, Handwritten Arabic, PAW, Segmentation.

الملخص

حتى يومنا هذا، لا تزال أنظمة التعرف على الخط تحتفظ بأهمية كبيرة نظرا للحاجة إليها في المجالات التالية: التعرف على محتوى الصكوك البنكية، تبادل ملفات الكمبيوتر عن بعد للتوثيق والتعرف على المستندات المكتوبة بخط اليد. فهي قادرة على تحويل صورة لنص مطبوع أو مكتوب يدويا إلى ملف إلكتروني يحتوي نفس النص يمكن تعديله بالآلة.

بالرغم من العديد ومختلف البحوث المنجزة في هذا المجال خلال السنوات الأخيرة قصد تحسين أنظمة التعرف على الخط المكتوب يدويا، لا تزال هناك العديد من النقائص بالأنظمة الغير لاتينية كأنظمة اللغة العربية. الكتابة الفنية العربية تمثل تحدي كبير لأنها تكتب مع التركيز ومراعاة الجانب الفني والجمالي للمظهر على حساب الأشكال الأساسية لمختلف للحروف. لهذا، هي تتسم بعدم انتظام، وتداخل وتقاطع في خطوط رسم الحروف.

في هذا العمل، نقترح طريقة متعددة المسح وبمرحلتين لتجزئة هذه الكتابة. والتي تعتمد على البحث عن نقاط أهمية. بعدها، تستخدم هذه النقاط لتتغلب على مشكل عدم الانتظام، والتداخل والتقاطع في خطوط الكتابة.

تم تجربة هذه الطريقة على مجموعة من الصور للكتابة الفنية، وقد أعطت نتائج جد مشجعة وتفتح أبعاد أخرى في مجال التعرف على الخط.

الكلمات المفتاحية: التعرف على الكتابة، كتابة اليد، PAW، التجزئة.

TABLE DES MATIERES

Remerciement	I
Dédicace	II
Résumé	III
TABLE DES MATIERES	VI
LISTE DES FIGURES	X
LISTE DES TABLEAUX	XII
Liste des Abréviations	XIII
INTRODUCTION GENERALE	01
Chapitre 01 : RECONNAISSANCE DE L'ECRITURE ARABE MANUSCRITE	03
1. Introduction	03
2. Les caractéristiques spécifiques de l'écriture arabe	03
3. L'art de la calligraphie Arabe "écriture arabe artistique"	10
4. Les Styles d'écriture calligraphique arabe	10
5. Qu'est ce qu'un OCR ?	12
6. Différents aspects de l'OCR :	13
6.1 Reconnaissance en ligne et hors-ligne	13
6.2 Reconnaissance globale et analytique	15
7. Architecture générale d'un système OCR	16
7.1 Acquisition	17
7.1.1 Modes d'acquisition en ligne/ hors ligne	17
7.2 phase de prétraitement	17
7.2.1 Binarisation	18
7.2.2 Lissage et Suppression du bruit	19
7.2.3 Normalisation	19
7.2.4 Redressement de l'écriture " la correction de l'inclinaison"	19
7.2.5 Squelettisation	20
7.3 Segmentation	21
7.4 Extraction de caractéristiques	21
7.4.1 Type de caractéristiques	22
7.5 Phase de Classification	22
7.5.1 L'apprentissage	23
7.5.2 Reconnaissance et décision	23
7.6 POST-TRAITEMENTS	23
8. Conclusion	23
CHAPITRE 02 : SEGMENTATION ET EXTRACTION DES CARACTERISTIQUES DE L'ECRITURE	24
1. Introduction	24
2. La Segmentation de documents et les différents niveaux de segmentation	24
2.1 Segmentation de la page	25
2.2 Segmentation de texte en lignes	26

2.3	Segmentation de lignes en mots	26
2.4	Segmentation de mots en caractères	27
3.	Les différentes approches de segmentation.....	28
3.1	Approches analytique à segmentation explicite.....	29
3.2	Approches analytique à segmentation implicite.....	29
3.3	Approche analytique mixte	29
3.4	Approche globale "Méthodes holistiques"	29
3.5	Approches hybrides	30
4.	Des techniques appliquées pour la segmentation de l'écriture cursive.....	30
4.1	Techniques appliquées pour la segmentation de texte en lignes.....	30
4.1.1	Segmentation par histogrammes de projection horizontale	30
4.1.2	Segmentation basée sur les fenêtres glissantes	31
4.1.3	Segmentation par écoulement d'eau	32
4.1.4	Segmentation basée peinture Piece-Wise.....	32
4.2	Techniques appliquées pour la Segmentation en mots et caractères	34
4.2.1	Segmentation sans classification	34
a)	Segmentation basée sur la squelettisation	34
b)	Technique de segmentation basée sur le traçage des contours	35
c)	Segmentation à partir des histogrammes de projection verticales.....	36
d)	Segmentation basée sur des réservoirs d'eau	36
4.2.2	Segmentation avec classification	37
a)	Segmentation par propagation d'affinité.....	37
b)	Segmentation à base de forêts aléatoires.....	38
c)	Segmentation à base de Spline en Plaque Mince	38
d)	Segmentation à base de réseaux de neurones.....	38
4.2.3	Comparaison des méthodes de segmentation	39
5.	Conclusion	40
CHAPITRE 03 : SEGMENTATION ET EXTRACTION DES CARACTERISTIQUES DE L'ECRITURE		41
1.	Introduction.....	41
2.	Les Problématiques inhérentes posée par l'écriture du cursif artistiques :	41
3.	La méthode de segmentation de l'écriture arabe artistique proposée.....	44
3.1	L'organigramme de la méthode proposée.....	45
3.2	Le principe de la méthode proposée:	47
3.2.1	Etape de l'Acquisition.....	47
3.2.2	Etape de prétraitements	47
3.2.2.1	Etape de binarisation	47
3.2.2.2	Etape filtrage et suppression des bruits	47
3.2.3	Méthode de segmentation proposée	48
1.	Etape 01 : Extraction Les différents composantes connexes	48
2.	Etape 02 : La Détermination différents points d'intérêts "les PNVs et PNHs".....	50
3.	Etape 03 : Filtrage et élimination des points incorrectes.....	51
4.	Etape 04: Découverte des cercles des caractères et des points de chevauchement et d'intersection	54

4.1 La Détermination la location des cercles	54
5. Etape 05: l'étape de segmentation des composants connexes en deux phase	55
5.1 Première phase.....	55
5.2 Deuxième phase	58
4. Conclusion	59
CHAPITRE 04 : SEGMENTATION ET EXTRACTION DES CARACTERISTIQUES DE L'ECRITURE	60
1. Introduction.....	60
2. Notre Système proposé:	60
3. L'Organigramme de l'architecture globale de notre système proposé	61
3.1 Etape de l'Acquisition.....	62
3.2 Etape de prétraitements	62
3.2.1 Etape de binarisation	62
3.2.2 Etape filtrage et suppression des bruits	63
3.3 La Segmentation par l'application de méthode proposée.....	64
4. Phase des tests & évaluation des résultats:	68
5. Conclusion	69
CONCLUSION GENERALE	70

LISTE DES FIGURES

• Figure 1.1: La cursivité de l'écriture arabe.	03
• Figure 1.2 : Lettres arabe ayant des Pointes diacritiques	05
• Figure 1.3 : Exemple des mots arabes incluant la Hamza, la Chadda et le Madda	06
• Figure 1.4 : Une même lettre avec des signes diacritiques différents.....	06
• Figure 1.5 : comparaison entre l'écriture arabe imprimé et le manuscrite Artistique	07
• Figure 1.6 : Illustration des mots arabes avec Ligature vertical	08
• Figure 1.7 : Ligature verticale (sans les points) utilisées.....	08
• Figure 1.8 : Des styles et des tailles différents pour le même mot écrite à main	08
• Figure 1.9 : Exemples de caractères avec boucles.....	09
• Figure 1.10 : Exemples de Différents chevauchements des lettres.....	09
• Figure 1.11 : Divers styles calligraphiques	10
• Figure 1.12 : Exemple de quelques styles calligraphiques de l'écriture arabe.....	12
• Figure 1.13 : Schéma synoptique d'un système de reconnaissance d'écriture arabe.....	12
• Figure 1.14 : Les Différents aspects de l'OCR	13
• Figure 1.15 : Description de la reconnaissance d'un caractère manuscrits	15
• Figure 1.16 : Architecture générale d'un système.....	16
• Figure 1.16 : Graphe de complexité des systèmes de reconnaissance.....	16
• Figure 1.18: Exemple d'allographes des caractères arabes	17
• Figure 1.19: Architecture générale d'un système (OCR).....	18
• Figure 1.17: Différentes opérations de prétraitement	18
• Figure 1.18: Exemple de binarisation	18
• Figure 1.19: Exemple de Lissage et Suppression du bruit	19
• Figure 1.20: Exemple de mots manuscrits avec des tailles différentes	19
• Figure 1.21: L'inclinaison globale du texte par rapport au cadre de l'image	20
• Figure 1.22: Exemple de squelettisation	20
• Figure 1.23: Les différents Modes de segmentations.....	21
• Figure 2.1 : Les différents niveaux du processus de segmentation.....	25
• Figure 2.2 : Exemple illustre le passage entre les étapes de segmentation	25
• Figure 2.3 : Détection des zones d'une page de document	26
• Figure 2.4 : Segmentation du texte en Lignes	26
• Figure 2.5 : exemple de Segmentation du texte en Lignes par projection horizontale.....	26
• Figure 2.6 : Segmentation du ligne en mots.....	27
• Figure 2.7 : exemple de Segmentation d'une Ligne en PAW par projection verticale.	27
• Figure 2.8 : Segmentation d'un mot en caractères.	28
• Figure 2.9 : organigramme des méthodes de segmentation selon R.G.Casey.....	28
• Figure 2.10 : Exemple de chevauchement entre des lignes du texte.	31
• Figure 2.11 : Exemple de segmentation du texte en lignes par projection horizontale.....	31
• Figure 2.12 : segmentation à base de fenêtre glissante	31
• Figure 2.13 : Exemple de la Segmentation par une technique écoulement d'eau	32
• Figure 2.14 : exemple de segmentation avec une technique Pice-Wise.	33

• Figure 2.15 : exemple d'image de texte lissée	33
• Figure 2.16 : Technique de Segmentation basée sur la squelettisation.	34
• Figure 2.17 : exemple illustre l'utilisation d'une Segmentation basée sur la squelettisation.....	34
• Figure 2.18 : Technique de segmentation basée sur le traçage des contours	35
• Figure 2.19 : illustration de la segmentation basée sur le traçage des contours.....	35
• Figure 2.20 : La segmentation d'une ligne de texte en pseudo-mots par Projection verticale.....	36
• Figure 2.21 : Détection des réservoirs "haut" et "bas".....	37
• Figure 2.22 : Segmentation "Drop fall" ascendant basée sur les réservoirs d'eau	37
• Figure 2.23 : Exemple illustre Une solution d'exemplaires basée sur le clustering par AP.....	38
• Figure 3.1: La cursivité d'un mot arabe manuscrite	42
• Figure 3.2: Exemples Illustre les caractéristique et des styles d'écriture artistique	43
• Figure 3.3: des Exemples Illustre les problèmes liés d'écriture artistique	43
• Figure 3.4: Organigramme illustre Le principe générale de la méthode proposée.....	46
• Figure 3.5 : exemple d'extraire Les différents composantes connexes	48
• Figure 3.6: Illustration de l'extraction les différentes composantes d'un texte artistique.....	49
• Figure 3.7: Les Masques de balayage vertical	50
• Figure 3.8: Les Masques de balayage Horizontal	51
• Figure 3.9: illustration de filtrage de points	51
• Figure 3.10: illustration de filtrage des points : entre deux points VDP et HDP	52
• Figure 3.11: illustration des cas de filtrage des points d'intérêt	53
• Figure 3.12: illustration la détermination les cercles des caractères	54
• Figure 3.13: illustration comment détecté les points de chevauchement	55
• Figure 3.14: le principe d'élimination les points chevauchement descendant	56
• Figure 3.15: la détection des intersections des caractères en utilisant les points d'intérêt.....	57
• Figure 3.16: le principe d'élimination les intersections entre les caractères	57
• Figure 3.17: des intersections entre des caractères à trois points d'intérêt seulement	58
• Figure 3.18: Extraction des connecteurs des caractères par les Points VFPs et HFPs.....	59
• Figure 4.1: L'Organigramme de l'architecture globale de notre système proposé.....	61
• Figure 4.2: Illustration de l'étape d'acquisition	62
• Figure 4.3: illustration de l'étape de binarisation de l'image acquit	63
• Figure 4.4: illustration de l'étape de binarisation de l'image acquit	63
• Figure 4.5: Illustration comment séparer les signes diacritiques à les Composants connexes.....	65
• Figure 4.6: Illustration comment la méthode extraire les segments de base d'un composants connexe.....	66
• Figure 4.7: illustration comment le système génère les composants de base après la segmentation.....	67

LISTE DES TABLEAUX

- *Tableau 1.1 : Différentes formes des caractères arabes selon leurs positions..... 04*
- *Tableau 1.2 : Classification des mots selon le Nombre des pseudo-mots qui les composées 05*
- *Tableau 1.3 : groupes de caractères arabes ayant la même corp 05*
- *Tableau 1.4: Les positions qu'occupe Hamza en association avec Alif, Waw et Ya 07*
- *Tableau 1.5 : Comparaison de la reconnaissance des caractères en lignes et hors lignes..... 14*
- *Tableau 2.1: Comparaison des méthodes de segmentation en mots et caractères 39*
- *Tableau 3.1 : La distinction entre les composantes connexes et les signes diacritiques..... 50*
- *Tableau 4.1: Les taux de sucées de notre système proposée 68*

Liste des Abréviations

- **OCR :** **Optical Character Recognition.**
- **PAW :** **Peace of Arabic Word.**
- **PPP :** **Points Par Pouce.**
- **RDF :** **Reconnaissance de Forme.**
- **SVM :** **Machine à Vecteur du Support.**
- **GIF** **Graphics Interchange Format**
- **BMP** **Bitmap**
- **JPEG** **Joint Photographic Experts Group**
- **AOCR** **Arabic Optical Character Recognition.**
- **PS** **Pen Size.**
- **PSP** **Primary Segmentation Points.**
- **SB** **l'ensemble des Segments de Base .**
- **SD** **l'ensemble des Signes diacritiques "El chakel , Les points , Les Marques additionnelles".**
- **RF:** **Random Forests / les forêts aléatoires**
- **SPs:** **Segmentation points**
- **CCs** **les composants connexes**
- **PPSL :** **Piece-wise Potential Separating Line.**
- **SPM** **La méthode de Spline en Plaque Mince .**
- **AP :** **Affinity Propagation**

Introduction général

Malgré tous les travaux et les avancés réalisés durant ces dernières décennies pour améliorer les systèmes de la reconnaissance de l'écriture. De tels systèmes reste un défi soulevé surtout pour les langues non latines comme l'arabe, vue la nature très complexe de ces langues. Principalement, deux types de systèmes de reconnaissance de l'écriture sont connus : reconnaissance hors-ligne et reconnaissance en-ligne. Le premier, semble beaucoup à la faculté de lecture de l'être humain, vue qu'il consiste à transformer une image de texte en un texte en format numérique. Autant que le second, prend les données du tracé directement à partir d'un stylo et une tablette de numérisation pour les transformées en un texte digitalisé. Plusieurs domaines démontrent l'importance des systèmes similaires, tels que l'authentification de chèques bancaires, l'échange à distance des fichiers informatiques pour les télécommunications, l'authentification et l'identification des manuscrits, la reconnaissance des cartes cartographiques et la lecture automatique de documents administratifs. Les techniques d'OCR (Optic Character Recognition) arabe ne sont pas encore bien développées, parce que la nature cursive de l'écriture arabe induit de nombreuses difficultés techniques.

L'écriture arabe artistique est d'une nature cursive et repose beaucoup plus sur la beauté des lettres et de l'écriture de manière esthétique sans préserver la forme standard du caractère. Une lettre ne possède pas une forme fixe et peut être à des formes différentes. En plus, des formes diacritiques ainsi que des formes similaires aux formes des caractères peuvent être insérés pour donner une vision de décoration plus élégante. Ces formes ajoutées peuvent dans des positions différentes au-dessous ou au-dessus des caractères. L'espacement entre les pseudo-mots et même entre les caractères est irrégulier, ce qui génère des chevauchements et des intersections entre les mots. En écriture arabe artistique, les cercles des caractères peuvent être avec des segments liés.

Pour ces raisons, la reconnaissance du manuscrit arabe et surtout l'écriture artistique reste un défi, et nécessite des travaux futurs. L'objectif de ce travail est d'étudier la possibilité d'utiliser un algorithme de segmentation structurelle multi-scans pour améliorer le taux de segmentation d'un texte manuscrit arabe artistique. Cet algorithme va résoudre les problèmes des chevauchements et des intersections entre les caractères.

Dans ce travail, nous allons proposer une contribution afin de remédier aux problèmes de chevauchements et intersections des caractères de l'écriture arabe manuscrite artistique. Premièrement, la méthode commence par l'identification des points d'intérêt en utilisant deux scans parallèle, l'un vertical et l'autre horizontal. Ensuite, ces points sont utilisés pour déterminer les

boucles dans l'écriture dans le but d'éviter des cas d'ambiguïté. Par la suite, les cas d'intersection sont traités à part. Finalement, le système peut identifier les points considérés comme connecteurs de caractères.

Le reste du travail est organisé comme suit :

Le premier chapitre présente un état de l'art sur les systèmes de reconnaissance de l'écriture manuscrite. D'abord, nous allons présenter les caractéristiques de l'écriture arabe. Ensuite, nous allons donner un aperçu sur les différentes étapes constituant un système de reconnaissance d'écriture manuscrite hors-ligne. Par la suite, nous allons décrire quelques travaux réalisés dans le cadre de la reconnaissance de mots arabes manuscrits.

Le deuxième chapitre présente un panorama sur la segmentation de l'écriture manuscrite. Dans ce chapitre, nous allons voir les étapes les plus importantes de la segmentation et de l'extraction des caractéristiques de l'écriture. Ensuite, nous présentons les différentes techniques et approches utilisées dans la littérature pour la segmentation du texte manuscrit cursive ainsi que les avantages et les inconvénients de chacune d'elles.

Le troisième chapitre va mettre l'accent sur la méthode de segmentation multi-scans (N-Scans) proposée. Premièrement nous allons voir les problématiques posées par l'écriture arabe manuscrite artistique. Ensuite, nous présentons en détail notre contribution présentée par une méthode de segmentation multi-scans (N-Scans) à deux phases de l'écriture manuscrite artistique.

Le quatrième chapitre présente la validation des résultats. Dans lequel, nous allons voir comment fonctionne globalement notre système. Parmi ces étapes nous présentons le langage de programmation utilisé ainsi que l'interface principale de l'application. Enfin, nous présentons les résultats obtenus ainsi que les discussions faites.

Nous terminons le travail par une conclusion sur les résultats obtenus par la méthode utilisée, et des perspectives de ce travail.

Chapitre 01: Reconnaissance Optique des Caractères « OCR »

1. Introduction

La reconnaissance et la segmentation de l'écriture arabe manuscrite ou imprimé reste encore aujourd'hui au niveau de la recherche et de l'expérimentation. malgré ça ,dans les années derniées elle a pris un nouvel essor et fait l'objet d'applications de plus en plus nombreuses dans des domaines différents. Parmi ces applications «le traitement automatique des dossiers et des formulaires administratifs, le tri postal, la lecture de chèques bancaires, la numérisation et sauvegarde du patrimoine culturel manuscrit, etc».

Dans ce chapitre nous présenterons les caractéristiques morphologiques spécifiques de l'écriture arabe et les Styles d'écriture calligraphique arabe, en suite nous définissons *quelques notions de l'OCR* (Optical Character recognition) et les différentes étapes intervenant dans un système OCR.

2. Les caractéristiques spécifiques de l'écriture arabe

L'écriture arabe possède des caractéristiques différentes d'autres langues (Latins, Chinois...) en structure et en mode de liaison entre les caractères formant un mot.

L'écriture arabe est par nature cursive de droite vers la gauche, la plupart des lettres sont interconnectés constituant des mots et bien aussi dans le cas de l'imprimé que du manuscrit, pour cela elle est difficile à segmenter et présente une grande variabilité.

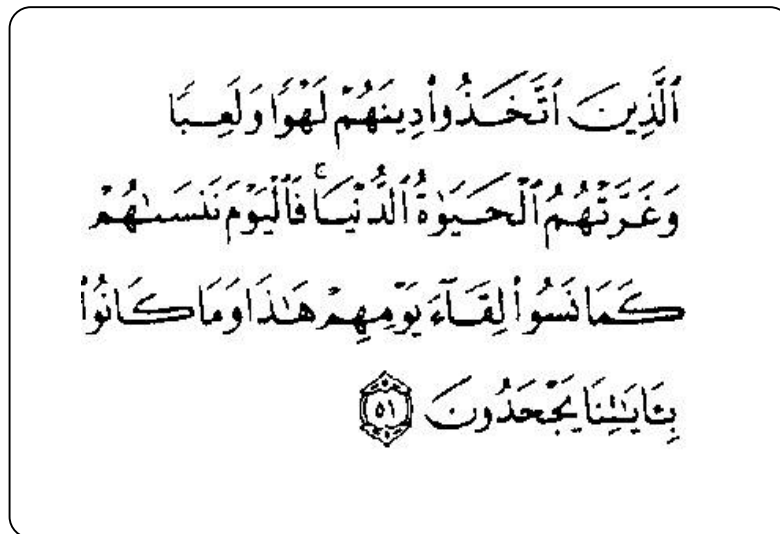


Figure 3.1: La cursivité de l'écriture arabe[12].

- L'écriture arabe est une écriture utilise un alphabet de 28 lettres, la forme de chaque caractère change selon la position dans le mot . Il y a 22 lettres écrites sous quatre formes différentes selon leur position dans les mots qu'elles soient isolées (I), au début (D), au milieu (M), ou à la fin (F) du mot, Et les six lettres restantes "أ", "ز", "ر", "د", "ذ", "و" sont ... caractérisées par ne pas attacher à aucun lettres successeurs écrivent seulement en deux formes : isolées ou à la fin ce qui introduisent une coupure dans le mots elle meme (Tableau1.1)[15,24].

La lettre	Les différentes formes d'un caractère				La lettre	Les différentes formes d'un caractère			
	Forme Isolé (I)	Forme au début (D)	Forme au milieu (M)	Forme à la fin (F)		Forme Isolé (I)	Forme au début (D)	Forme au milieu (M)	Forme à la fin (F)
Alif	ا	ا	ا	ا	Dhad	ض	ض	ض	ض
Ba	ب	ب	ب	ب	Tad	ط	ط	ط	ط
Ta	ت	ت	ت	ت	Thad	ظ	ظ	ظ	ظ
THa	ث	ث	ث	ث	Ain	ع	ع	ع	ع
Jim	ج	ج	ج	ج	Gyn	غ	غ	غ	غ
Ha	ح	ح	ح	ح	Fa	ف	ف	ف	ف
Kha	خ	خ	خ	خ	Qaf	ق	ق	ق	ق
Dal	د	د	د	د	Kaf	ك	ك	ك	ك
Dhal	ذ	ذ	ذ	ذ	Lam	ل	ل	ل	ل
Ra	ر	ر	ر	ر	Mim	م	م	م	م
Zay	ز	ز	ز	ز	Nun	ن	ن	ن	ن
Sin	س	س	س	س	Hae	ه	ه	ه	ه
Shin	ش	ش	ش	ش	Waw	و	و	و	و
Sad	ص	ص	ص	ص	Ya	ي	ي	ي	ي

Tableau 1.1 : Différentes formes des caractères arabes selon leurs positions[4,24].

- De ce fait, un mot unique arabe peut être contient un ou plusieurs pseudo-mots **PAW** (**peace of arabic word**) Séparé par des espaces, chaque pseudo-mot composé d'une séquence de lettres liées, comme peut également être un caractère isolé (*Tableau 1.2*)[11,18].

Nombre des pseudo-mots				
05 PAWs	04 PAWs	03 PAWs	02 PAWs	01 PAWs
الفردوس	الورد	معلوماتية	تقنيات	سد
الأزهار	الدينار	مرطبات	الله	عملية
الجزائرية	الرحمان	الرحيم	قيوم	قسنطينة

Tableau 1.2 : Classification des mots selon le Nombre des pseudo-mots qui les composées [15]

- Dans l'écriture arabe, il y a 15 lettres utilisées incluent des points diacritiques. Ces points Le nombre maximum de ces points qui peut avoir une lettre est de trois points et apparaissent seulement au dessus ou en dessous du corps du caractère, mais jamais en haut et en bas simultanément (Figure 1.2), leur rôle est de permettre la distinction entre certains groupes de lettres et parfois sont indispensables pour la différenciation entre les mots (*Tableau 1.3*)[4,24].

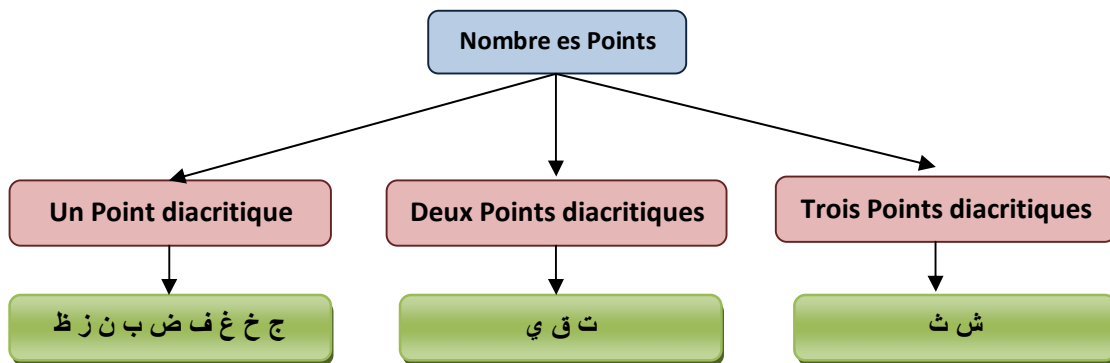


Figure 1.2 : Lettres arabe ayant des Pointes diacritiques[4]

{ د ذ } → د	{ ر ب ت ث ز ي } → ر	{ س ش } → س
{ ص ض } → ص	{ ز ر } → ر	{ ح خ } → ح
{ و ف } → و	{ ط ظ } → ط	{ ع غ } → ع
{ ه هـ } → هـ	{ ي ي } → ي	{ ه هـ } → هـ

Tableau 1.3 : groupes de caractères arabes ayant la même corp mais avec un Nbr et/ou emplacement de diacritiques différents [17,9]

- [Un caractère arabe peut contenir un trait vertical (TAA ط), un trait oblique (KAF ك) ou un zigzag (ALIF ا)] [4].

- L'écriture arabe est riche en diacritiques, un signe diacritique est une composante secondaire d'une lettre, qui vient la compléter ou en modifier le sens. Les signes diacritiques peuvent être des points ou d'autres signes tels que hamza (ء), chadda (ّ), madda (ـ) ... etc [4].



Figure 1.3 : Exemple des mots arabes incluant la Hamza, la Chadda et le Madda [4].

- Ces signes peuvent se situer au-dessus ou au-dessous du caractère, mais jamais en haut et en bas simultanément. Il existe **8 signes** qui peuvent se placer au-dessus de la ligne d'écriture : telle que *fatha* (َ), *damma* (ُ), *kasra* (ِ), *Soukoun* (◌). En plus « *tanwin* » peuvent être formés à partir d'un *double fatha* (ً), d'un *double damma* (ٌ) ou d'un *double kasra* (ٍ).

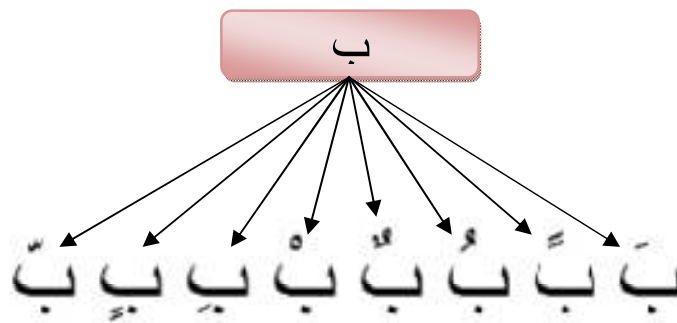


Figure 1.4 : Une même lettre avec des signes diacritiques différents

- **La hamza** « ء » a une orthographe spéciale qui dépend de règles grammaticales, ce qui multiplie les formes nécessaires à sa représentation, puisqu'elle peut s'écrire seule ou sur le support de trois voyelles (*alif*, *waw* et *ya*) dont elle suit le code (**Tableau 1.4**). La considération du symbole « ~ » qui s'écrit uniquement sur le support du caractère « ا » (**Tableau 1.4**).

Caractère	Initial	Médiane	Finale	Isolé
Alif + ~	-	-	أ	آ
Alif + ء	-	-	أ	أ
			آ	آ
Waw + ء	-	-	و	و
Ya + ء	ي	ي	ي	ي

Tableau 1.4: Les positions qu'occupe Hamza en association avec Alif, Waw et Ya [4,10,]

- **La ligne de base :** Une caractéristique distinguée de l'écriture arabe, Les caractères arabes sont connectés par une ligne de base horizontale dite encore ligne de référence, et l'espace est utilisé comme séparateur de mots. La ligne de base possède un nombre maximum de pixels noirs dans le texte (voire *la figure 1.5*). Cette caractéristique peut être découverte simplement sur l'écriture arabe imprimé, mais pas dans l'écriture manuscrite artistique.

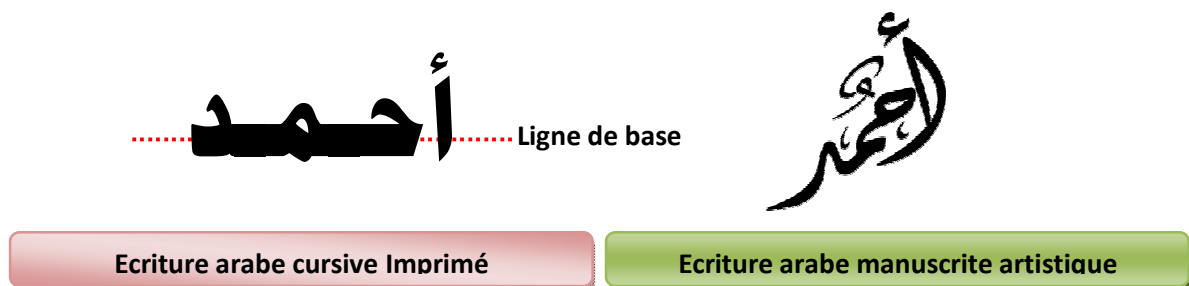


Figure 1.5 : comparaison entre l'écriture arabe imprimé et le manuscrite Artistique en l'existence d'un ligne de base

- **Les ligatures verticales:** Dans certaines fontes arabes ou l'écriture arabe manuscrite plusieurs caractères peuvent être écrits de façon combinée. Les ligatures verticales sont des superpositions verticales de deux, trois ou quatre caractères sont utilisées pour des raisons d'esthétique comme un style d'écriture manuscrite et artistique. (*Figure 1.7*).

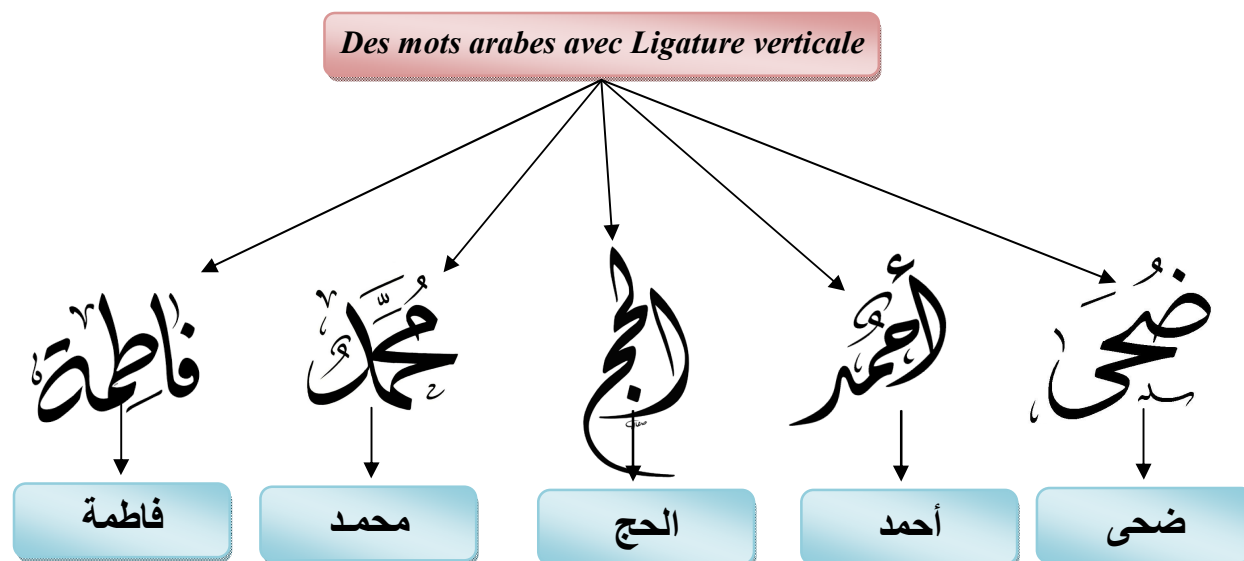


Figure 1.6 : Illustration des mots arabes avec Ligature vertical.

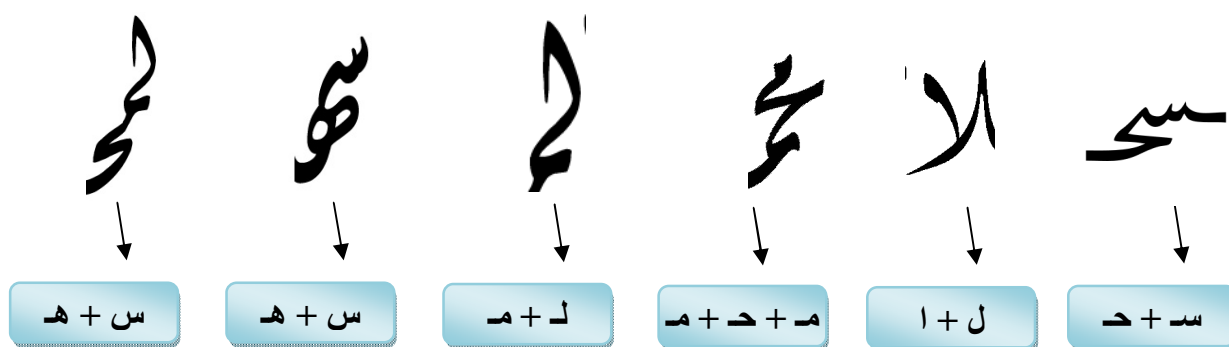


Figure 1.7 : Ligature verticale (sans les points) utilisées.

- Le même mot dans un texte arabe manuscrite pourrait avoir des styles et des tailles différents pour le même auteur ainsi que pour différents auteurs, comme le montre *la figure 1.8*.

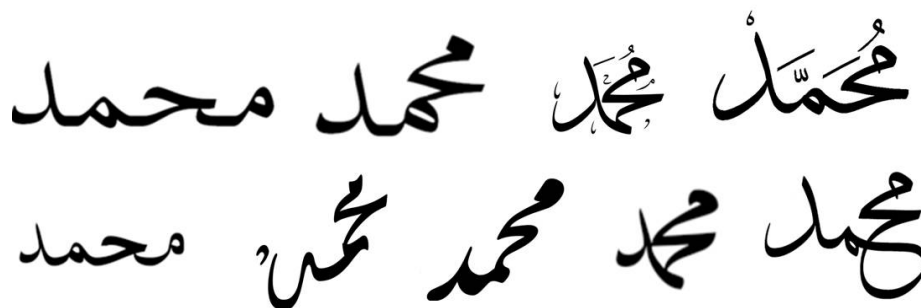


Figure 1.8 : Des styles et des tailles différents pour le même mot écrite à main.

- Certains caractères arabes incluent une boucle qui peut avoir différentes formes
(Figure 1.9)



Figure 1.9 : Exemples de caractères avec boucles[10].

- **Le chevauchement entre des caractères** : En outre, dans l'écriture arabe manuscrite à main ou artistique, des caractères peuvent être chevaucher ou intersectés entre eux. (Figure 1.10).

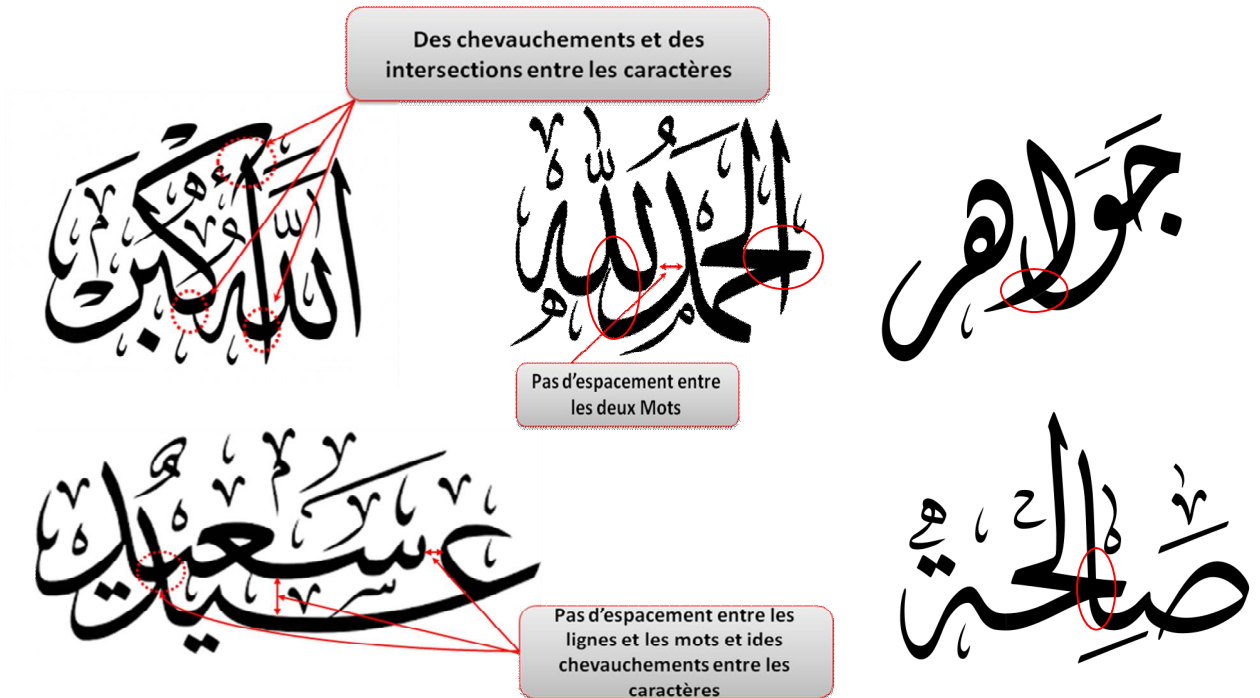


Figure 1.10 : Exemples de Différents chevauchements des lettres.

- Dans les textes arabes manuscrits, certains caractères semblent similaires, bien qu'ils soient différents, et il est difficile même pour l'œil humain de trouver la différence. En outre, le même caractère peut sembler différent dans ses diverses formes, telles que (غ, ف). En outre, la grande similitude entre certains des caractères manuscrits rend la classification de ces caractères un autre défi.

3. *L'art de la calligraphie Arabe "écriture arabe artistique"*

La calligraphie Arabe, aussi connu comme la calligraphie islamique est la pratique artistique de l'écriture en alphabets Arabes « Alharf Alarabi ». La calligraphie est connue principalement dans les pays partageant un patrimoine culturel Islamique et Arabe commun. Cette forme d'art est basé sur l'écriture en lettre arabe, qui pendant longtemps a été utilisé par tous les musulmans dans leurs langues respectives. Ils l'ont utilisé pour écrire principalement les paroles de Dieu « Le saint coran ».

La calligraphie « belle écriture » est la pratique artistique basé sur le côté Esthétique de l'écriture en alphabets Arabes « Alharf Alarabi », La calligraphie est considérée plus exactement comme un art de bien dessiner ou d'écrire d'une belle manière les caractères, au même titre que le dessin ou la peinture. Cette dernière sert souvent à donner l'impression d'une écriture imagée.

4. *Les Styles d'écriture calligraphique arabe :*

Il existe plusieurs styles d'écriture calligraphique développés dès les premiers temps de l'islam, Elle varie selon les milieux et les régions, d'une extrême simplicité formelle à la complexité exhaustive de l'écriture arabe. Chaque style arabe est régi par des lois particulières. D'un style à un autre, les proportions d'une même lettre et son dessin peuvent changer considérablement.

Les trois plus importantes sont le style Koufique, le Diwani et le Farsi. Nous citons quelques styles par exemple : Tholothi, Nesghi, Requeh, Dewani, Farsi, Othmani et Koufique [1, 25]



Figure 1.11 : Divers styles calligraphiques[25].

- **Le Diwani** : d'origine Turque, et ayant connu son summum durant le règne Ottoman, ce style de calligraphie arabe est tout en courbes, idéal pour les rêveurs et les épris de liberté et se définit par l'élongation des caractères et son allure ornementale prononcée [1,4,25].

- **Le Koufique** : ce style d'écriture (anciennement appelé "Hiri" et issu de l'écriture syriaque) tire son nom de la ville iraquienne de Koufa, il est très utilisé en Syrie et en Irak. Cette écriture illustre certains très beaux corans du XIème siècle. Il se caractérise par une certaine sobriété et beaucoup de lignes verticales. il combine harmonieusement lignes droites, angles et formes rondes. Il sera alors largement utilisé pour orner des bâtiments et des monuments [1,4,25].

- **Le Naskhi:** un des styles les plus anciens avec l'écriture koufique, respectant le caractère esthétique de l'écriture arabe, le style classique naskhi rassemble souplesse du style perse et harmonie de l'écriture koufique. Il est aujourd'hui utilisé pour la copie (Naskh) des manuscrits. Il a été aussi adapté à l'imprimerie, à la machine à écrire et à l'ordinateur. Il est aujourd'hui le style le plus employé dans les livres et les journaux.. [1,4,25].
- **Le Farsi:** style typiquement perse, qui tend vers l'effacement des lettres angulaires pour un effet plus doux. Il a été créé par les calligraphes de la Perse (Fares) pour les recueils de poésie. Il est devenu le style prédominant chez les persans, et les turcs.
- **Le Maghribi :** autrefois utilisé dans les pays du Maghreb-Arabe, en Espagne Andalousie et au Soudan. Il a été largement utilisé pour transcrire le Coran au Maghreb-Arabe. Il tend, aujourd'hui, à être supplanté par le style Naskh en Afrique du Nord [1,4,25].
- **Le Riqa :** ou « Petite Feuille » dérive du Naskhi et du Thuluth. L'aspect géométrique de ses lettres et particulièrement les fioritures des finales, s'apparente largement à celles du Thuluth, mais elle est bien plus petite et dotée de courbes plus arrondies et ses Alifs ne sont jamais écrits avec des barbelures. Le centre des boucles des lettres est toujours rempli, les lignes horizontales sont très courtes et les ligatures agencées avec densité, les finales étant souvent rattachées aux initiales. C'est de nos jours l'écriture manuscrite la plus employée dans le monde arabe [1,4,25].
- **Le Thuluth:** est une calligraphie statique et monumentale, Il est utilisé, de nos jours, pour les titres des chapitres et des livres, et à des fins décoratives dans les manuscrits et les inscriptions monumentales. On la juge comme la plus importante des écritures ornementales. [1,4,25].
- **Le Mohaqqaq :** était à l'origine une écriture dont les lettres étaient moins angulaires que le Koufique, avec des ligatures amplement séparées; l'ensemble était « produit avec méticulosité » comme son nom le signale. Cette calligraphie arabe acquit une certaine rondeur qui la rendit plus facile à tracer et elle devint l'écriture privilégiée des scribes et pour la correspondance. Il se caractérise par sa simplicité par rapport aux autres styles [1,4,25].

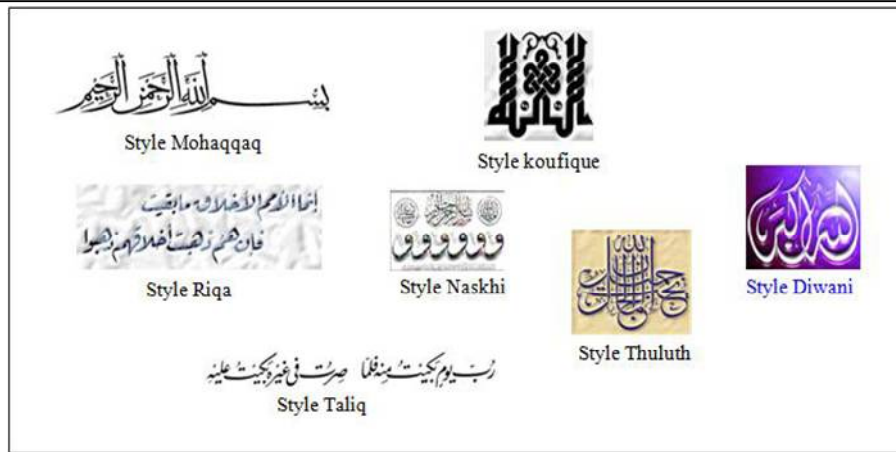


Figure 1.12 : Exemple de quelques styles calligraphiques de l'écriture arabe [4]

5. Qu'est-ce qu'un OCR ?

La Reconnaissance Optique de Caractères appelée **OCR** (**O**ptical **C**haracter **R**ecognition), est le domaine qui permet de convertir des documents différents, tels que les images numérisées de documents papier, en documents modifiables par l'ordinateur, et d'extraire les données d'une image pour être manipulable par l'ordinateur.

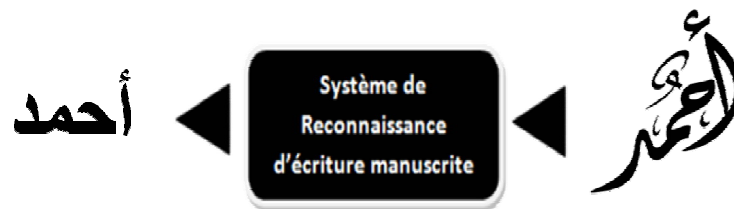


Figure 1.13 : Schéma synoptique d'un système de reconnaissance d'écriture arabe [20].

- Domaines d'application d'OCR :

- L'une des premières utilisations de cette science fut dans le domaine postale La reconnaissance automatique des adresses écrites sur les lettres. « dans le tri automatique du courrier en fonction du code postal ou de l'adresse manuscrite écrites sur les lettres », « et dans l'automatisation le transfert dans les bonnes zones postales ».
- Dans l'authentification de chèques bancaires,
- Dans l'échange à distance des fichiers informatiques pour les télécommunications,
- Dans l'authentification et l'identification des manuscrits,
- Dans la reconnaissance des cartes cartographiques.
- Dans l'extraction automatiquement le contenu et des données des formulaires et des documents et la lecture automatique de documents administratifs [8] .

6. Différents aspects de l'OCR :

[Il existe plusieurs modes de classification des systèmes OCR parmi lesquels on peut citer :

- Les systèmes qualifiés de « en-ligne » ou « hors-ligne » suivant le mode d'acquisition.
- Les approches globales ou analytiques selon que l'analyse s'opère sur la totalité du mot, ou par segmentation en caractères.
- Les approches statistiques, structurelles ou stochastiques relatives aux traits caractéristiques extraits des formes considérées.] [1]

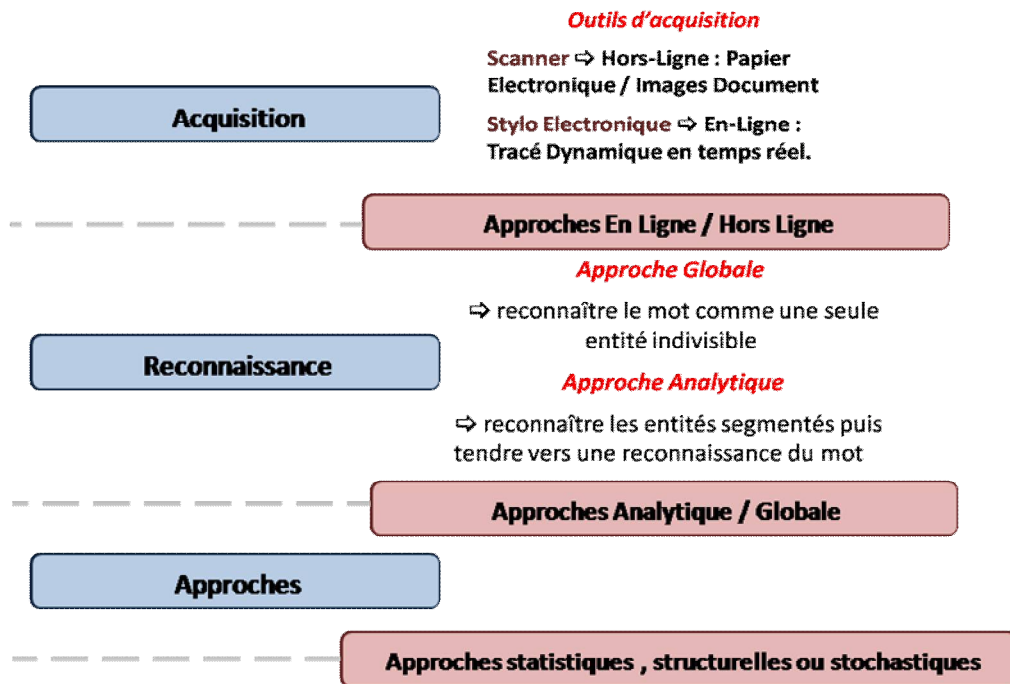


Figure 1.15 : Les Différents aspects de l'OCR [2,6,10].

6.1 Reconnaissance en ligne et hors-ligne

la reconnaissance de l'écriture fait par deux approches différentes, approche hors-ligne et approche en-ligne , chacun de ces deux approches ayant ses outils propres d'acquisition et ses algorithmes correspondants de reconnaissance, ils se diffèrent par la façon d'acquérir les tracés de l'écriture (en-ligne ou hors-ligne), et par la présence ou non des informations décrivant l'ordre de construction de chaque tracé ; [13, 15, 1].

a) La reconnaissance En-Ligne (on-line)

Ce mode de reconnaissance s'opère en temps réel (c'est à dire pendant l'écriture) au fur et à mesure qu'ils sont écrits à la main . il s'agit de saisir un texte avec un stylo et une table à digitaliser pour le transmettre à un traitement de texte et les informations se présentent sous

forme d'une suite ordonnée de points définis par leurs coordonnées, l'ordre d'arrivée étant important pour la classification. [1,6,7,13,10,15,16].

b) La reconnaissance hors-ligne (off-line) :

Elle convient aux documents imprimés et les manuscrits déjà rédigés "hors ligne". La reconnaissance dans cette approche fait après l'acquisition, et les informations recueillies se présentent sous la forme d'une image discrète constituée d'un ensemble des pixels. L'écriture prend l'aspect d'un signal spatial bidimensionnel numérisé [1,6,7,13,10,15,16].

La Reconnaissance hors ligne caractérisé par l'absence d'informations temporelles, elle destiné à des applications qui ne nécessitent pas un traitement en temps réel telles que les applications bancaires ou postales.[7, 16, 1].

c) Comparaison entre les deux approches [16]:

Critère de comparaison	L'APPROCHE DE RECONNAISSANCE	
	En Ligne	Hors Ligne
Outil d'acquisition	- Stylo électronique plus tablette graphique	- Scanner ou caméra.
Bruit d'image	- Faible	- Existence d'un bruit important.
Informations disponibles	- La position. - La direction de mouvement. - Les points de fin. - Les points de début. - Ordre des traits	- Absence d'informations contextuelles.
Exigence des données brutes	- Echantillons	- Points
Taux de reconnaissance	- Plus élevé	- Plus bas.
Précision	- Plus élevé	- Plus bas.

Tableau 1.5 : Comparaison de la reconnaissance des caractères en lignes et hors lignes.[6]

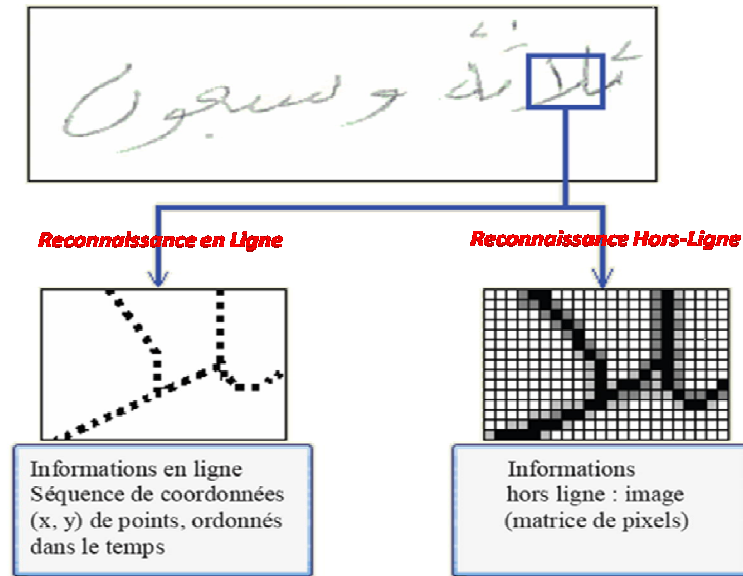


Figure 1.16 : Description de la reconnaissance d'un caractère manuscrits [20]

6.2 Reconnaissance globale et analytique :

En se basant sur le processus de segmentation, il y a deux approches de reconnaissance ont été appliquées.

a) L'approche globale :

L'approche globale essaye de reconnaître la représentation intégrale des mots de l'image d'entrée " Considère le mot comme une seule entité indivisible " et de le décrire indépendamment des caractères qui le constituent.

L'avantage est de garder le caractère dans son contexte avoisinant, ce qui donne une efficacité de modéliser des variations de l'écriture . Cependant elle est pénalisante par la taille mémoire, le temps de calcul et la complexité du traitement qui croient linéairement avec la taille du lexique considéré, d'où une limitation du vocabulaire [15,1 ,10].

b) L'approche analytique :

Dans cette approche, le mot est segmenté en caractères ou en fragments morphologiques significatifs appelé graphèmes. La reconnaissance du mot consiste à reconnaître les entités segmentées puis tendre vers une reconnaissance du mot, cette approche est la seule applicable dans le cas de grand vocabulaire

Un processus de reconnaissance selon cette approche est basé sur une alternance entre deux phases : la phase de segmentation et la phase d'identification des segments [15, 1].

7. Architecture générale d'un système OCR (Optical Character Recognition)

Un système OCR se présente sous forme d'un ensemble de modules en partant de l'acquisition des caractères à reconnaître pour arriver à leur reconnaissance effective (*figure 1.19*).

Un système de reconnaissance fait appel généralement cinq étapes, ces étapes ne sont pas tous obligatoires, la première étape est le prétraitement où différentes opérations peuvent être effectuées, la deuxième est la segmentation, elle peut être explicite ou implicite suivie par l'étape d'extraction de caractéristiques, ensuite vient l'étape de reconnaissance qui est l'étape la plus importante où plusieurs classifieurs peuvent être utilisés, enfin l'étape de post-traitement qui est une étape optionnelle servant à confirmer les résultats obtenus. Dans ce qui suit, nous nous intéressons aux systèmes de reconnaissance hors-ligne et l'étape de segmentation [1,6].

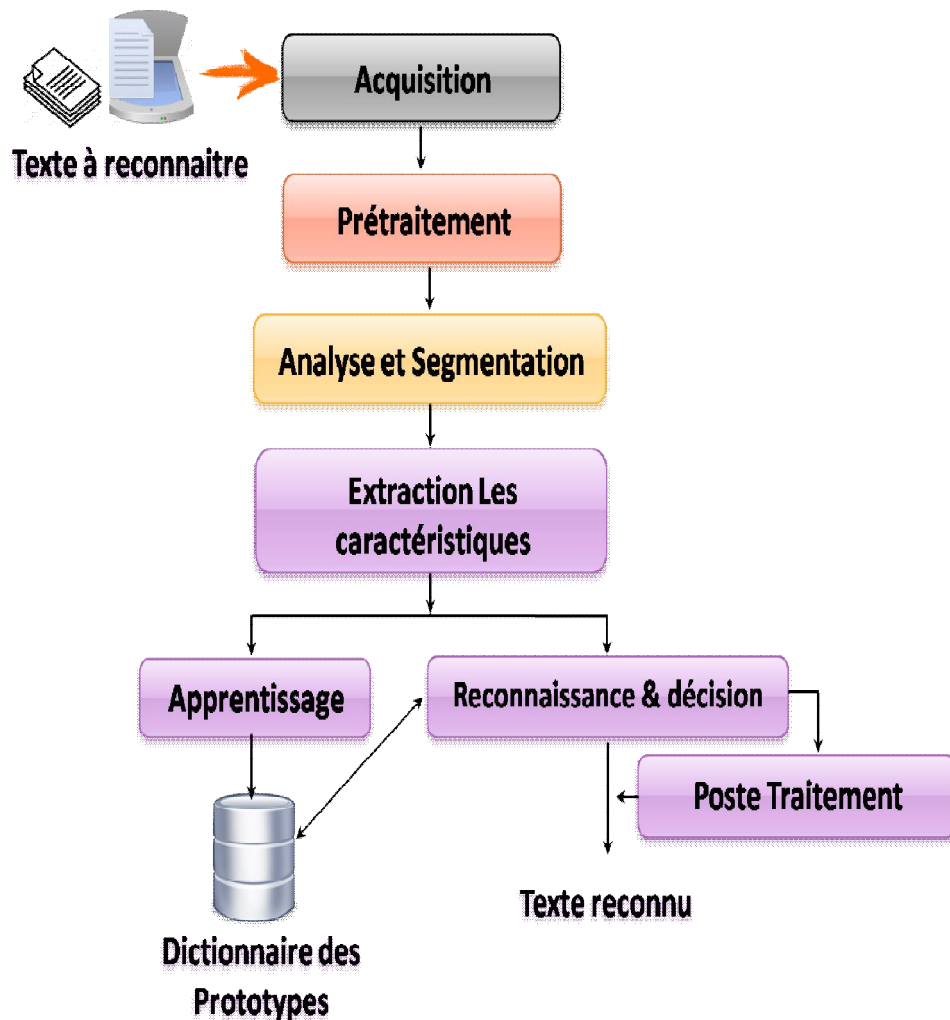


Figure 1.17: Architecture générale d'un système (OCR) [15,6]

7.1 Acquisition

L'acquisition permet la conversion du document papier sous la forme d'une image numérique à l'aide d'un scanner ou des caméras linéaires. Elle est caractérisée par deux paramètres importants : la résolution et le niveau d'éclairage. [15, 1]

○ *La résolution*

La résolution optimale d'une image dépend de l'épaisseur du trait d'écriture. Ainsi, pour les traitements ultérieurs puissent s'appliquer correctement, il faut que le trait d'écriture ait une épaisseur minimale de 3 pixels. La résolution souvent utilisée de 300dpi. Pourtant, quand la taille de l'écriture est petite, il faut augmenter la résolution. [15, 1]

○ *Le Niveau d'éclairage*

Un éclairage élevé (du numériseur) réduit le bruit, mais fait disparaître les traits minces. Donc il faut choisir un niveau d'éclairage optimal selon la qualité du document physique. [15, 1]

7.1.1 Modes d'acquisition en ligne/ hors ligne

Il y a deux modes d'acquisition de l'écriture *Acquisition hors ligne et en ligne*.

○ *Acquisition de l'écriture hors ligne :*

Dans le contexte de l'écriture "hors ligne", les systèmes d'acquisition les plus courants sont essentiellement des scanners ou des caméras linéaires (barrettes CCD). La résolution de l'image numérisée influence les étapes ultérieures d'un système de lecture automatique. Il est communément admis que la résolution optimale d'une image est fonction de l'épaisseur du trait d'écriture.

○ *Acquisition de l'écriture en ligne :*

Dans le cas de l'acquisition "en ligne", les dispositifs d'acquisition les plus répandus sont des tablettes à numériser ou des "papiers électroniques". L'échantillonnage du tracé délivre une série de coordonnées décrivant la trajectoire du stylet au cours du temps. Les tablettes les plus perfectionnées permettent également d'avoir accès aux informations de pression et d'inclinaison du stylo.

7.2 Phase de prétraitement

La phase des prétraitements consiste à préparer l'image acquise à la phase suivante dont leur but est de réduire le nombre de données et n'en garder que les informations importantes et les plus significatives de la forme représentée et les opérations de prétraitement qui sont généralement utilisées : le redressement de l'écriture, le lissage et la réduction du bruit, la normalisation et la squelettisation. [1,11, 15, 16].

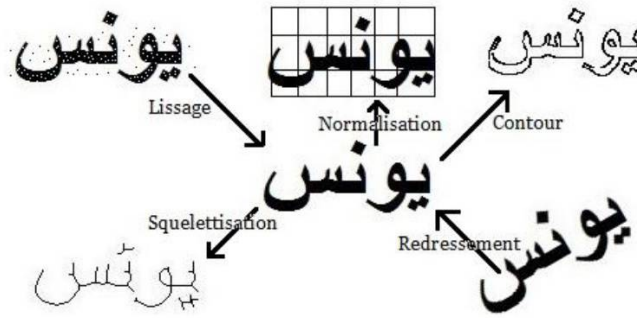


Figure 1.20: Différentes opérations de prétraitement [15]

7.2.1 Binarisation

La binarisation a pour but de convertir l'image numérique en niveau de gris issue du scanner en une image bimodale (composée de 2 valeurs 0 et 1, plus simple à traiter.) en utilisant un seuil correcte dans laquelle l'objet et le fond sont séparés., un simple seuillage global est utilisé après analyse de l'histogramme des niveaux de gris lorsque les documents papier sont de bonne qualité et de fond blanc. Tous les pixels plus lumineux que le seuil sont pris comme blancs, les autres comme noirs. Mais lorsque le fond est texturé ou si le document est dégradé (pliures, taches...), une analyse plus fine est nécessaire [11, 15, 16, 1].

L'opération de seuillage utilise la fonction de mappage suivante : Supposons que l'image d'entrée est $inp(x,y)$ et de sortie est $Out(x,y)$ donc [16] :

$$Out(x,y) = \begin{cases} 0 & \text{si } inp(x,y) < \text{seuil} \\ 1 & \text{si } inp(x,y) > \text{seuil} \end{cases}$$

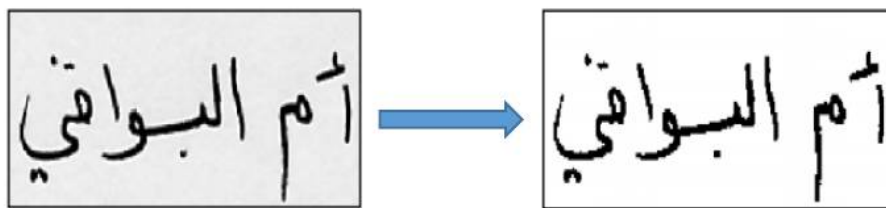


Figure 1.20: Exemple de binarisation

Il existe deux types de seuillage : **Seuillage global et seuillage adaptative** , Le premier type consiste à prendre un seuil identique pour toute l'image où chaque pixel de l'image est comparé à ce seuil : ceux de niveaux de gris inférieurs sont mis à 0 « noir », ceux supérieurs mis à 1 « blanc » et la deuxième consiste à faire varier le seuil localement d'une position à une autre, en fonction des niveaux de gris des pixels de document , Les méthodes basées sur le seuillage local sont plus robustes aux dégradations si la qualité du contenu n'est pas constante dans toute la page, des

taches parasites peuvent apparaître si le fond est bruité ou non homogène, et lorsque la page contient des encarts de couleur mais leur temps de calcul est plus important.

7.2.2 Lissage et Suppression du bruit

L'image d'entrée peut être entachée de bruits et contient une distorsion soit à une absence de points ou à une surcharge de points, dus lors de l'opération d'acquisition et à la qualité du document originale. Le lissage et le filtrage des bruits consiste à détecter et à éliminer les pixels du bruit qui le représentent par l'application des opérations qu'on appelle opérations de bouchage et de nettoyage. L'opération de nettoyage permet de supprimer les petites tâches et les excroissances de la forme, pour le bouchage, il s'agit d'égaliser les contours et de boucher les trous internes à la forme du caractère en lui ajoutant des points noirs (la valeur d'un pixel est remplacée par la valeur d'une fonction appliquée à ce pixel et à ses voisins).[11, 16, 1]



Figure 1.21: Exemple de Lissage et Suppression du bruit [4]

7.2.3 Normalisation

Cette opération a pour but de réduire tous les types de variations, et pour aboutir à des données normalisées pour rendre l'écriture la plus indépendante possible du scripteur.

La normalisation de la taille permet de ramener les images de mots à la même taille [11, 1].

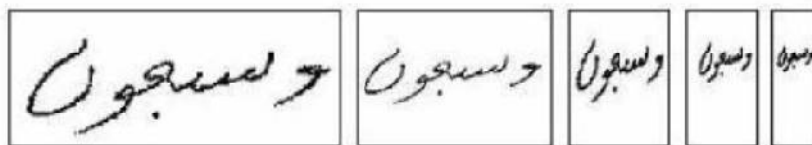


Figure 1.22: Exemple de mots manuscrits avec des tailles différentes [4]

7.2.4 Redressement de l'écriture " la correction de l' inclinaison "

l'inclinaison de l'image généralement est la conséquence d'un mauvais positionnement du document sur le scanner, elle est mener à des problèmes de segmentation du texte en lignes et en caractères. Pour une bonne exploitation et analyse mieux des documents, la redressement des images(appelée correction de "skew") est important, elle consiste à redresser horizontalement les lignes et corriger l'inclinaison, de telle sorte que les lignes de texte soient parallèles aux bords

et la pente d'un mot soit corrigé ou l'inclinaison des lettres soit redresser dans un mot afin de faciliter la segmentation.

il suffit d'appliquer une rotation isométrique de l'image d'angle θ . [11, 13, 15].

$$\begin{pmatrix} x' \\ y' \end{pmatrix} = \begin{pmatrix} \cos(\theta) & \sin(\theta) \\ -\sin(\theta) & \cos(\theta) \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}$$



Figure 1.23: L'inclinaison globale du texte par rapport au cadre de l'image [4].

7.2.5 Squelettisation :

La squelettisation est une opération qui permet de passer d'une image à sa représentation en "fil de fer" « ligne » pour être plus facile à traiter, Elle consiste à obtenir une épaisseur égale à 1 du trait d'écriture et de se ramener ainsi à une écriture linéaire représente l'information indépendamment de l'épaisseur initiale de l'écriture.

Le squelette doit préserver la forme, connexité, topologie et extrémités du tracé, et ne doit pas introduire d'éléments parasites ; et donc simplifier la représentation, et permet d'extraire des caractéristiques importantes, comme les intersections et le nombre de tracés, leurs positions relatives [11, 13, 15, 1].



Figure 1.24: Exemple de squelettisation [4]

7.3 Segmentation

A la suite des pré-traitements, il est nécessaire de segmenter l'image binaires traité pour obtenir des entités élémentaires "*imassettes*" qui seront analysées dans la phase ultérieure.

Il existe deux techniques permettant la mise en œuvre de la segmentation : segmentation explicite et segmentation implicite.

a) Segmentation explicite

C'est une stratégie de segmentation antérieure à la reconnaissance basée sur des critères topologiques et n'est pas remise en cause pendant la phase de reconnaissance, elle utilise des points caractéristiques dans le mot ; tels que les minima locaux du contour supérieur, les espaces ou encore les points d'intersection.

Elle doit d'être d'une grande fiabilité car la moindre erreur remet en cause la totalité des traitements ultérieurs.

b) Segmentation implicite

Elle consiste à effectuer un découpage a priori de l'image en intervalles de grandeur régulière d'après les modèles de lettres. Contrairement à la segmentation explicite, il n'y a pas de pré-segmentation du mot. La segmentation s'effectue pendant la reconnaissance et est guidée par cette dernière en cherchant dans l'image, des composantes ou des groupements des graphèmes qui correspondent à ses classes des lettres. [13].

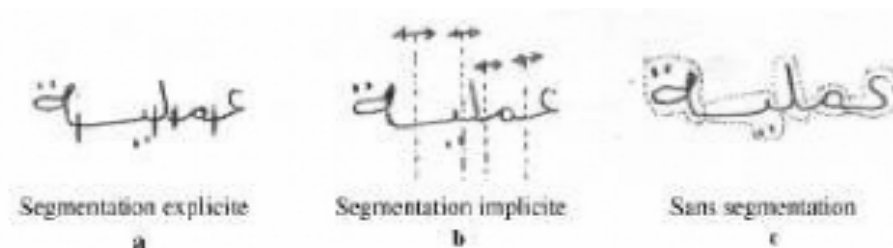


Figure 1.26: Les différents Modes de segmentations.

7.4 Extraction de caractéristiques

C'est une phase critique consiste à extraire l'information pertinente en représentant les données d'entrée (mots, caractères, graphèmes) en un vecteur de primitives de dimension fixe, pour différencier un objet d'un autre dans l'étape de la classification et la prise de décision.

En effet, un mauvais choix des primitives influence négativement et nettement les résultats même si on utilise un classifieur très performant.

7.4.1 Type de caractéristiques

Les différentes techniques d'extraction des caractéristiques sont classées en fonction des types des primitives :

a) Caractéristiques structurelles

Les primitives structurelles sont généralement extraites non pas de l'image brute, mais à partir d'une représentation de la forme par le squelette ou par le contour [4,18].

b) Caractéristiques statistiques

Le mot représenté par des mesures statistiques de l'image. Il peut utiliser la distribution des pixels dans différentes régions de l'image, ou bien des histogrammes (nombre de points noirs par colonne, par ligne, ou dans d'autres directions) [4,18].

c) Caractéristiques globales

L'image est considérée globalement sans chercher à distinguer les différentes zones , le codage fait globalement et ne fait pas intervenir la position spécifique d'éléments particuliers de l'image. [4,18].

d) Caractéristiques morphologiques

L'extraction des caractéristiques morphologiques s'appuie sur une étude des positions relatives des différentes composantes noires et blanches de l'image. Alors le mot est décrit en termes de composantes blanches et noires, de cavités (parties blanches partiellement entourées de noir) et de boucles (parties blanches entièrement entourées de noir) [4,18].

e) Caractéristiques métriques

Cette catégorie comprend des caractéristiques basées sur des mesures physiques de l'image. Outre des caractéristiques assez simples, comme la hauteur, la largeur et le rapport de ces deux grandeurs.. ect [4,18].

f) Caractéristiques adaptatives

Les caractéristiques adaptatives sont obtenues directement de l'image et requièrent une phase d'apprentissage. Autrement dit, le système opère sur une représentation proche de l'image d'origine et doit lui-même construire et optimiser l'extracteur de caractéristiques [4,18].

7.5 Phase de Classification

C'est la phase de passage de l'espace de codage vers l'espace de décision , La classification est la création d'une règle de décision pour transformer les attributs caractérisant les formes en appartenance à une classe et déterminer la classe de l'objet en entrée., elle regroupe deux tâches : l'apprentissage et la reconnaissance et décision [4].

7.5.1 L'apprentissage

Cette étape permet de construire un dictionnaire de prototype. Il s'agit de regrouper en classes plusieurs prototypes dont les caractéristiques se rapprochent. Il existe deux types d'apprentissages : supervisé et non supervisé [4].

7.5.2 Reconnaissance et décision

C'est l'étape la plus importante, La décision est l'ultime étape de reconnaissance, telle que à partir de la description en paramètres du caractère traité le système détermine le modèle le plus proche à la description de la forme du caractère parmi les modèles de référence en présence, ceux qui lui sont les plus proches [4].

La reconnaissance peut conduire à un succès si la réponse est unique et conduire à une confusion si la réponse est multiple. elle peut conduire également à un rejet de la forme si aucun modèle ne correspond à sa description [4].

7.6 POST-TRAITEMENTS

la classification peut aboutir à plusieurs candidats possibles, alors le post-traitement est la phase pour effectuer une sélection de la solution optimale en utilisant des niveaux d'informations plus élevés (syntaxiques, lexicales, sémantiques...) relatifs à la fréquence d'apparition des caractères dans une chaîne, aux lexiques et à d'autres informations contextuelles. Le post-traitement a pour but d'améliorer le taux de reconnaissance des mots.

Le post-traitement se charge également de vérifier si la réponse est correcte (même si elle est unique) en se basant sur d'autres informations non disponibles au classifieur [4].

8. Conclusion

Dans ce chapitre Nous présentons les principales propriétés morphologiques et typographiques de l'écriture arabe manuscrite, en suite nous présentons une étude générale sur la reconnaissance d'écriture manuscrite (les principales Approches et les aspects liés à ce domaine) et ensuite l'architecture générale d'un système OCR reflète d'une manière générale le processus de reconnaissance d'un Texte manuscrite. Donc la reconnaissance optique de l'arabe reste une tâche encore non résolue, les problèmes majeurs dans ce domaine se ramènent à la cursivité de l'écriture et à la sensibilité de certaines caractéristiques topologiques de l'arabe à la dégradation, en l'occurrence les points diacritiques et les boucles.

Chapitre 02: Segmentation et extraction des caractéristiques de l'écriture

1 Introduction

Dans le chapitre précédent, nous avons présenté une vue sur le domaine d'OCR et l'architecture globale d'un système OCR, ces différents types, les avancés dans ce domaine.

Le but de ce chapitre, est de présenter et expliquer la plus importante étape qui est la segmentation et extraction de caractéristiques de l'écriture et de présenter les techniques ou approches utilisées dans la littérature pour la segmentation du texte manuscrit cursive " un aperçu de l'état de l'art " en particulier la segmentation du mot en caractères, ainsi que les avantages et les inconvénients de chacune d'elles , afin de justifier le choix de notre première contribution, qui consiste en la proposition d'une approche de segmentation et les hypothèses sur lesquelles nous avons construit notre idée de segmentation multi-scans N-Scans

2 La Segmentation de documents et les différents niveaux de segmentation

En la reconnaissance hors-ligne de l'écriture manuscrite, les données à traiter sont sous forme des images , la segmentation est une phase très cruciale consiste à extraire des différents zones ou segments d'images et à segmenter le texte en sous-unités "des graphèmes, des lettres ou bien des sous-mots" utiles pour des traitements futurs. Elle a un effet sur le taux de reconnaissance, une bonne segmentation va déterminer les bons segments et par conséquent un bon taux de reconnaissance, à l'inverse, une mauvaise segmentation quant à elle va entraîner une chute du taux de reconnaissance.

Généralement, il existe quatre niveaux de segmentation, comme suit:

- ✓ Segmentation de la page,
- ✓ Segmentation de texte en lignes,
- ✓ Segmentation de lignes en mots,
- ✓ Segmentation de mots en caractères.

La figure 2.1 montre le passage entre ces étapes de segmentation.

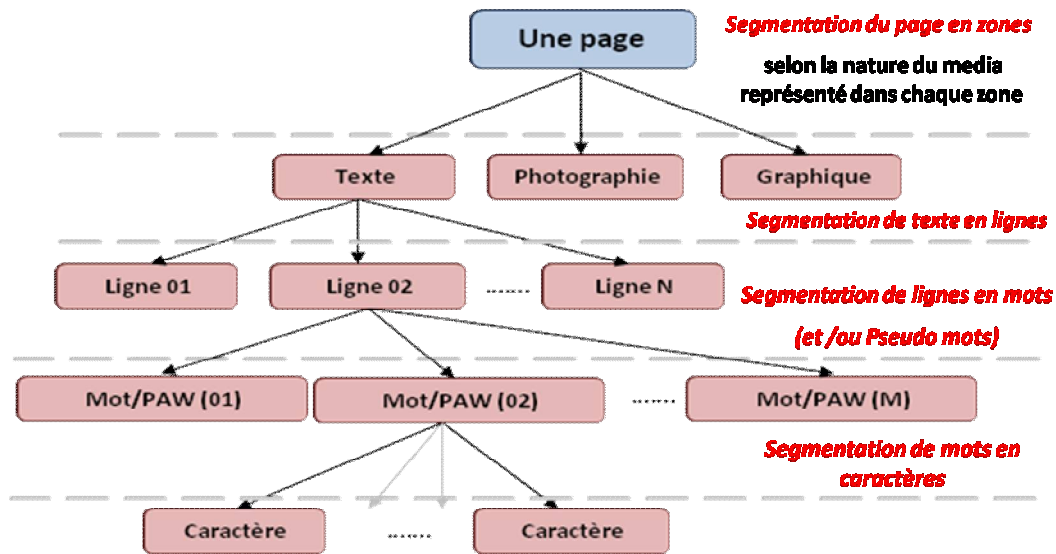


Figure 2.1 : Les différents niveaux du processus de segmentation [2]

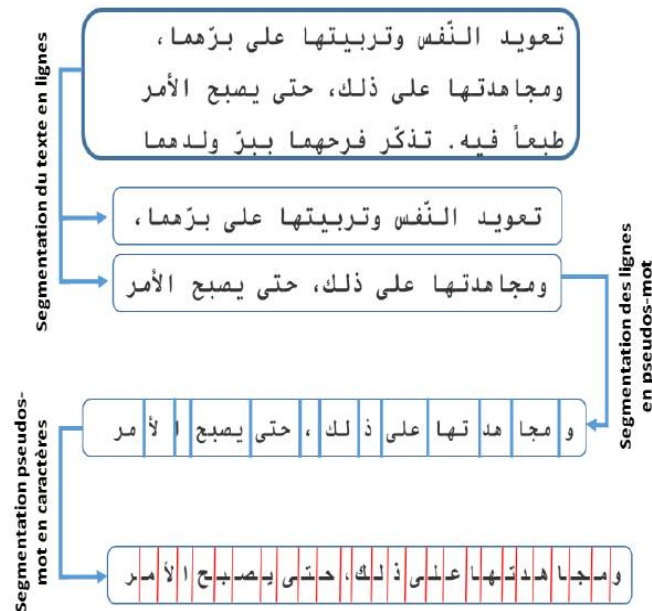


Figure 2.2 : Exemple illustre le passage entre les étapes de segmentation

2.1 Segmentation de la page :

Pour reconnaître un texte, il faut son extraction de la page, alors la segmentation de la page est très importante, elle consiste à retrouver la structure physique du document en délimitant les différentes parties homogènes (les graphiques, formules mathématiques, blocs de texte ...) et localiser dans chaque page les zones d'information conformément à leur apparence physique (voir figure 2.3).

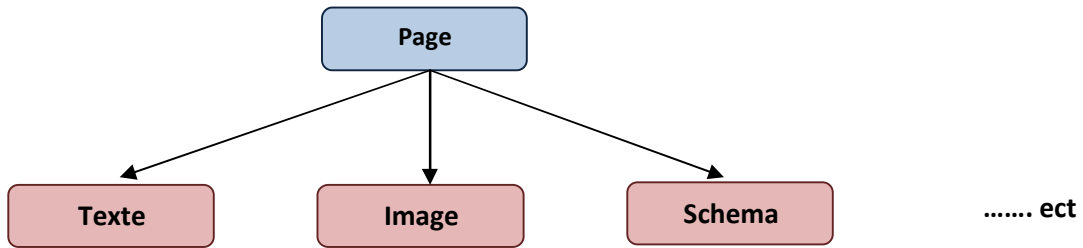


Figure 2.3 : Détection des zones d'une page de document [2,3,6].

2.2 Segmentation de texte en lignes:

il existe certaines méthodes utilisées pour localiser les lignes de texte à cet effet, le plus utilisé c'est la projection horizontale pour déterminer les lignes d'un texte donné.

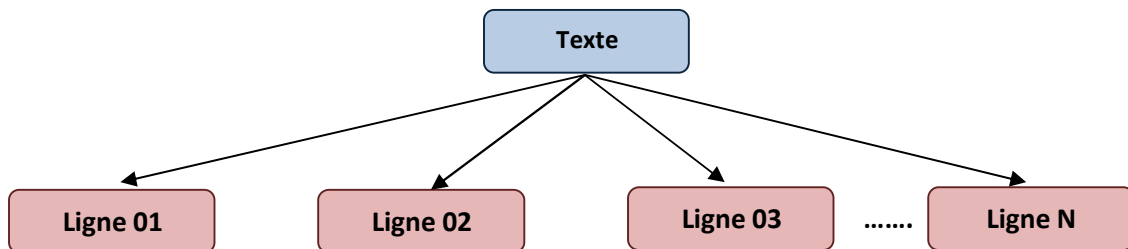


Figure 2.4 : Segmentation du texte en Lignes [2,3,6].

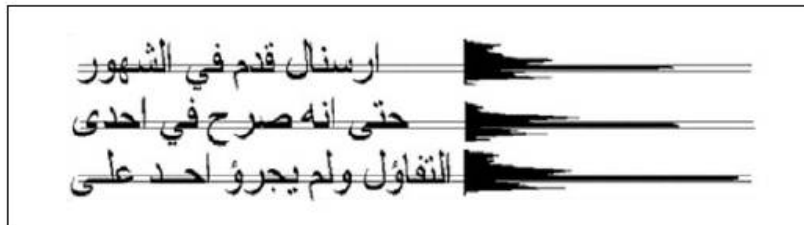


Figure 2.5 : exemple illustre la Segmentation du texte en Lignes par projection horizontale.

2.3 Segmentation de lignes en mots:

Après la segmentation des lignes, la ligne est segmentée en mots. Cette étape consiste à extraire les différents mots ou les pseudo mots d'une image de ligne en se basant sur l'espace entre les mots et les espaces dans les mots elles même mot en sous-mots.

Il y a deux approches de segmentation de lignes en mots : une approche descendante il faut déterminer l'histogramme des projections verticales des lignes pour détecter les espaces entre les mots et pouvoir les séparer. Cependant, cette méthode n'est pas efficace dans le cas où les PAWs se chevauchent verticalement. et une approche ascendante d'autres techniques sont utilisées telles que : le suivi du contour, détermination du squelette ou la détermination des composantes connexes . . .

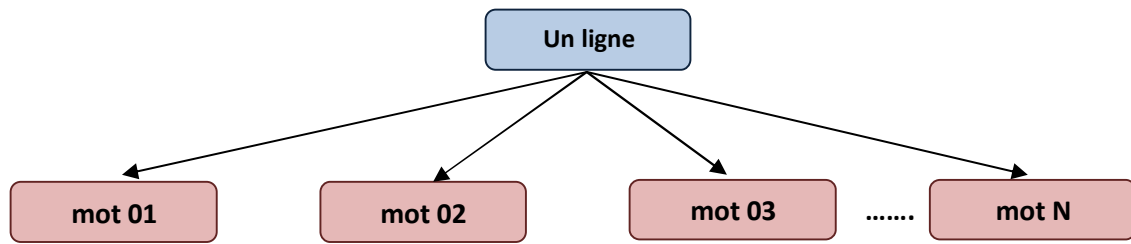


Figure 2.6 : *Segmentation du ligne en mots [2,3,6].*

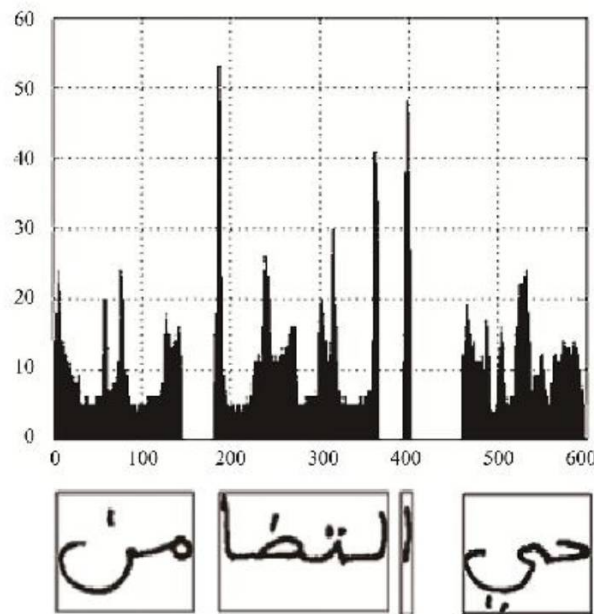


Figure 2.7 : *Exemple de Segmentation d'une Ligne en PAW par projection verticale.*

2.4 Segmentation de mots en caractères:

La Segmentation de mots en caractères (ou en graphèmes) est l'étape la plus délicate dans tout le processus d'un système de reconnaissance d'écriture. Une erreur dans la segmentation des caractères produira des erreurs dans la reconnaissance.

Cette étape implique la segmentation du mot en caractères(ou en graphèmes) individuels comme des sous-images de symboles individuels ".

La tâche de segmentation consiste à localiser deux points de segmentations, puis extraire le caractère entre eux. les points de segmentation sont identifiés à la fin d'un caractère et au début de la suivante. Les difficultés rencontrées à ce niveau sont souvent plus complexes à cause de la diversité des formes du caractère arabe, de la courte liaison qui existe entre les caractères successifs, de l'allongement des ligatures horizontales et de la présence des ligatures verticales

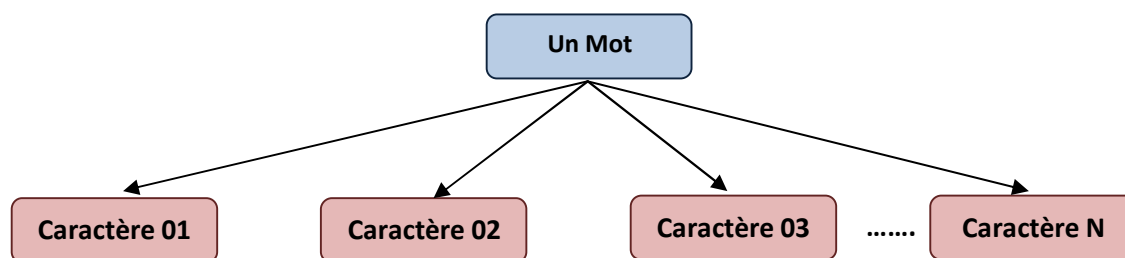


Figure 2.8 : Segmentation d'un mot en caractères [2,3,6].

3 Les différentes approches de segmentation

Il existe différentes approches permettant la mise en œuvre de la segmentation, il les classifie trois stratégies de segmentation plus d'autres méthodes hybrides à base des trois stratégies de base [18] :

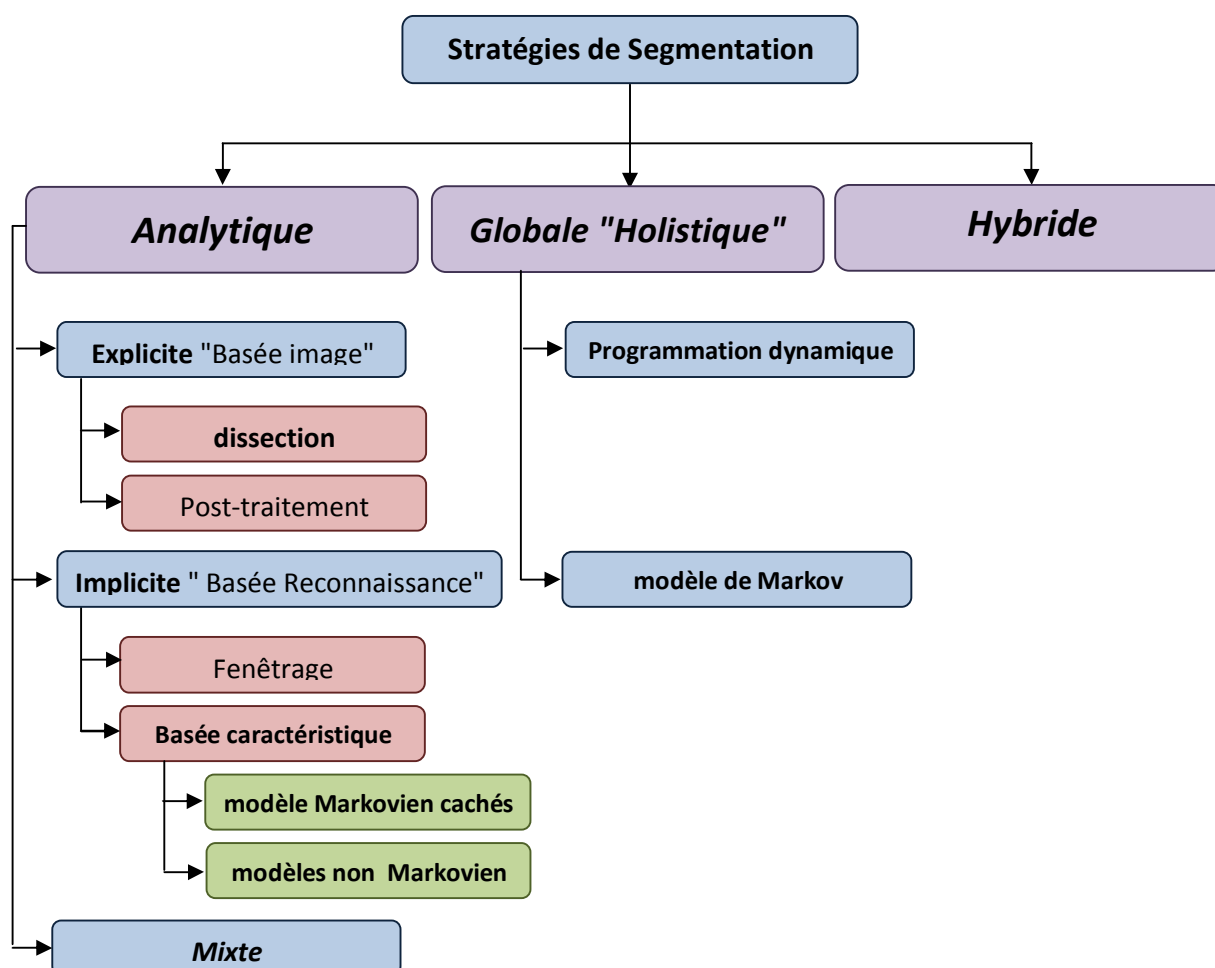


Figure 2.9 : Organigramme des méthodes de segmentation selon R.G. Casey [1,2,3,15].

3.1 Approches analytique à segmentation explicite "L'approche classique":

Dans laquelle les segments sont identifiés à base de propriétés de ressemblance des caractères par l'application d'une technique de découpage dite "dissection" selon des caractéristiques générales et des critères topologiques ou morphologiques sans invoquer la classification.

Le résultat de cette étape est la segmentation du mot en entités de base appelées "graphèmes", les graphèmes générés sont reconnus individuellement. Ensuite, le mot est reconstruit à base des informations contextuelle [2,3].

Dans cette approche, il existe deux méthodes permettant de choisir la segmentation finale de la séquence de lettres: les méthodes dites de segmentation puis reconnaissance et les approches de segmentation/reconnaissance. Les premières choisissent les meilleurs points de segmentation sans l'aide de la reconnaissance, alors que les méthodes de segmentation/reconnaissance sont basées sur l'utilisation du moteur de reconnaissance de lettres pour valider et classer les hypothèses de segmentation[2,3].

3.2 Approches analytique à segmentation implicite "basée reconnaissance" :

L'approche à segmentation implicite (ou continues) utilisé pour surmonter la difficulté du choix des points de segmentation, on utilisant le principe que tous les points du tracé sont des points de segmentation potentiels et la segmentation et la reconnaissance sont réalisées conjointement (de manière intégrée). Le système recherche dans l'image, des composants ou des groupements de graphèmes qui correspondent à ses classes de lettres.

3.3 Approche analytique mixte " Sur-Segmentation ":

Est une approche mélange des deux stratégies précédentes. D'abord, une phase de pré-segmentation est appliquée pour sur-segmenter l'image. Puis la meilleure combinaison de reconnaissance est choisie, en considérant les combinaisons possibles de 1, 2, . . . n graphèmes.

3.4 Approche globale "Méthodes holistiques" :

Dans cette approche le mot entier est reconnu comme un tout " comme un entité " par sa forme générale, sans le besoin de segmentation en caractères (ni explicite, ni implicite), la reconnaissance dans cette approche dépend d'un lexique (une liste de modèles de mots) , et chaque mot du lexique utilise un modèle. Ces méthodes est mieux applicable où le lexique est défini statiquement et restreint comme c'est le cas dans le traitement automatique des chèques.

L'inconvénient majeur de ces méthodes est que leur utilisation est toujours retreinte à un lexique limité et prédéfini.

3.5 ***Approches hybrides:***

Les approches hybrides combinent plusieurs stratégies de reconnaissance différentes afin d'exploiter les forces et les faiblesses d'approches complémentaires (stratégies ascendantes et descendantes). la force de Cette approche est de combiner les avantages des deux méthodes pour atteindre de meilleurs résultats. Généralement, les approches ascendantes sont adoptées en écriture parce qu'ils sont capables de mieux faire face au le bruit et la variation de l'écriture [2,3].

4 ***Des techniques appliquées pour la segmentation de l'écriture cursive:***

La segmentation de l'écriture cursive est une opération très délicate et coûteuse, leur résultat influe considérablement sur le rendement des prochaines phases. Dans cette section, nous présentons des méthodes de segmentation. Ces méthodes sont classés selon qu'ils sont utilisés pour segmenter un texte en lignes ou pour segmenter un mot en caractères.

4.1 ***Techniques appliquées pour la segmentation de texte en lignes:***

L'une des tâches les plus importantes et les plus difficiles dans l'analyse de documents textuels est de segmenter sans erreurs un texte en lignes précises, Plusieurs méthodes ont été développer dans ce contexte, parmi les on trouvant :

4.1.1 ***Segmentation par histogrammes de projection horizontale:***

C'est la éthode la plus utilisée par les systèmes de reconnaissance, Elle consiste à faire une segmentation du texte en combinant l'utilisation d'une projection horizontale, et la classification des lignes selon leurs poids (*voir la figure 2.11*) [3].

Selon cette méthode, le document acquis par scanner peut contient trois types de lignes :

1. Lignes diacritiques : elles se caractérisent par des petits poids. Ces lignes sont considérées comme bruit [3].

2. Lignes normales : elles possèdent des poids moyens. Cette approche calcule les poids moyens de toutes les lignes du document [3].

3. Lignes chevauchées : elles possèdent des grandes hauteurs, ce sont le résultat du chevauchement des caractères, mais aussi des caractères collés des lignes adjacentes (*voir la figure 2.10*). Pour déterminer le nombre des lignes dans un bloc chevauché il faut diviser la hauteur du bloc sur la hauteur d'une ligne, et ensuite diviser le bloc sur le nombre des lignes obtenu [3].

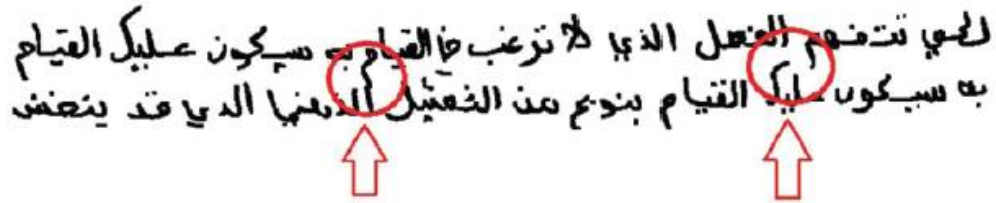


Figure 2.10 : Exemple de chevauchement entre des lignes du texte [3].



Figure 2.11 : Exemple de segmentation du texte en lignes par projection horizontale.

4.1.2 Segmentation basée sur les fenêtres glissantes:

Elle permet de diviser l'image traitée en colonnes par une fenêtre de taille fixe (également peut être variable). Cette fenêtre est déplacée avec un pas. Ensuite, un diagramme de projection horizontale est calculé à chaque déplacement de la fenêtre dans le but de détecter les lignes du texte [2,3]. Cette technique est simple, robuste au bruit et indépendante des connexités. Néanmoins, la détermination de la taille de la fenêtre appropriée n'est pas facile. En plus, quand il y'a beaucoup de segments liés, on risque de joindre deux entités différentes[2,3].

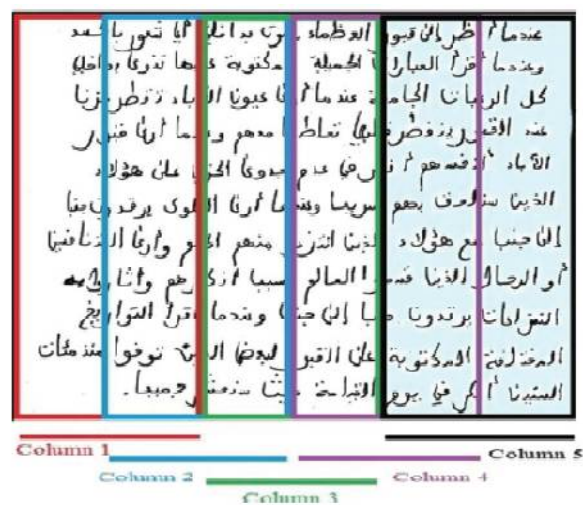


Figure 2.12 : Segmentation à base de fenêtre glissante [2,3].

4.1.3 Segmentation par écoulement d'eau:

Cette algorithme basé sur la supposition que des débits d'eau hypothétique à partir de quelques angles de l'image du document de gauche à droite et de haut en bas. Ce qui engendre des régions mouillées et d'autres non mouillées. le flux d'eau hypothétique devrait combler les écarts entre les lignes du texte consécutives. Par conséquent, les régions non mouillées restantes sur l'image indiquent les lignes du texte. Les régions non mouillées sont étiquetées afin d'extraire les lignes du texte. Une fois cet étiquetage terminé, l'image va être divisée en deux types de bandes. Le premier contient les lignes du texte. Le second présente l'espacement entre les lignes [3].

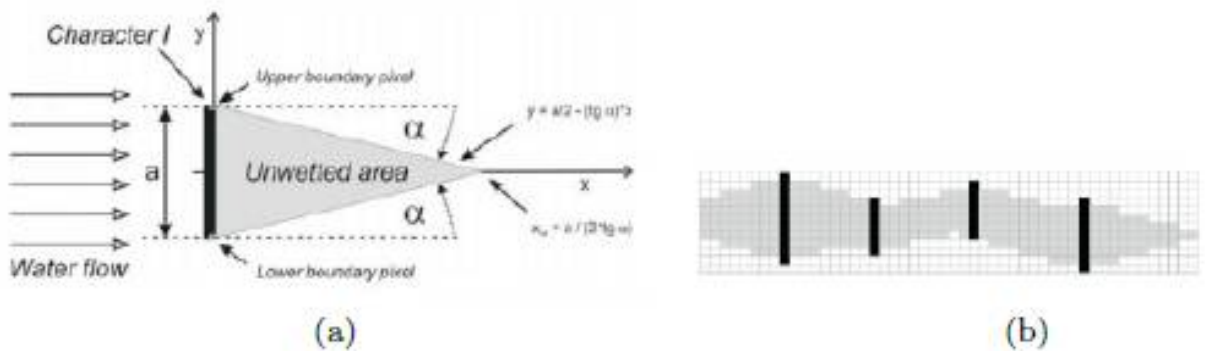


Figure 2.13 : Exemple de la Segmentation par écoulement d'eau entre des caractères ligaturés [3].

4.1.4 Segmentation basée peinture Piece-Wise:

Cette technique permet de décomposer verticalement les blocs de texte traité en bandes parallèles. Ou chaque ligne de chaque bande est colorée par une intensité de gris, qui est la valeur d'intensité moyenne des valeurs de gris de tous les pixels présents dans cette bande de lignes. (voir la figure 2.14). Les espaces blancs / noirs dans chaque bande de l'image lissée sont analysés pour obtenir une courte ligne de séparation (PPSL : Piece-wise Potential Separating Line) entre deux espaces noir consécutifs. Les PPSLs sont concaténées pour produire une segmentation de texte en lignes (voir la figure 2.15) [3].



Figure 2.14 : Exemple de segmentation avec une technique Pice-Wise [3].

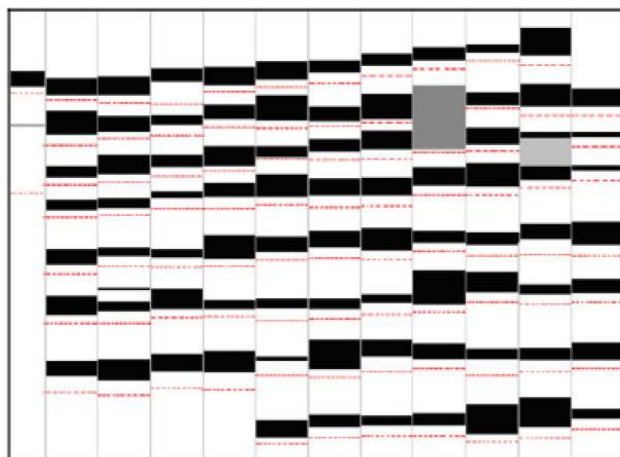


Figure 2.15 : Exemple d'image de texte lissée [3].

4.2 Techniques appliquées pour la Segmentation en mots et caractères:

A ce niveau le système analyse l'image du texte ou l'image d'un segment de ligne pour extraire les différents mots ou/et caractères composants le texte. Les méthodes de ces classes peuvent être classées en deux catégories : *Segmentation sans classification* et *Segmentation avec classification*.

4.2.1 Segmentation sans classification:

ces méthodes n'utilise pas des bases de connaissances et des classificateurs afin d'alléger la segmentation. Généralement, elles sont basées sur les histogrammes de projections verticales et des balayages avec marquages colorés de l'image de texte, ou même sur des balayages d'une représentation en squelette en contour de l'image source.

a) Segmentation basée sur la squelettisation:

La squelettisation est une technique peut aider à résoudre le problème de segmentation de caractères parceque la squelette contient et présente presque tous les informations essentielles d'une forme [3,18,27]. Dans cette technique, le système créer une image contenant le squelette du texte à traiter et à partir du squelette généré cherche d'identifier les points de liaison entre les caractères en introduisant des calculs de courbures et d'angles basées sur des seuils prédéfinis. *La figure 16* montre un exemple d'un mot arabe et son squelette [3,18,27].

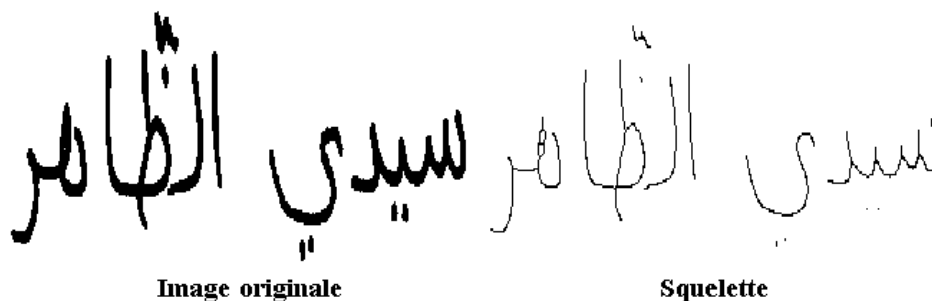


Figure 2.16 : Technique de Segmentation basée sur la squelettisation[18].

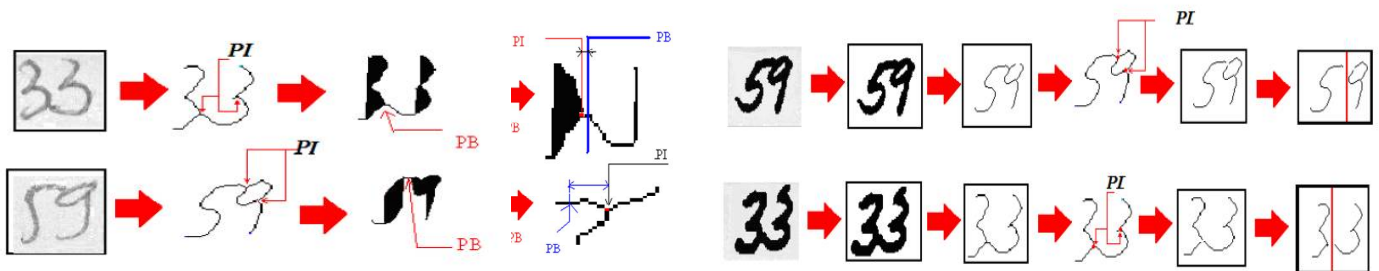


Figure 2.17 : Exemple illustre l'utilisation d'une Segmentation basée sur la squelettisation.[27]

b) Technique de segmentation basée sur le traçage des contours :

C'est une technique basée sur le traçage des contours d'un mot pour la segmentation . Les résultats de cette technique sont très remarquables par rapport aux autres techniques par ce que Le contour reste le meilleur choix à utiliser pour isoler les caractères connectés, et ceux en chevauchement [18,3,16] .

Le principe , le système génère à partir de l'image source le tarçage du contour extérieur du corps principal du mot. ensuite, le système fait un balayage et une analyse sur le traçage du contour et les règles topologiques est réalisé afin de détecter les points de liaison des caractères ou graphèmes en se basant sur les locaux du contour inférieur de chaque pseudo-mot (voir *la figure 2.18*). Ces derniers sont faciles à ajuster lorsque la qualité de l'écriture est bonne, comme ils peuvent y avoir des comportements erratiques lorsque l'écriture est de mauvaise qualité [18,3,16].

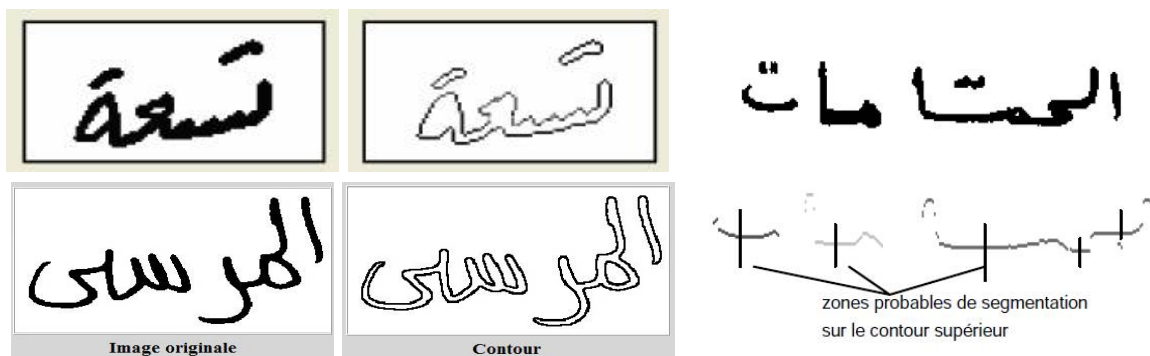


Figure 2.18 : Technique de segmentation basée sur le traçage des contours [18].

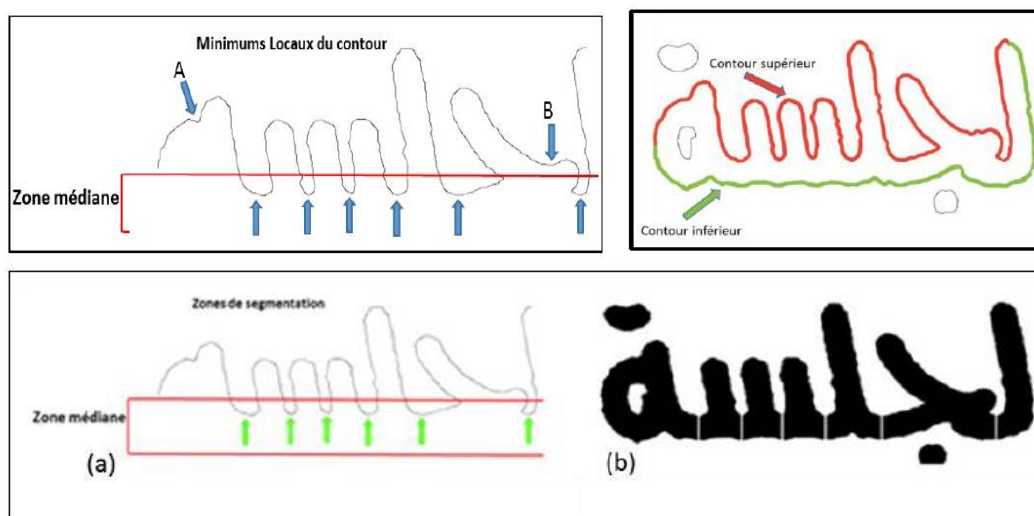


Figure 2.19 : Illustration de la segmentation basée sur le traçage des contours.

Parmi les problèmes rencontrés lors de l'analyse du contour pour la segmentation est la discontinuité de celui-ci, et cela peut être dû au bruit, c'est pour cela que la phase de prétraitement doit éliminer au maximum le bruit présent dans l'image. Un autre problème dans cette technique est la détection de la ligne de base devient très difficile [18,3].

c) *Segmentation à partir des histogrammes de projection verticales*

La projection verticale est utile pour segmenter les mots, les sous-mots et les caractères, tandis que la projection horizontale est habituellement utilisée pour la segmentation de ligne et l'extraction de ligne de base. Le but de cette méthode est de simplifier un système OCR en réduisant les informations bidimensionnelles (2D) en une seule dimension (1D), elle est basée sur le calcul des histogrammes de projection vertical pour identifier ces différents caractères en se basant. Les droites de segmentation sont celles qui contiennent le moins de pixels noirs, avec une contrainte d'espacement régulier dans l'image (voir **figure 2.20**). Néanmoins, Cette méthode montre des limites lorsque les lettres sont très proches ou enchevêtrées [2,3,6].

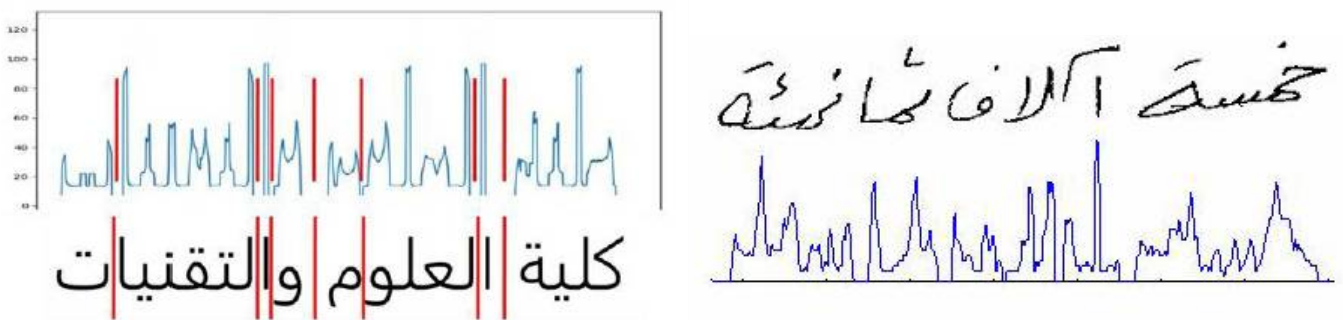


Figure 2.20 : La segmentation d'une ligne de texte en pseudo-mots par Projection verticale.

d) *Segmentation basée sur des réservoirs d'eau "Algorithme Water Reservoir"*.

Cette technique est utilisée dans la segmentation des manuscrits, Les caractères sont segmentés à base des points de la région de base du réservoir.

Le principe de cette méthode est basée sur la recherche des « Water reservoir » ou réservoirs d'eau, Un « water reservoir » est utilisé pour illustrer la région où les lettres se touchent. Les réservoirs sont obtenus en considérant les zones résultantes d'un lâcher d'eau sur la composante, depuis le haut ou le bas de celle-ci (*Figure 2.21*). si l'eau coule de haut en bas du caractère, les régions de la cavité (trou, vide) des caractères où l'eau est stockée sont considérées comme des réservoirs (*Figure 2.22*). les réservoirs de haut sont les réservoirs obtenus quand l'eau coule du haut en bas. les réservoirs utilisées seulement ceux qui sont plus grand d'un seuil vont être utilisés dans des traitements futurs.[3,27]

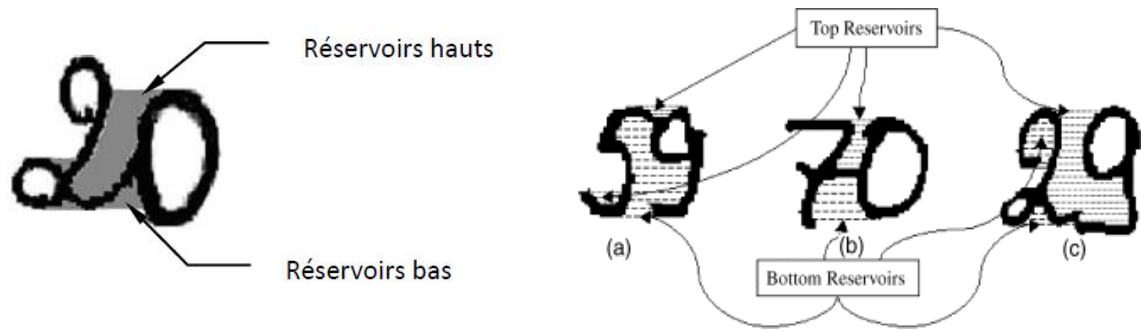


Figure 2.21 : Détection des réservoirs "haut" et "bas" [27].

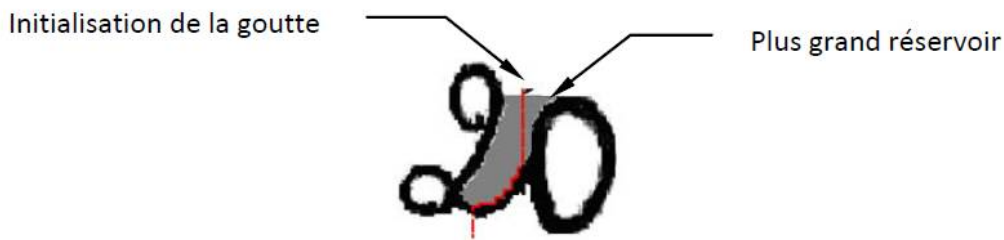


Figure 2.22 : Segmentation "Drop fall" ascendant basée sur les réservoirs d'eau [27]

4.2.2 Segmentation avec classification:

Les techniques de segmentation de cette catégorie font introduire l'utilisation d'une base de connaissance (des différents caractères ou même graphèmes) avec un classifieur (tel que ANN, HMM, . . . etc) afin d'obtenir un meilleur taux de segmentation.

a) Segmentation par propagation d'affinité (*Affinity Propagation*):

C'est une méthode de clustering consiste à trouver et de regrouper un ensemble d'exemplaires (prototypes) dans les données et d'assigner d'autres points de données aux exemplaires les plus appropriés, puis à partitionner l'ensemble de données en clusters. L'objectif de propagation d'affinités c'est de maximiser la somme des similarités (affinités) entre les points de données et ses exemplaires. [3,24]

Par exemple l'extrait et le regroupement les composantes connexes en clusters pour segmenter le texte en mots, ces derniers correspondent aux composantes principales des pseudo-mots avec leurs diacritiques (*figure 2.23*). tous les espaces entre deux clusters successifs sont considérés comme des candidats de séparation inter-mots et que la présence du caractère *alif isolé* donne une information importante pour guider une segmentation en mots par l'utilisation un réseau de neurones pour classer les espaces entre clusters en deux classes : séparateur de mots ou non. [24]

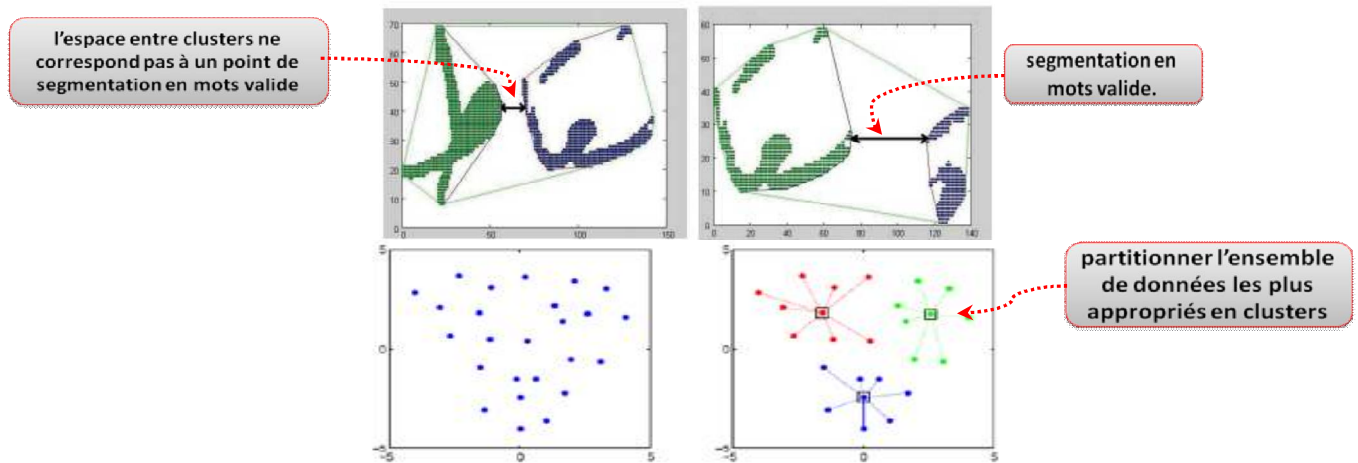


Figure 2.23 : Exemple illustre Une solution d'exemplaires basée sur le clustering par AP [3,24]

b) Segmentation à base de forêts aléatoires:

C'est une approche de segmentation des mots manuscrits arabes en utilisant les forêts aléatoires (**RF**: *Random Forests*). Elle fonctionne en deux phases : La première phase dissèque les mots en différents composants sans laisser des caractères liés, ensuite la l'extraction des caractéristiques de zoning, en raison de leur force dans la capture des caractéristiques locales et globales, pour les utiliser dans l'apprentissage des classificateurs efficaces de forêts aléatoires. Ensuite, les classificateurs vérifient chaque segment de caractère et valide les points de segmentation (SPs: segmentation points) [3].

c) Segmentation à base de Spline en Plaque Mince

C'est une approche de segmentation de texte en lignes. L'approche segmente les composants connexes (CCs) en deux étapes. Premièrement, elle cherche à trouver dans un dictionnaire un modèle similaire avec une segmentation correcte pour un CC. La recherche est basée sur le contexte de la forme et La méthode de Spline en Plaque Mince (SPM). Ensuite, elle segmente la CC en utilisant le point central des pièces de modèle similaire [3].

d) Segmentation à base de réseaux de neurones

Les réseaux de neurones (perceptrons multicouches) sont les classificateurs les plus utilisés dans le domaine de l'intelligence artificielle, ils ont connu un essor important grâce l'algorithme de retro-propagation du gradient. Ce classifieur a trouvé application dans beaucoup de domaines tels que la reconnaissance de caractères, reconnaissances de visages...etc. Ce mécanisme discriminant a été intégré comme outil de vérification des formes[05,16].

Un réseau de neurones est un graphe orienté pondéré. Les nœuds de ce graphe sont des automates simples appelés neurones formels. Les neurones sont dotés d'un état interne, l'activation,

par lequel ils influencent les autres neurones du réseau. Cette activité se propage dans le graphe le long d'arcs pondérés appelés liens synaptiques

En OCR, les primitives extraites sur une image d'un caractère (ou de l'entité choisie) constituent les entrées du réseau. La sortie activée du réseau correspond au caractère reconnu. Le choix de l'architecture du réseau est un compromis entre la complexité des calculs et le taux de reconnaissance.

Par ailleurs, le point fort des réseaux de neurones réside dans leur capacité de générer une région de décision de forme quelconque, requise par un algorithme de classification, au prix de l'intégration de couches de cellules supplémentaires dans le réseau. [5,3,20,16].

4.2.3 Comparaison des méthodes de segmentation :

Méthode de segmentation	Critère de comparaison			
	seuils	Balayage	Support des caractères enchevêtrés	reconnaissance
A partir de squelette	prédéfinis	Rapide	Oui	Non
A partir de contour	Dynamiques	Lourd	Oui	Non
A partir des histogrammes	prédéfinis	Très rapide	Non	Non
Basée sur les réservoirs	prédéfinis	Lourd	Oui	Non
Basée sur fenêtres glissantes	prédéfinis	Lourd	Oui	Non
Propagation d'affinité	prédéfinis	Très Lourd	Oui	Oui
Basée forêts aléatoires	prédéfinis	Très Lourd	Oui	Oui
Basée sur les Spline en Plaque Mince	prédéfinis	Très Lourd	Oui	Oui
basée sur les réseaux de neurones	Dynamiques	Très Lourd	Oui	Oui

Tableau 2.1:: Comparaison des méthodes de segmentation en mots et caractères [6]

5 *Conclusion*

La segmentation de l'écriture cursive est une phase très cruciale parmi les phases de la reconnaissance du manuscrit parce qu'elle extrait les principaux segments d'image nécessaire pour des traitements ultérieurs. Le taux de réussite de la segmentation augmente progressivement à partir de l'écriture imprimée à l'écriture manuscrite jusqu'à l'écriture manuscrite cursive où la difficulté devient plus importante.

Dans ce chapitre nous avons exposé les différents niveaux de segmentation et des méthodes utilisées dans ce processus, ces méthodes ont connu beaucoup de progrès ces derniers temps mais elles restent insuffisantes parce que les mots arabes manuscrits, les sous-mots, et les limites entre les caractères sont difficiles à identifier avec l'existence de coupures et des chevauchements dans les segments des caractères ce qui rend la segmentation des documents Arabes très difficile.

La difficulté dans la réalisation d'une segmentation performante dépend généralement de la nature du document à lire et de sa qualité.

Chapitre 03 : Technique de segmentation n-scans à deux phases

1. Introduction

Malgré l'essor de développement et l'évolution des technologies de reconnaissance de la langue arabe AOCR, elle reste insuffisant et presque inadaptable pour l'écriture arabe artistique parce qu'il est très difficile d'identifier les mots et les pseudo-mots et les limites entre les caractères avec l'existence des chevauchements entre les caractères et des caractères en ligature verticale et des intersections entre les segments des lettres ...etc.

Dans ce travail, notre contribution est de proposer une méthode de segmentation structurelle adaptable et confient à la segmentation de l'écriture arabe manuscrite artistique, Où nous avons cherché à améliorer dans une méthode de segmentation basée multi-scans (N-Sans) et de proposer des nouveaux règles de segmentation.

Cette méthode est applicable à l'écriture arabe manuscrite mais n'est pas adaptable sur l'écriture artistique, ceci est dû à la difficulté illimité et aux niveau de complexités élevé (les chevauchements, les interdépendances entre les mots et les lettres et à l'abondance des signes diacritiques dans l'écriture artistique.

2. Les Problématiques inhérentes posée par l'écriture du cursif artistiques :

Une image d'un texte ou d'une mot cursive est représentée par un signal bidimensionnel dans lequel, Ce n'est pas possible d'extraire les informations d'ordonnancement décrivant la trajectoire d'un écriture cursive, le problème majeur de tout méthode de segmentation est comment déterminer les points potentielle de la localisation précise du début et de la fin d'une lettre, Et comment déterminer qu'un point du tracé d'écriture d'un mot appartient à cette caractère ou à un autre caractère ,qui est s'avère très difficile à réaliser, voire impossible [3].

La seule façon de restituer l'ordre cohérent du tracé manuscrit est de trouver et de développer une méthode fiable de segmentation les mots en lettres.



Figure 3.1: La cursivité d'un mot arabe manuscrite

Généralement, l'écriture cursive artistique comme toute écriture manuscrite se caractérise par des problèmes suivants :

1. L'écriture arabe artistique est par nature cursive et repose plus sur la beauté des lettres et de l'écriture de manière esthétique sans préserver la forme standard du caractère. Une lettre ne possède pas une forme fixe et peut être à des formes différentes, ie un caractère n'est pas tracé de la même manière. (comme illustre **la figure 3.1**) .
2. Les documents arabes artistiques sont caractérisés par la présence fréquente des signes « El Chakle » et points diacritiques et des marques additionnelles au-dessus ou au-dessous des caractères comme des entités isolées, pas seulement pour distinguer entre les caractères possédant la même forme de base et pour modifier l'accent du caractère et de donner aux mots leur signification réelle dans le texte, mais également pour donner une élégance supplémentaire et de remplir joliment le dessin de l'écriture artistique. Ces signes sont parfois chevauchés avec une lettre, ce qui rend le processus de segmentation très difficile et change éventuellement de la forme du caractère à une autre forme inconnue. (comme illustre **la figure 3.1**).
3. Les points diacritiques peuvent être loin au-dessus ou au-dessous de la forme de base ce qui parfois crée une confusion entre les caractères adjacents, ce qui rend la segmentation des documents Arabes très difficile à distinguer ou conduit parfois à la confusion dans la distinction entre des caractères similaires de la même forme. (comme illustre la figure 3.1 (d))
4. *L'Espacement entre les pseudo-mots et même entre les caractères est irrégulier*: l'écriture artistique est caractérisée généralement par le manque de respect pour l'espacement entre les lettres et les mots. Elle est également caractérisée par le manque de respect de l'ordre dans l'écriture des mots constituant le texte (elle ne respecte pas la successivité entre les mots constituant le texte l'un après

l'autre de droite vers a gauche) mais écrire habituellement le mot verticalement au dessus sur autres avec des chevauchement et des intersections entre les caractères qui les constituent.

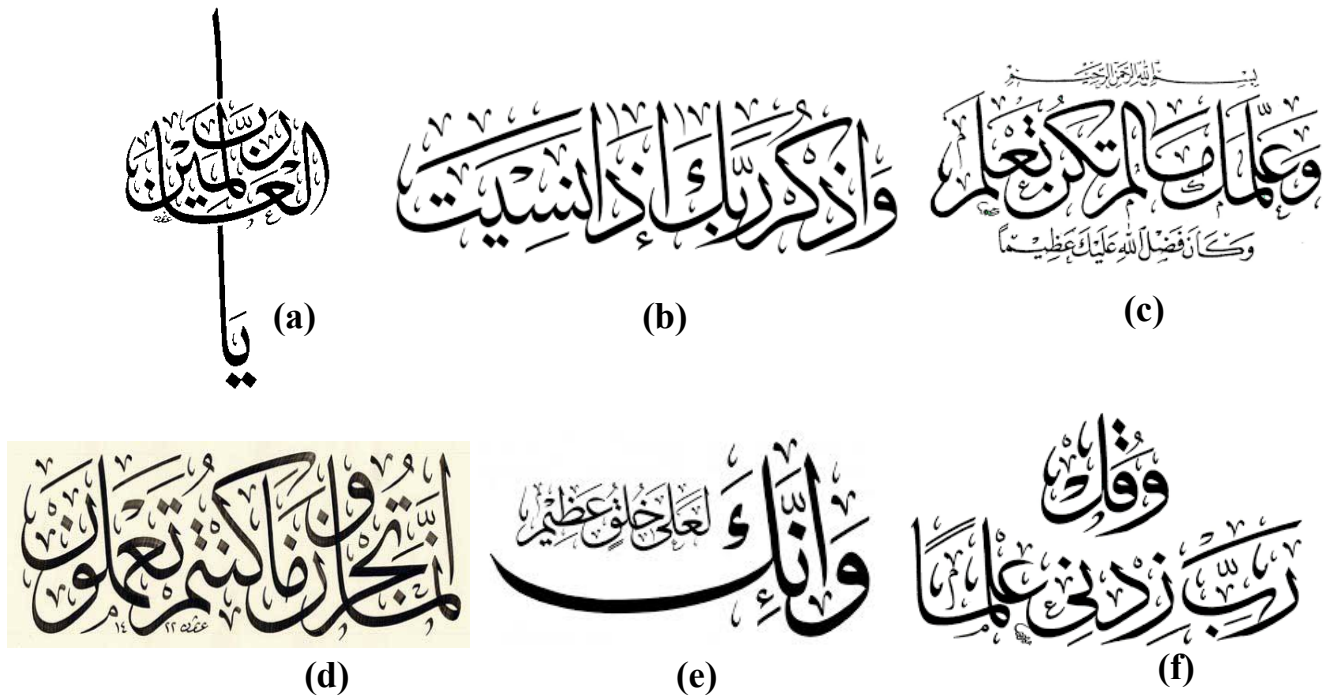


Figure 3.2: Exemples Illustre les caractéristique et des styles d'écriture artistique

5. En écriture arabe artistique présente souvent un grand problème dans le domaine de reconnaissance et la segmentation, telle que il y a des chevauchements et des intersections entre les mots, cela due à la non respect de l'espacement entre les lignes et l'absence de ligne de base d'écriture, et la superposition entre des mots que les constituants et l'existence des chevauchements entre les caractères constituant les mots avec leurs extensions descendantes à cause de la préservation de l'esthétique de l'écriture et du style d'écriture., ceci rendre presque tous les méthodes de segmentation développées sont limités et pas appropriés à les appliquer pour l'écriture arabe artistique.

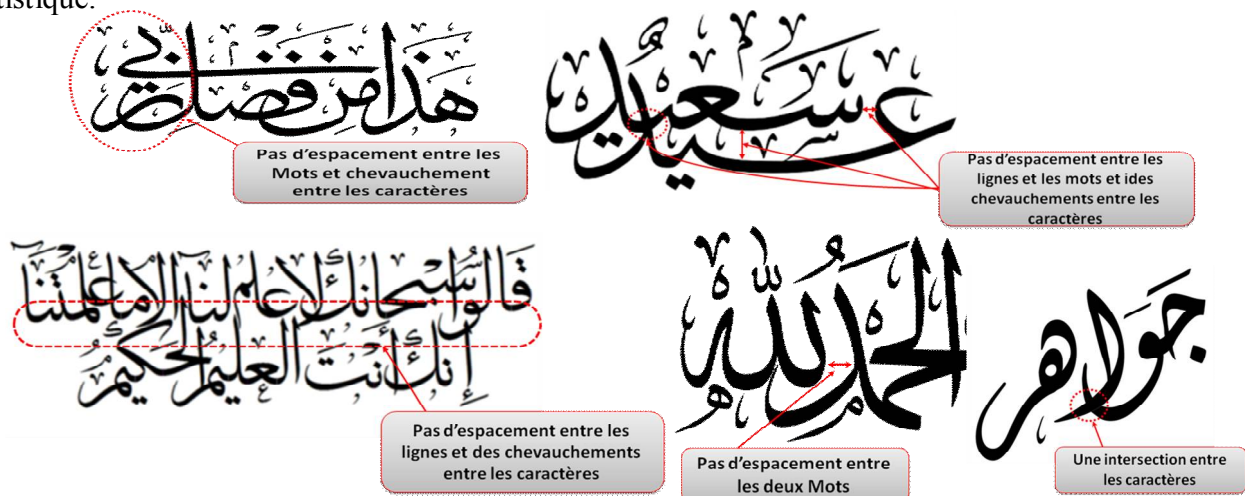


Figure 3.3: Des exemples Illustre les problèmes liés d'écriture artistique

6. En écriture arabe artistique, Les cercles des caractères peut être avec des segments liés. les caractères s'écrivant avec des cercles tels que : ق, ف, ع, ه, م, peuvent avoir le problème de liaison des segments composants ces cercles .

7. **Les Caractères sous-segmentés:** Un caractère sous-segmenté est le résultat des liaisons supplémentaires entre les segments (traits) représentant le caractère. Nous pouvons appeler aussi chevauchement de caractères.

8. **Caractères sur-segmentés:** Parmi les problèmes de l'écriture manuscrite artistique, l'existence des petites coupures dans les segments représentant le caractère ce qui cause le problème de caractère sur-segmentés.

9. **Caractères non-clair ou ambigu:**

L'écriture cursive, cause parfois des caractères possédant une représentation graphique non-claire et par conséquent une ambiguïté dans l'interprétation du caractère (comme les caractères ت et ر ou ف et م). Et parfois l'écriture de certain caractères comme le caractère "س" par des dents invisibles ce qui modifié complètement la forme générale reconnue du caractère .

10. **Des signes diacritiques additionnels écrits de même forme d'un caractère :**

Parfois, le calligraphe utilise des symboles supplémentaires pour décorer l'écriture et lui donner une vision esthétique: parmi ces symboles, des symboles ont la même forme de lettres, ce qui donne une ambiguïté et entrave le processus de reconnaissance des lettres du texte et rend difficile leur distinction entre eux.

3. La méthode de segmentation de l'écriture arabe artistique proposée

Le travail présenté dans ce mémoire a pour objectif principal de proposer une méthode structurelle adaptable pour la segmentation de l'écriture manuscrite arabe artistique, Alors que toutes les techniques et méthodes de segmentation existantes et développées dans ce domaine sont spécialisés au traitement de l'écriture arabe imprimé ou manuscrite mais inadaptable et ne sont pas capable de segmenter et extraire les principaux segments d'image contenant un texte manuscrite artistique.

Le problème majeur est que l'écriture manuscrite artistique présente une grande difficulté et niveau de complexité élevé comme nous le montre précédemment (des chevauchements et des interdépendances entre les mots et les lettres et à l'abondance de Les signes diacritiques, pas d'alignement claire et d'espacement irrégulière entre les PAW ... etc)

La méthode de segmentation proposée est une méthode structurelle multi-scans (N-Sans) à deux phases capable de segmenter un texte arabe artistique et d'extraire les différents segments en appliquant de nouvelles règles de segmentation en des étapes.

Nous nous sommes appuyés dans notre méthode proposé sur l'application d'une succession de scans (N-Scan) afin d'identifier les différents points d'intérêts dans le texte présenté dans l'image analysé, ces points sont utilisées pour identifier tous les changements dans la Courbe d'écriture affichée et extraire les différents caractéristiques structurelle qui décrivent la forme ou le style de l'écriture du texte artistique en terme de sa topologie et sa géométrie en donnant ces propriétés globales et locales et comment écrire des caractères et les relier les uns aux autres dans le texte analysé et de manière intelligente.

3.1 L'organigramme de la méthode proposée:

Dans ce partie nous allons présenter les différents étapes de notre méthode de segmentation proposé .

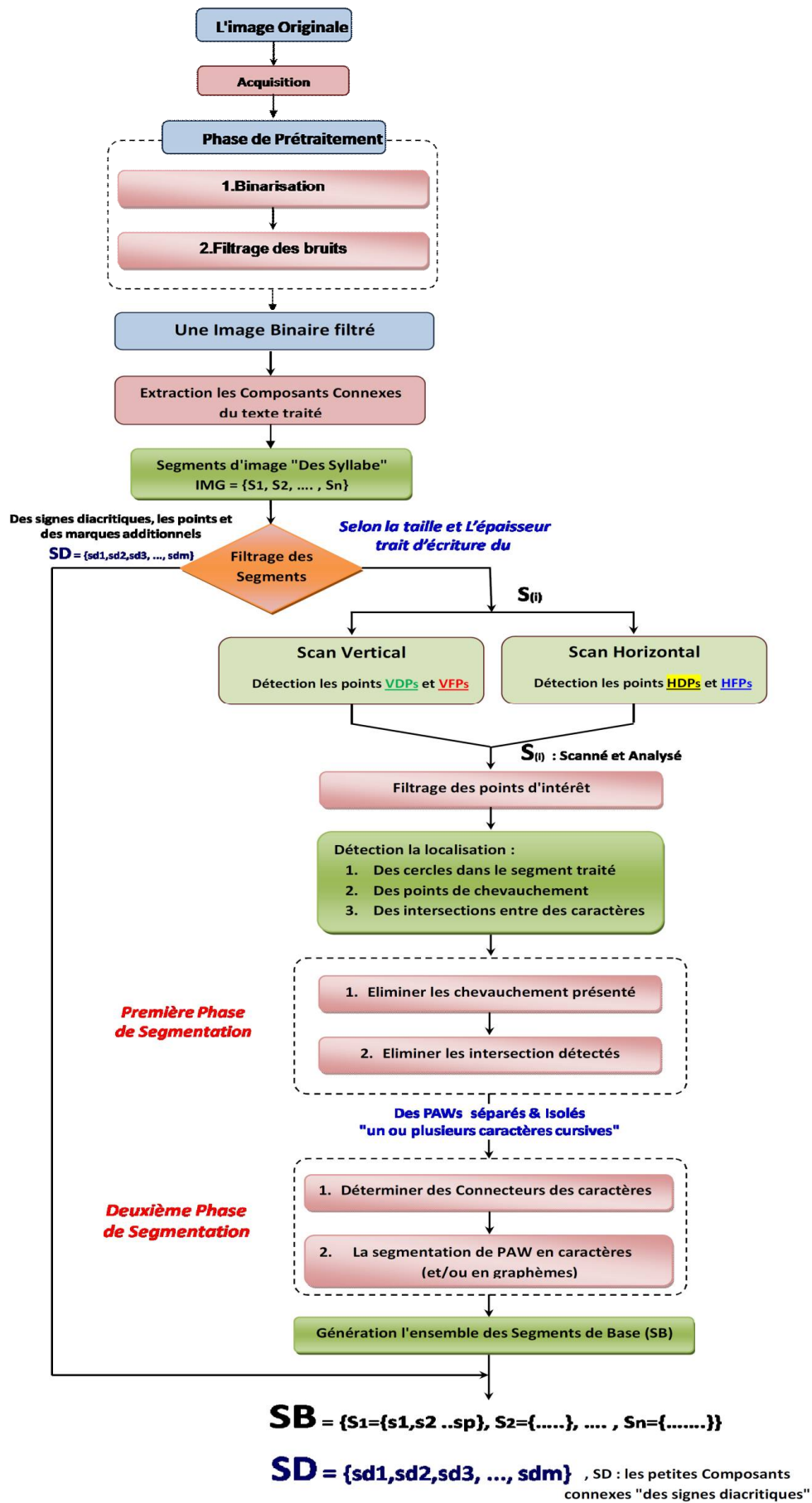


Figure 3.4: Organigramme illustre Le principe générale de la méthode proposée

3.2 Les étapes et le principe de la méthode proposée:

3.2.1 Etape d'Acquisition

L'acquisition est l'étape d'entrée du système proposé fait par une unité d'acquisition: Scanner, Camera.. etc (ou à partir d'une base de données "image numérique") , c'est une étape simple et très importante car elle influence sur les étapes suivantes . Elle consiste à charger l'image du papier et la conversion du document papier sous la forme d'une image numérique en un format d'image connu (tel que : jpeg, gif, bmp. . . etc.) en couleur ou en niveau de gris ou la transformer en un matrice de pixels avec un minimum de dégradations possibles.

3.2.2 Etape de Prétraitements :

L'étape de prétraitement est une étape très important car elle consiste à améliorer la qualité de l'information acquit et préparer les données issues du capteur à la phase de segmentation . Il s'agit essentiellement de réduire le bruit superposé aux données dû aux conditions d'acquisition ou encore à la qualité du document d'origine et essayer de ne garder que l'information significative de la forme représentée. Nous proposons que cette étape passe par deux étapes principales, à savoir:

3.1.2.1 Etape de binarisation :

[L'étape de binarisation est de transformer une image en niveau de gris en une image en noir et blanc, donc de séparer l'information utile du fond de l'image à l'aide d'une méthode de seuillage] [18], Nous utilisons le seuillage global identique pour toute l'image. Où chaque pixel de l'image est comparé à ce seuil et prend la valeur blanc ou noir selon qu'il est supérieur ou inférieur. Cette méthode est une méthode rapide. Cette transformation utilise les formules de conversion RBG/YUV, comme suit :

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

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

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

3.1.2.2 Etape filtrage et suppression des bruits:

Il consiste à utiliser des masques de pixels afin d'éliminer les taches de l'image binaire pour représenter en mieux les points d'intérêt. Le système élimine chaque pixel noir entouré par des pixels blancs ,les pixels isolés (nettoyage). Telle que un point noir (1) est mis en blancs (0) s'il n'y a pas assez de points noirs autour de lui .

3.2.3 Méthode de segmentation proposée:

Le principe de fonctionnement de la méthode proposée est de suivre plusieurs étapes : (comme illustre *la figure 3.4*) :

1. Etape 01 : Extraction Les différentes composantes connexes :

Pour extraire les différents composants du texte artistique traité, Le système proposé commence par un balayage de l'image binaire de droite à gauche afin d'avoir le premier point noir (pixel noir), ensuite il utilise les huit points voisins et une fonction récursive pour traquer le reste du PAW. Le résultat de cette étape est la génération et l'extraction des différentes composantes connexe constituant le texte artistique traité sous formes d'un ensemble des segments (*voire la figure 3.5*).

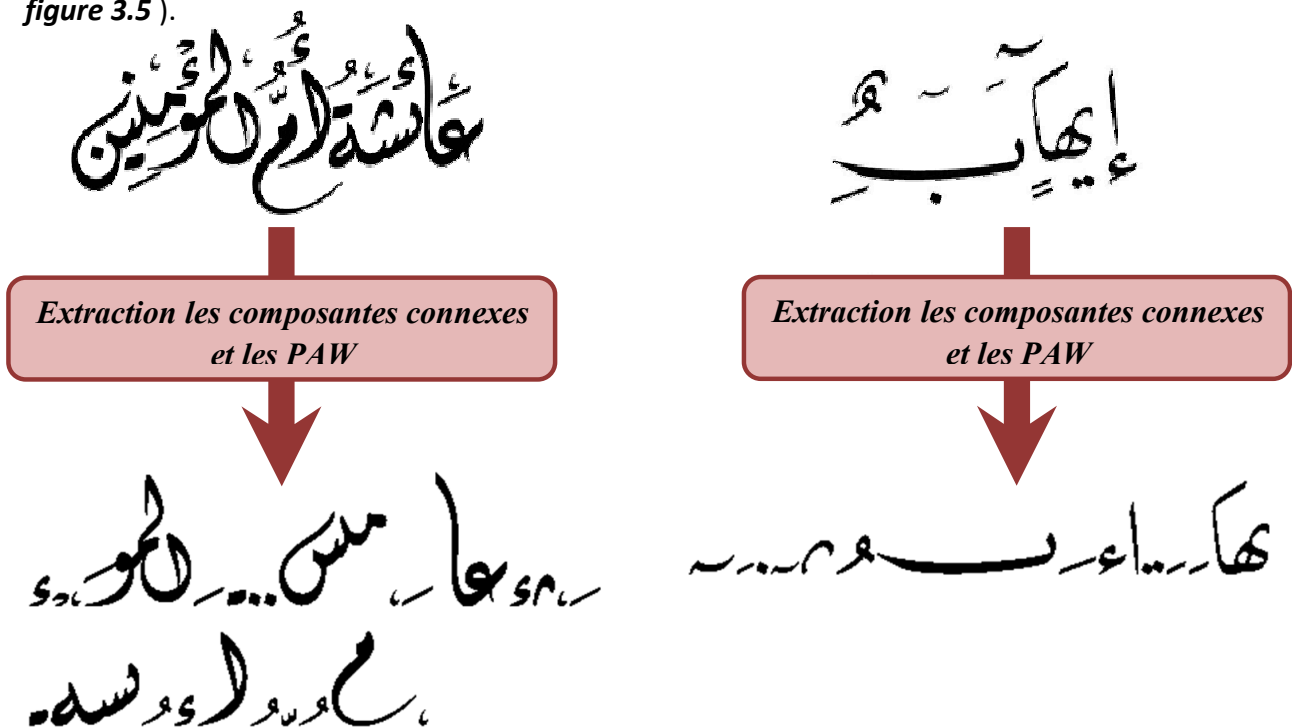


Figure 3.5 : exemple d'extraire Les différents composantes connexes d'un texte artistique

Les composants connexes peuvent être des PAWs chevauchés et/ou interconnecté par l'intersection des extensions de caractères qui les compose, comme et peuvent être des composants isolés "des PAWs, des signes diacritiques « El Chakle » et les points diacritiques et des marques additionnelles".

1.1 La distinction entre les composantes connexes et les signes diacritiques extraites:

Généralement, il y a deux classes des composants qui constituent le texte artistique :

- **La première classe** : les composants connexes qui sont des unités regroupant une PAW ou plusieurs PAW chevauchés ou croix entre eux par des extensions de certain caractère qui les

compose. un PAW isolé peut être un caractère isolé ou une unité connexe regroupant une ou plusieurs lettres sous forme d'une séquence (voir le tableau 3.1).

- **La Deuxième classe** : les petites composantes isolées qui sont les points et les signes diacritiques et les marques additionnelles. (voir le tableau 3.1).

On peut différencier entre les composantes connexes (les différents Syllabes du texte) et les signes diacritiques ("El_Chakle, les points et les marques additionnels) par la taille et l'épaisseur d'écriture de chacune, alors il est possible d'utiliser ces informations comme une règle.

1.2 Illustration de l'extraction des différentes composantes connexes du texte artistique :

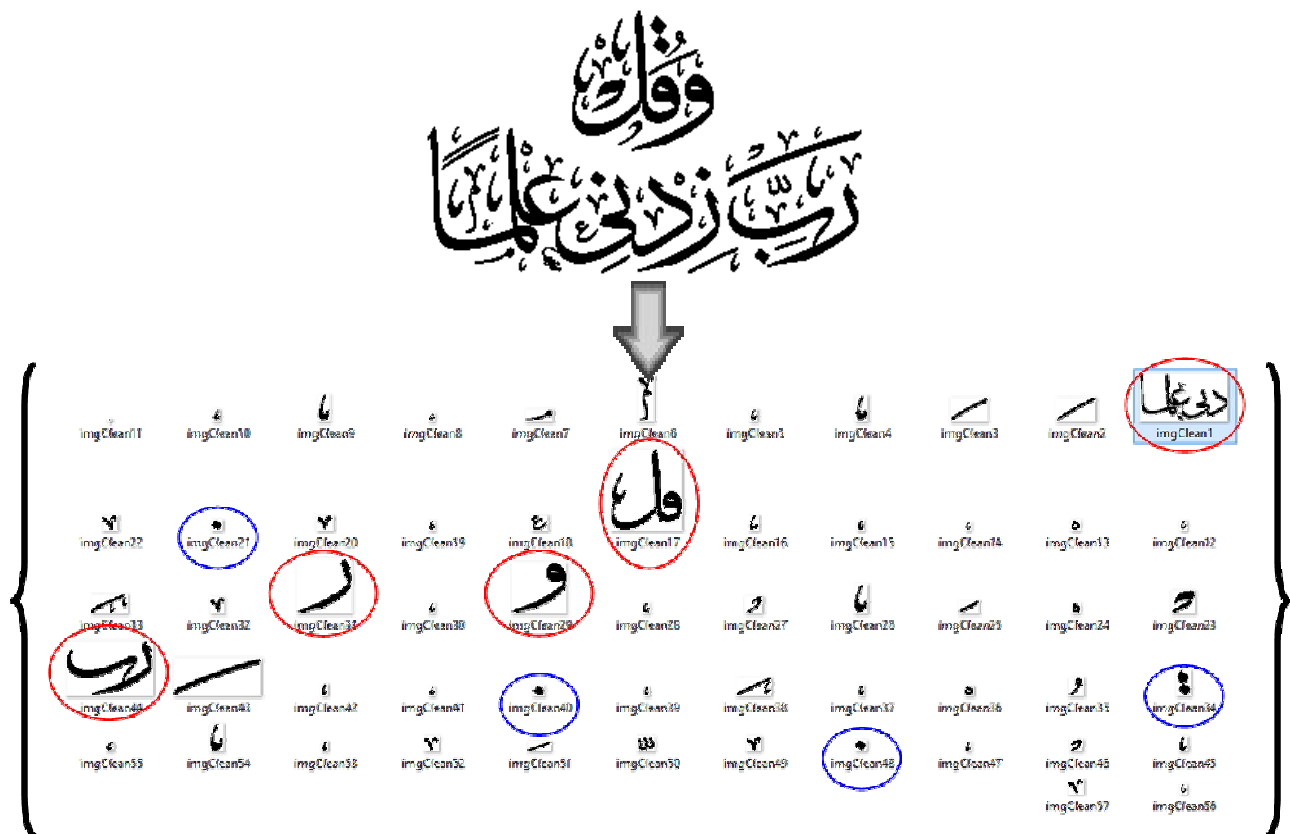


Figure 3.6: Illustration de l'extraction des différentes composantes d'un texte artistique

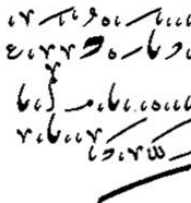
Les signes diacritiques additionnel		les composantes connexes du texte		
« El_Chakle » & les signes additionnels	Les points diacritiques	Les PAWs Isolés		Les PAWs Chevauchés entre eux
	◌َ	و	ك	دلى
	◌ِ			
	◌ُ	ر	ك	ر
	◌ْ			

Tableau 3.1 : La distinction entre les composantes connexes du texte traité et les signes diacritiques

2. Etape 02 : La Détermination différents points d'intérêts "les PNVs et PNHs"

Pour chaque segment $S(i)$ extraite dans la première étape, on déterminer et identifier les différents points d'intérêts dans le composant connexe qu'il contient en utilisant parallèlement un scan vertical et un scan horizontal, en utilisant des masques prédéfinies, Les points d'intérêt détectés sont des points de nature verticale (PNVs) nommé points d'inclinaison qui permet de détecter les liaisons ou connecteurs des caractères et des points de nature horizontale (PNHs) nommé des points de structure permettant d'identifier les limites du caractère et l'apparence de leur forme .

Pour déterminer les Points de Nature Verticale " PNVs ", le système proposé utilise deux masques d'analyse (*présentés dans la figure 3.7*) pour déterminer les deux types de points verticaux: Points de Fusion Verticaux (VFPs) en couleur rouge et Points de Division Verticaux (VDPs) en couleur vert.

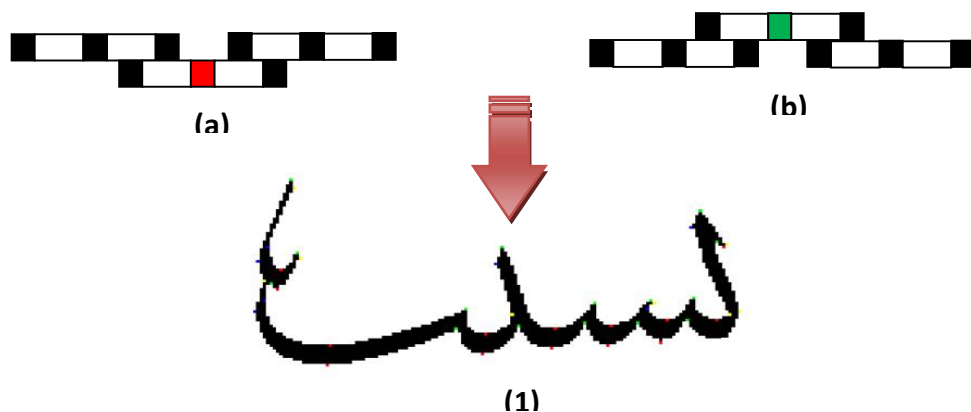


Figure 3.7: Les Masques de balayage vertical ((a) masque de fusion (b) masque de division)

Pour déterminer *les Points de Nature Horizontale " PNHs "*, De la même façon, Le système proposé utilisant des masques de balayage horizontal (*comme le montre la figure 3.8*) pour déterminer les deux types de points horizontaux: *Points de Fusion Horizontaux (HFPs)* en couleur bleu et *Points de Division Horizontaux (HDPs)* en couleur jaune.

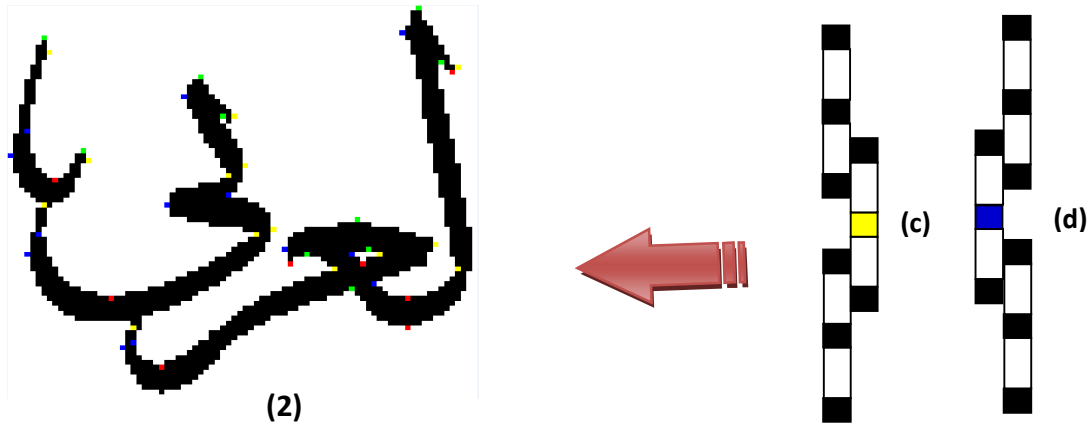


Figure 3.8: Les Masques de balayage Horizontal ((c) masque de fusion (d) masque de division)

3. Etape 03 : Filtrage et élimination des points incorrects :

Les points générés peuvent contenir des points incorrects (ou additionnels), ce qui nécessite une étape de filtrage pour éliminer ces points additionnels, Le but principale de cette étape est d'éviter l'ambiguïté et la difficulté dans le processus de segmentation.

3.3 Des règles Filtrage des points incorrectes

- a) Elimination l'un des points proche (comme le montre *la figure 3.9*) quand $T \& L < \text{pen_size}$,
telle que (T " thickness ": est l'épaisseur) et L "length" : est la longueur des lignes des segments.

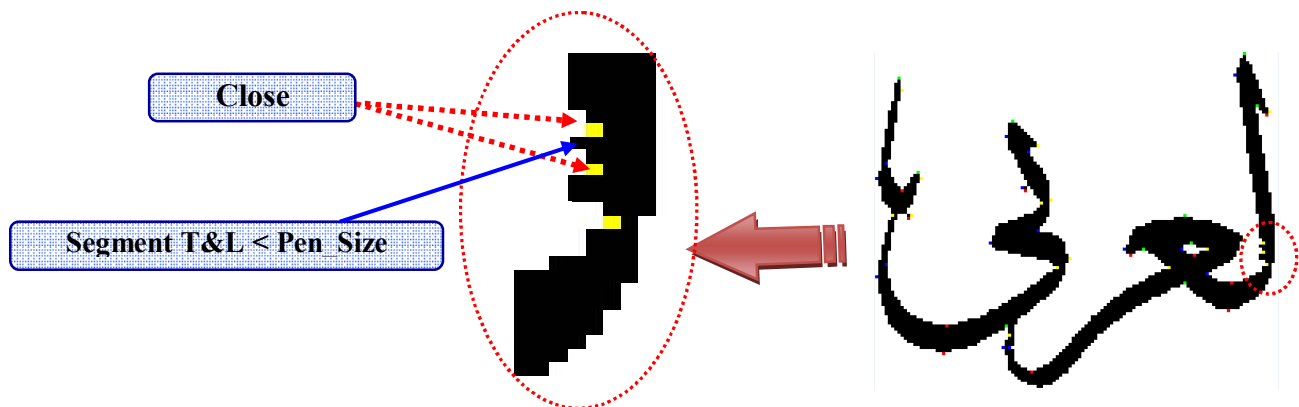


Figure 3.9: illustration de filtrage de points

- b)** Dans le cas où deux points du type VFPs (ou deux points du type VDPs) consécutifs horizontalement sans la présence d'autres points d'autre type parmi eux, nous supprimons l'un d'entre eux comme une opération de filtrage.
- c)** Dans le cas où deux points différents un de type VDP et l'autre de type HDP sont en convergence ou très proches (comme le montre *la figure 3.10*), Nous suggérons de garder le point **HDP** et de convertir le couleur du point VDP en blanc (élimination).

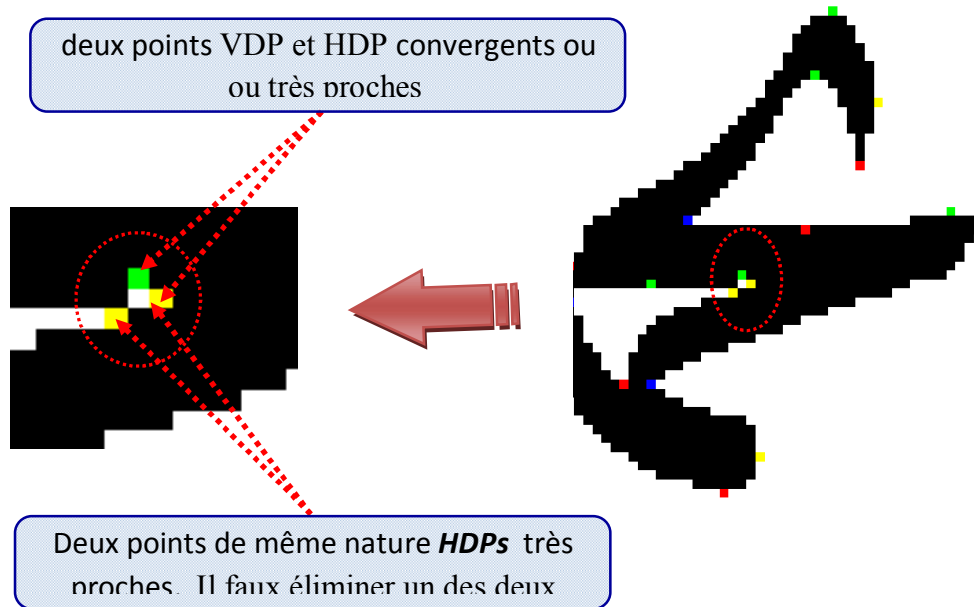


Figure 3.10: illustration de filtrage des points entre deux points différents VDP et HDP

- d)** Dans le cas de la convergence d'un point de type HFP avec un autre point de type VFP (comme le montre *la figure 3.11*), Nous avons suggéré de garder le point de type HFP et de convertir le couleur du point VFP en Blanc, parce que on donnant la priorité à la segmentation horizontale pour l'élimination de la ligature verticale des caractères.
- e)** Dans le cas où nous trouvons un point **HFP** (coloré en bleu) non connecté à aucune pixel noir des côtés gauche et droit, alors nous avons suggéré de changer ce point par un blanc pixel (une segmentation).
- f)** Dans le cas où nous trouvons un point **HFP** (coloré en bleu) non connecté à aucune pixel noir des côtés supérieur et inférieur, alors nous avons suggéré de changer ce point par un blanc pixel (une segmentation). Et le même pour un point de type VFP (coloré en rouge).

RQ : Si le nombre des points VFPs et HFPs filtrés est égale à zéro ("VFPs = 0" et "HFP = 0"), le PAW courant est considéré comme de Structure Simple (SS). Sinon, il cherche à trouver les connecteurs des caractères.

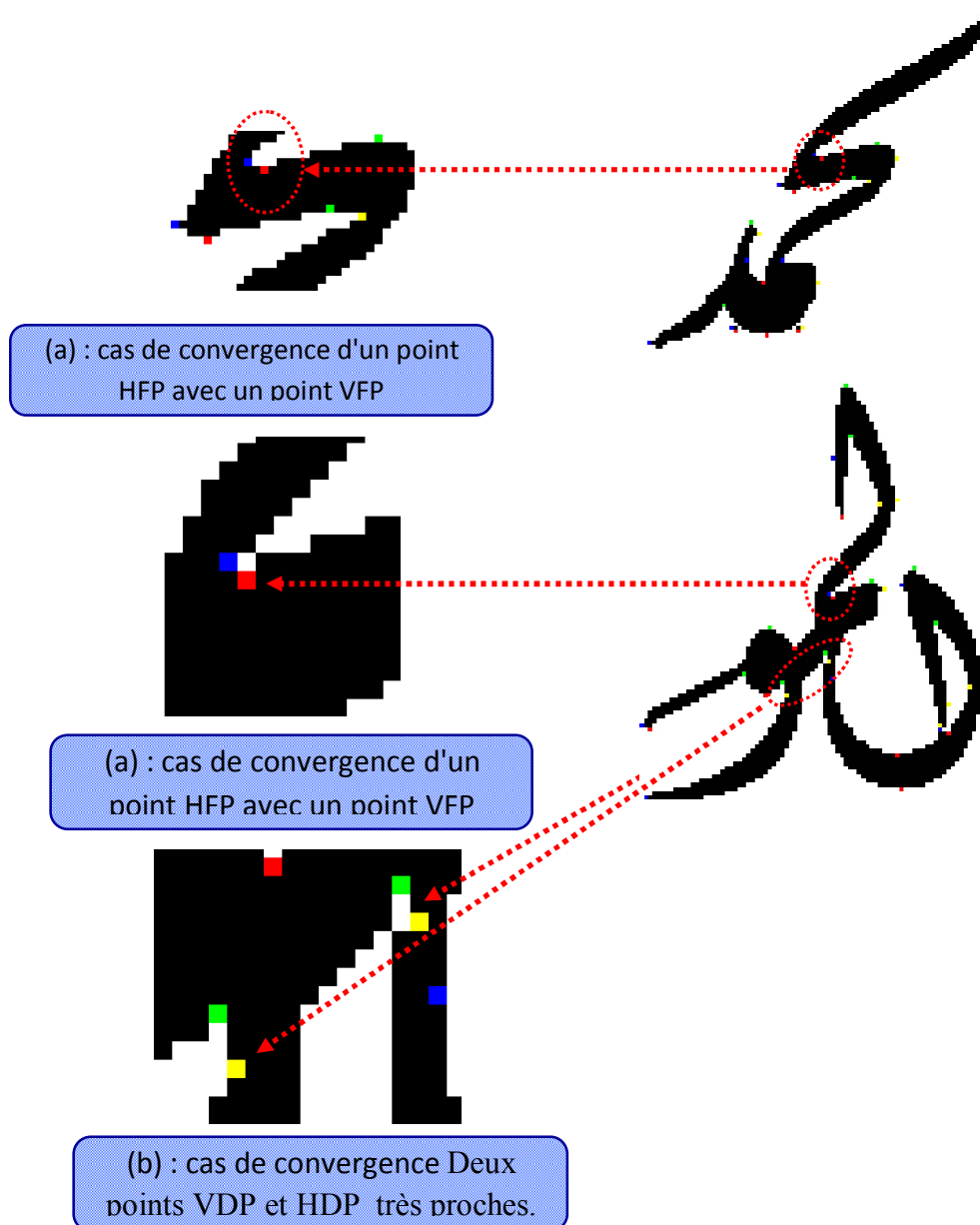


Figure 3.11: illustration des cas de filtrage des points d'intérêt

4. *Etape 04: Découverte des cercles des caractères et des points de chevauchement et d'intersection.*

Cette étape est une étape préalable à la segmentation, telle que pour chaque segment $S(i)$ n'on pas de structure simple et par l'ordre :

- ☒ Identification (s'il existe) de la Localisation des cercles des caractères (comme les caractères (ط, ق, ف, و, خ) qui composent les composants connexes
- ☒ Détermination les points potentiel de chevauchement descendants entre les caractères qui ont des extensions descendantes aux autres caractères appartient de même mot ou au autre pseudo-mot.
- ☒ Détermination la Localisation des intersections entre des caractères qui composent les pseudo-mots.

4.1 *La Détermination la Localisation des cercles :*

Nous avons suggéré la détermination les cercles dans les composants connexes " les syllabe du texte traité " pour éviter de segmenter le caractère elle-même et d'éviter l'ambiguïté dans l'application des règles de segmentation proposés et d'éviter des erreurs pendant la processus de segmentation.

Après vérification, nous avons trouvé que les boucles de caractères sont toutes distingués par la présence de quatre points consécutifs sur son périmètre de nature différente et avec un ordre bien défini et fixe comme suite : un point HDP (de couleur Jaune) dans le coté du cercle droite ensuite un point VFP (de couleur rouge) au bas de la boucle, et un point HFP (de couleur bleu) au coté gauche et un point VDP (de couleur vert) au sommet de la boucle. Comme le montre les figures suivants.

Le type des points constituant la boucle est obtenu au fur et à mesure qu'on fait des opérations de division ou fusion.

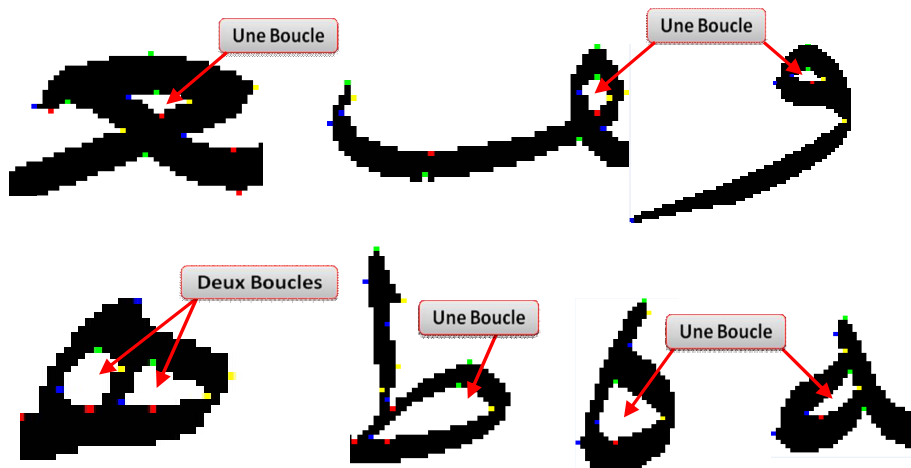


Figure 3.12: illustration la détermination les cercles des caractères en utilisant les points d'intérêt

5. Etape 05: l'étape de segmentation des composants connexes en deux phases:

5.1 Première phase : la segmentation des composants connexes en des PAWs qui les compose, par la Séparation entre les PAW chevauchés entre eux et l'élimination des pixels de chevauchement entre les caractères chevauchés en descendante, et En séparant les caractères interconnectés entre eux par l'élimination les pixels d'interconnexions entre les caractères .

1) Elimination des points de chevauchement descendant présenter dans le texte analysé:

Dans l'écriture artistique, le chevauchement descendant se produit lorsqu'une lettre à extension descendant est attachée à un segment d'une autre lettre du même mot ou d'un autre mot.

Dans chaque chevauchement descendantes entre les caractères, on remarquant qu'il existe toujours un point vert (**VDPs**) ou un point jaune (**HDPs**) avec épaisseur du trait d'écriture nul ou presque nul " épaisseur d'un ou deux pixels" et toujours précédé par un point bleu (**HFPs**) et un point rouge (**VFPs**) par la suite "de manière successive", Comme montre *la figure 3.13* Suivante.

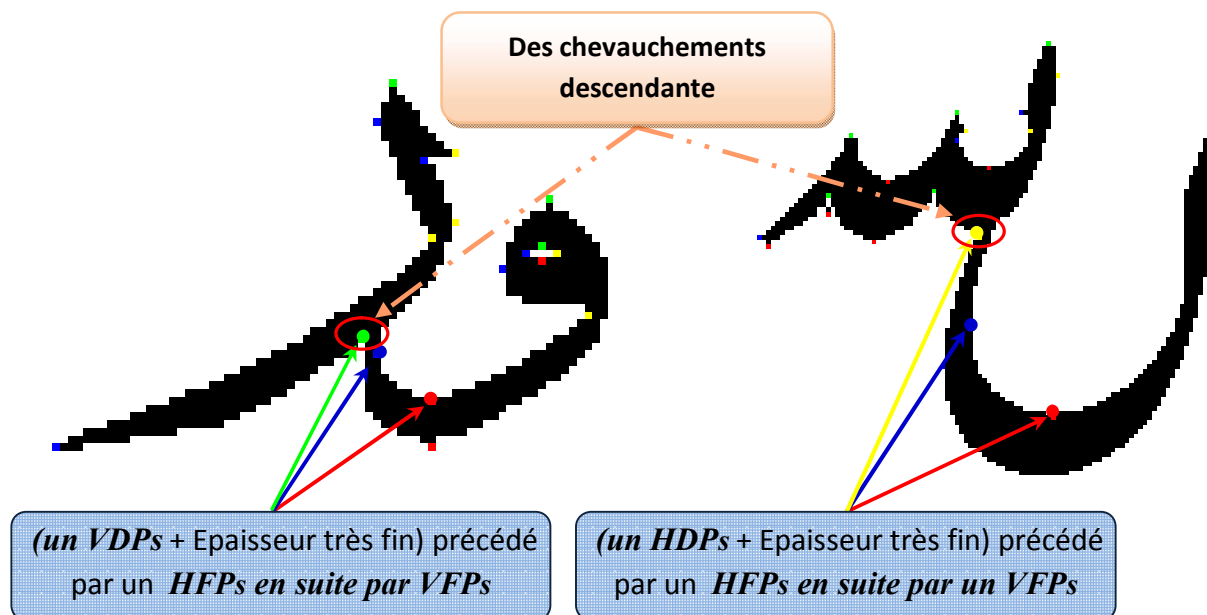


Figure 3.13: illustration comment détecté les points de chevauchement

Pour éliminer le chevauchement descendant et dissocier les caractères chevauché, on éliminant les pixels noirs appartenant au premier caractère adjacents à la deuxième caractère. Le résultat de cette étape est la décomposition d'un composant connexe complexe à des PAW isolé et non chevauchés.

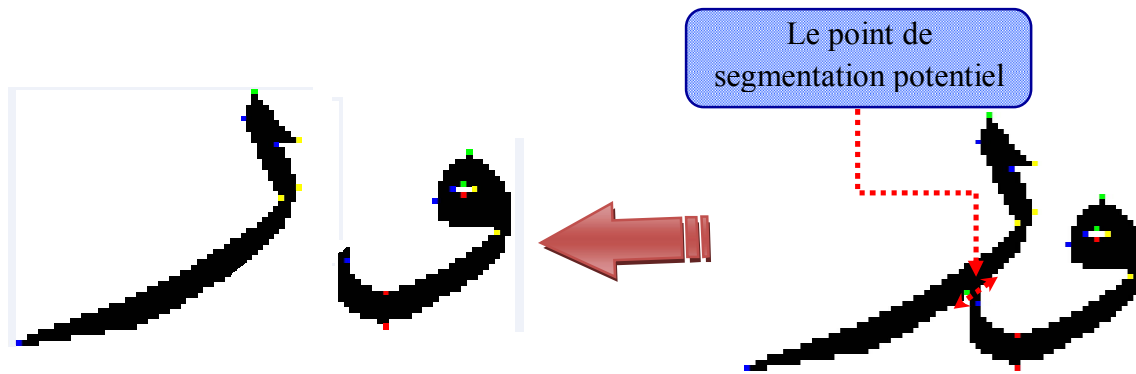


Figure 3.14: le principe d'élimination les points chevauchement descendant

RQ : Un Cas exceptionnel très important, Le point VFPs qui précède un point de chevauchement descendant ne considère pas un point de segmentation, parce que elle n'est pas un connecteur des caractères, elle est un extrémité terminale d'un caractère.

2) L'Elimination et la Séparation des intersections entre les caractères intersectés présentés dans le Segment traité:

Les intersections entre les segments des lettres est parmi les caractéristiques qui caractérise l'écriture artistique, L'intersection résulte toujours lorsque il y à un caractère qui a une extension horizontale où il croise d'autres extensions verticales d'autres caractères.

Nous trouvons deux types de base: l'un sous la forme de (+) quand il y a deux lignes se croisant verticalement et l'autre sous la forme de (x) parce que l'existence de l'une des lignes transversales est oblique.

Pour détecter et localiser les intersections dans un texte artistique, on trouve toujours les quatre points par l'ordre aux coins de l'intersection (le point rouge VFPs dans le coin supérieur du côté droit en suite le point bleu HFPs dans le coin inférieur droit, puis le point vert VDPs dans le coin inférieur gauche et le point jaune HDPs dans le coin supérieur gauche), voire **les figures 3.15 et 3.16**. Et parfois nous trouvons trois points seulement au lieu de quatre, parfois ces cas se retrouvent dans le cas d'intersections de type x car certains des angles sont très ouverts. Comme le montre **la figure 3.17**.

C'est pourquoi nous allons Utilisez une technique supplémentaire "**la méthode de squelettisation**" dans ces cas pour vérifier la présence vraiment une intersection ou non et évitez autant que possible les erreurs dans le processus de segmentation.

Dans la segmentation, Nous choisissons la segmentation verticalement dans le segment du caractère qui s'étend horizontalement Du point VFPs (colorer en rouge) au point HFPs (colorer en bleu) et Du point HDPs (colorer en jaune) au point VDPs (colorer en Vert). Comme le montre *la figure 3.16*.

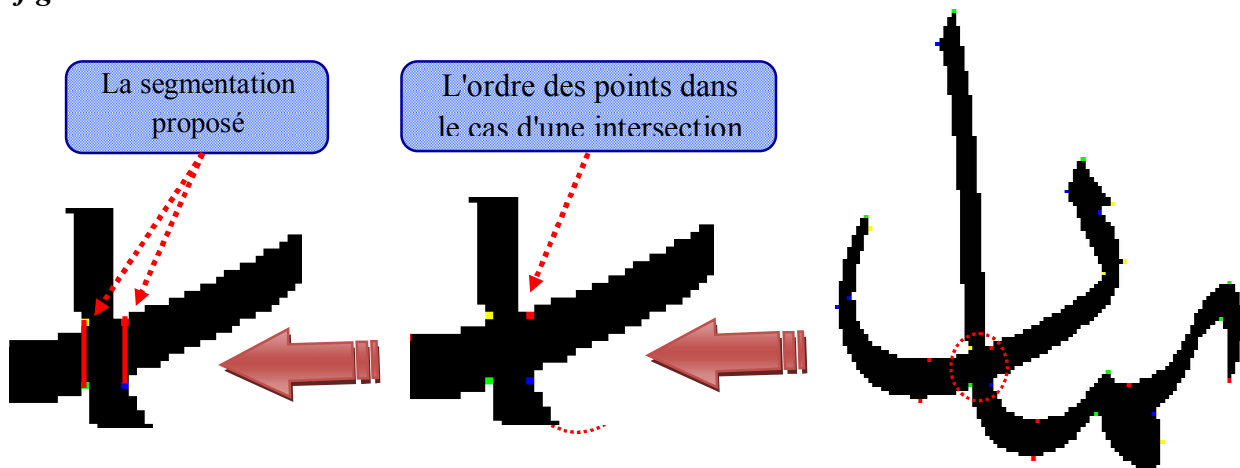


Figure 3.15: la détection des intersections des caractères en utilisant les points d'intérêt

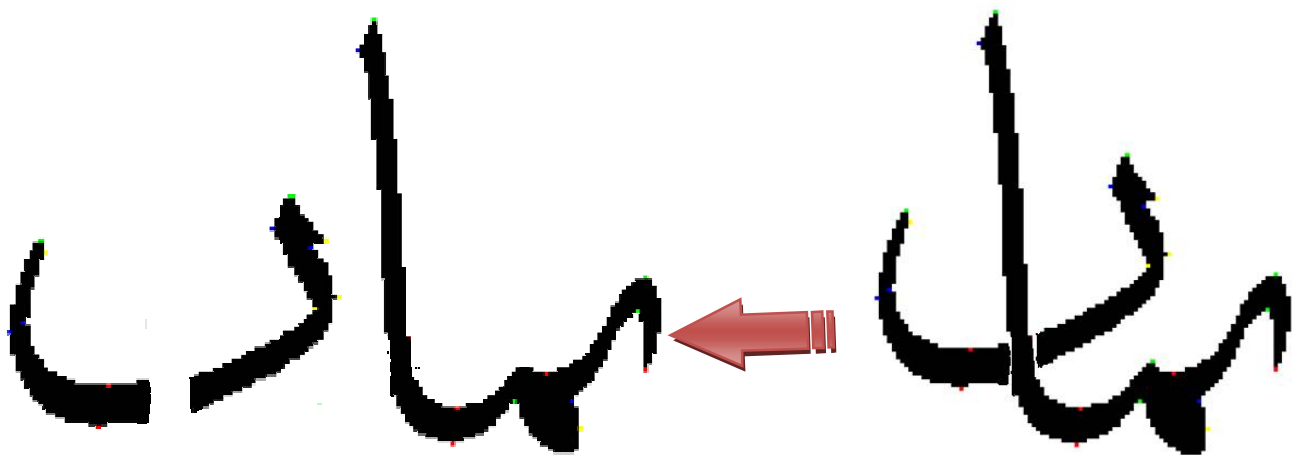


Figure 3.16: le principe d'élimination les intersections entre les caractères

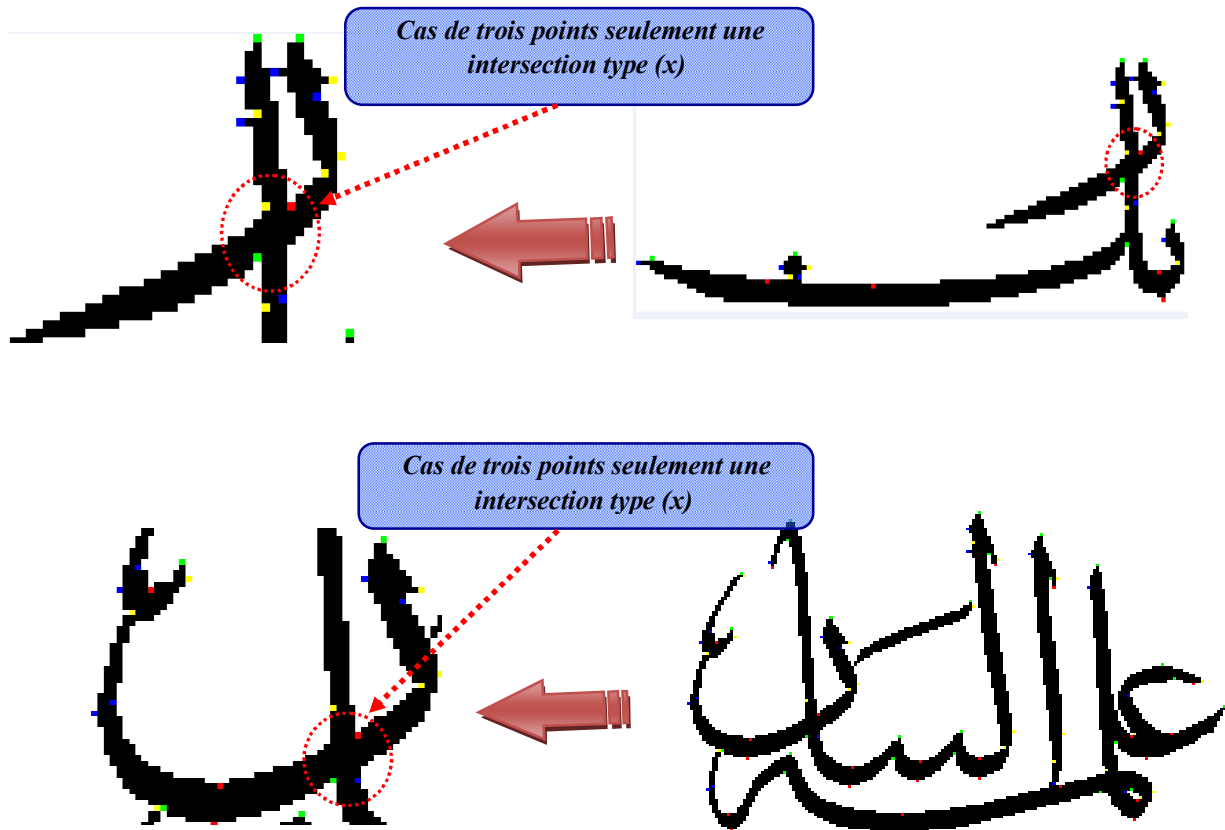


Figure 3.17: des intersections entre des caractères à trois points d'intérêt seulement

5.2 Deuxième phase : la détermination des Connecteurs de caractères (CCs) et la segmentation des PAWs en caractères (ou des graphèmes):

C'est l'étape finale dans le processus de segmentation proposé, pour trouver les connecteurs des caractères nous appliquons les règles suivantes:

1. Premièrement, il faut garantir que les Points VFPs proposé comme des points potentiel de segmentation à
2. le premier contrainte, Il faut souligner que les points VFPs (Les points rouge) potentiels comme des points de segmentation doivent être en place là où l'épaisseur de trait est supérieure ou égale à Pen size (**épaisseur de trait $\geq$ Pen size**)
3. le deuxième contrainte, les Points VFPs et HFPs potentiels comme des points de segmentation doivent être hors des cercles.
4. **La règle 01 :** les Points VFPs qui sont toujours précédés de points de type HDPs (ou VDPs) et vérifient que les deux conditions précédentes sont des connecteurs de caractères, l'exception de cette règle sera Uniquement dans un cas où le point VFP suivi

d'un point de type HFPs et ensuite d'un point de type HDP ou VDP " avec un épaisseur trait fin" .

5. **La règle 01** : dans le cas d'un ligature entre des caractères, les points HFPs qui sont toujours précédés de points de type VDPs ou HDPs et non précédé par un connecteur VFP sont considérés des connecteurs de caractères (voir *la figure 3.18*).

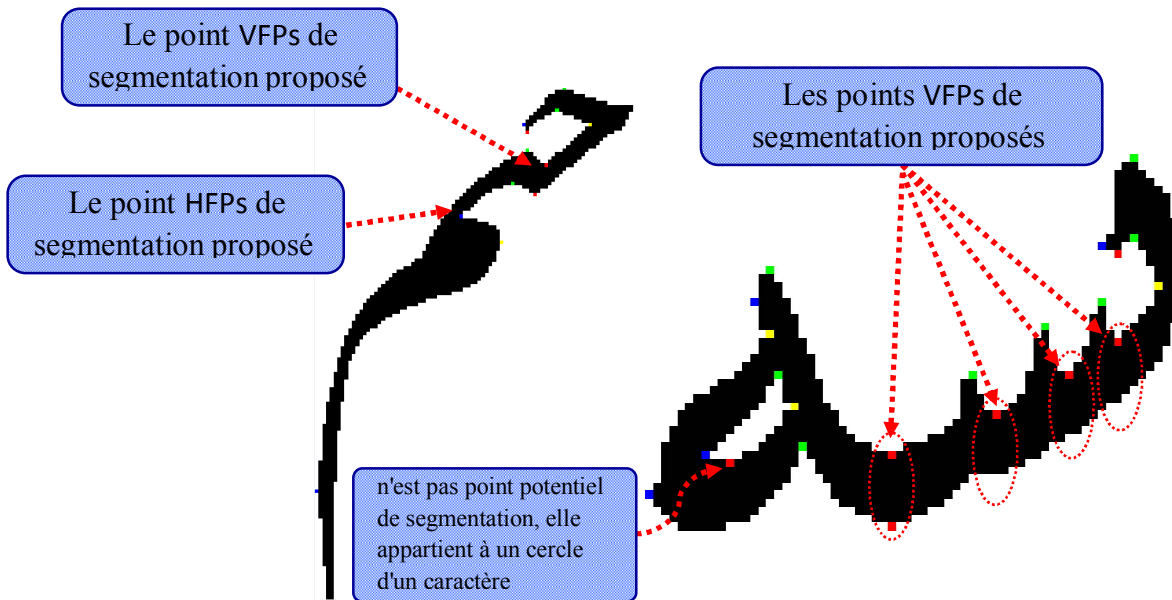


Figure 3.18: Extraction des connecteurs des caractères par les Points VFPs et HFPs

4 Conclusion

Dans ce chapitre, nous avons commencé par présenter les caractéristiques et la problématique posée par L'écriture arabe artistique.

Un texte arabe artistique caractérise par la présence une grande difficulté dans leur nature, dans la méthode d'écriture, (pas d'espacement régulière entre les mots, PAWs et pas de possibilité de contrôler les lignes d'écriture et l'existence des chevauchements et des ligatures entre les caractères ...) et en l'absence de critères de référence pour la reconnaissance de l'écriture artistique, ce qui rendre le développement d'une méthode de segmentation est très difficile à réaliser

Ensuite, nous illustrons l'architecture globale du système proposé et présentons en détail notre contribution par l'amélioration dans une méthode des segmentations structurale multi-scans ou N-Scans, en proposons des nouvelles règles à suivre et fait la segmentation à deux phases pour être capable et adaptable à segmenter l'écriture arabe manuscrite artistique. Cette méthode permet de préparer le premier ensemble des segments représentant par la suite l'état initial du puzzle et une base pour autres systèmes de reconnaissance de l'écriture arabe.

Chapitre 04 : Mise en oeuvre notre contribution, résultats et évaluation

1. Introduction

Dans le chapitre 02, nous avons présenté les différentes méthodes de Segmentation utilisées dans ce domaine , Et prouvé qu'il n'est pas spécialisé dans le traitement et la division de l'écriture artistique en raison de la nature , de la complexité et de la difficulté de cette écriture , et ainsi dans le chapitre précédent nous avons présenté notre contribution de proposer et créer une méthode de segmentation structurelle N-Scans à deux phases pour la segmentation du texte arabe artistique et leur principe de fonctionnement en détail et ceci pour atteindre les résultats souhaités

Dans ce chapitre nous allons présenter les détails de notre système proposée de segmentation , ainsi que la validation des résultats .Nous allons commencer par montrer l'architecture globale du système proposé, où nous allons voir des exemples sur l'application des différentes étapes de la méthode de segmentation structurelle N-Scans à deux phases jusqu'à la génération les différents segments "des Imagettes" constituant le texte artistique traité, Lesquels sont ensuite utilisés dans d'autres systèmes comme ensemble des segments initiale pour la reconnaissance du texte artistique et d'autres traitements de l'écriture artistique. Enfin, nous allons faire la validation des résultats et présenter un bilan des résultats de segmentation de notre système .

2. Notre Système proposé:

Le système commence par l'acquisition de l'image contenant le texte source par l'intermédiaire d'une unité d'acquisition telle que : scanner, camera ou à partir d'une base de données (comme illustre **la figure 4.2**). Ensuite, il passe par une phase de pré-traitement pour améliorer la qualité de l'image acquise ou importé (Binarisation, Filtrage des bruits, ..) et extrait une image binarisé net sans bruit pour la préparée à la phase de segmentation. Ensuite La phase de segmentation à deux phases consiste à extraire les principaux segments du texte artistique traité nécessaire pour des traitements ultérieurs.

3. *L'Organigramme de l'architecture globale de notre système proposé:*

Dans ce partie nous allons présenter l'architecture globale et les différents étapes de notre système de segmentation développé . Nous présenterons par des exemples l'application de chaque étape du système en appliquant la méthode de segmentation structurale N-Scans à deux phases.

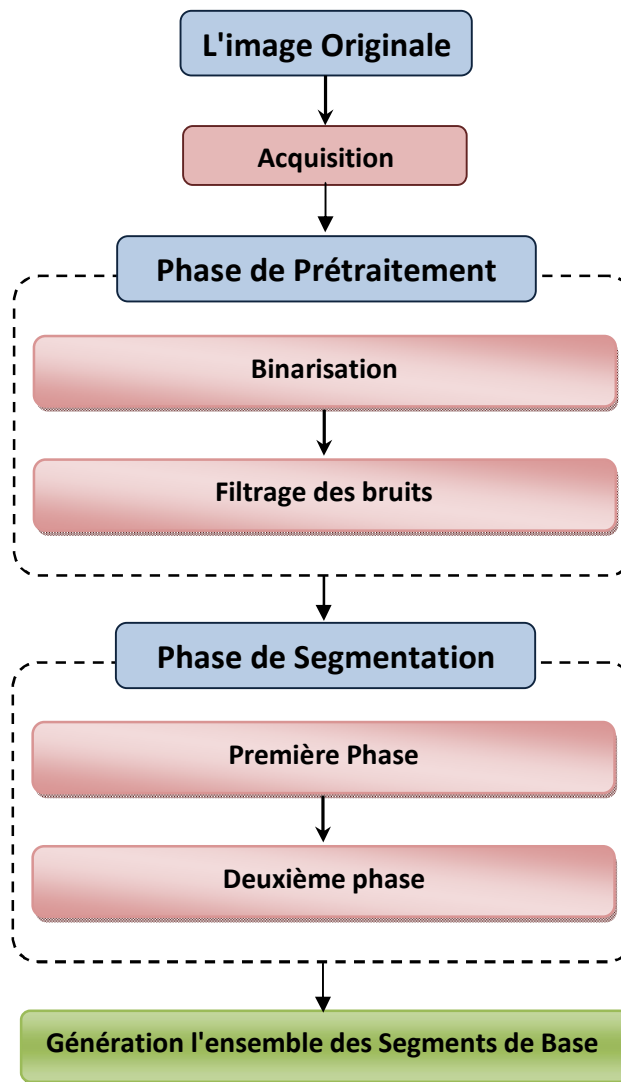


Figure 4.1: *L'Organigramme de l'architecture globale de notre système proposé*

3.1 Etape d'Acquisition

l'acquisition de l'image c'est l'entrée de ce système se fait par une unité d'acquisition: Scanner, Camera.. etc (ou à partir d'une base de données "image numérique"). Elle consiste à charger l'image du papier et la conversion du document papier sous la forme d'une image numérique en un format d'image connu (tel que : jpeg, gif, bmp. . . etc.) .



Figure 4.2: Illustration de l'étape d'acquisition

3.2 Etape de Prétraitements :

Le prétraitement est une étape très important visant à améliorer la qualité de l'information et préparer les données issues du capteur à la phase de segmentation .

3.2.1 Etape de binarisation :

[L'étape de binarisation est de transformer une image en niveau de gris en une image en noir et blanc, nous utilisons le seuillage global où chaque pixel de l'image est comparé à ce seuil et prend la valeur blanc ou noir selon qu'il est supérieur ou inférieur. Cette transformation utilise les formules de conversion RGB/YUV, comme suit :

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

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

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

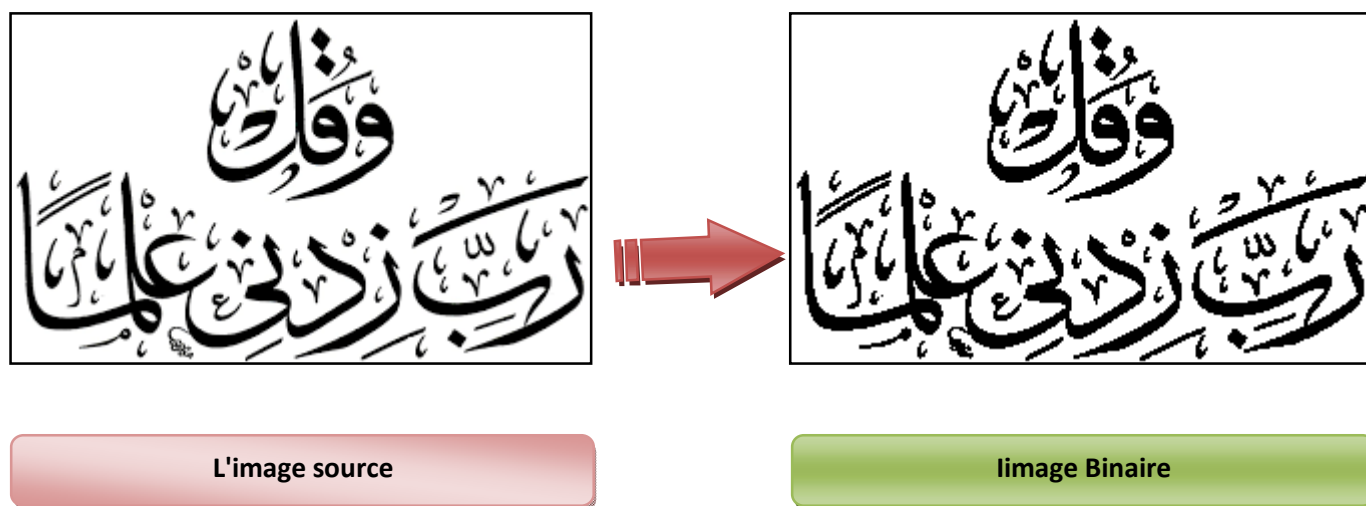


Figure 4.3: illustration de l'étape de binarisation de l'image acquit

3.2.2 Etape filtrage et suppression des bruits:

Il consiste à utiliser des masques de pixels afin d'éliminer les taches de l'image binaire pour représenter en mieux les points d'intérêt. Le système élimine chaque pixel noir entouré par des pixels blancs, les pixels isolés (nettoyage). Telle que un point noir (1) est mis en blancs (0) s'il n'y a pas assez de points noirs autour de lui.



Figure 4.4: illustration de l'étape de binarisation de l'image acquit

3.3 La Segmentation par l'application la méthode de la segmentation proposée N-Scans à Deux Phases :

Selon notre méthode de segmentation proposée , le système avant tous, fait un balayage et une analyse de droite à gauche sur l'image binarisé , nous utilisons les huit points voisins du pixel et une fonction récursive créée pour traquer le reste du PAW . Le résultat de cette étape est la génération et l'extraction des différentes composantes connexe le texte artistique traité sous formes des Imagettes constituant sous formes d'un ensemble des segments (la figure 4.4 comme exemple) , les segments générés dans le premier pas sont stockés dans un fichier nommé " ImageClean " Ensuite , Il sépare ces segments en deux ensembles par "la taille du segment". le premier ensemble sont les grands Composants connexes qui peut être des PAW ou des PAWs chevauchés . La deuxième ensemble sont les signes diacritique et les petites segments et les transférés et stockées dans un autre fichier nommé " NegligeableImages ".

La deuxième étape, après la séparation entre les grands composants connexes et les signes diacritique et les marques additionnels , le système fait un balayage horizontal et vertical pour identifier et déterminer les points d'intérêt pour chaque Imagette sélectionnée dans la première ensemble (de taille grand relativement).

si les points sont identifiés , Ensuite pour chaque Imagette contenant un composant connexe le système fait un filtrage partiel des points pour éviter l'ambiguïté dans l'exécution des règles de segmentation parce que si les points non filtrés le système tombe dans des cas de défaillance et dans des perturbations dans l'exécution des règles de segmentation entre les caractères.

Ensuite, Après avoir le filtrage des points d'intérêt , le système applique les règles créées pour l'annulation des chevauchements descendants ensuite appliquer les règles d'élimination les intersection entre les caractères,

Ensuite, Après avoir le filtrage des points d'intérêt , le système applique les règles créées pour l'annulation des chevauchements descendants ensuite appliquer les règles d'élimination les intersection entre les caractères, cela permet la dissension des composants connexes en des PAW isolés (pas de chevauchement et des intersection entre les caractères) .

ensuite le système exécute les règles de détermination des connecteurs entre les lettres dans chaque PAW obtenue et extrait les différents segments de PAW en un dossier isolé nommé par même nom du PAW traité.

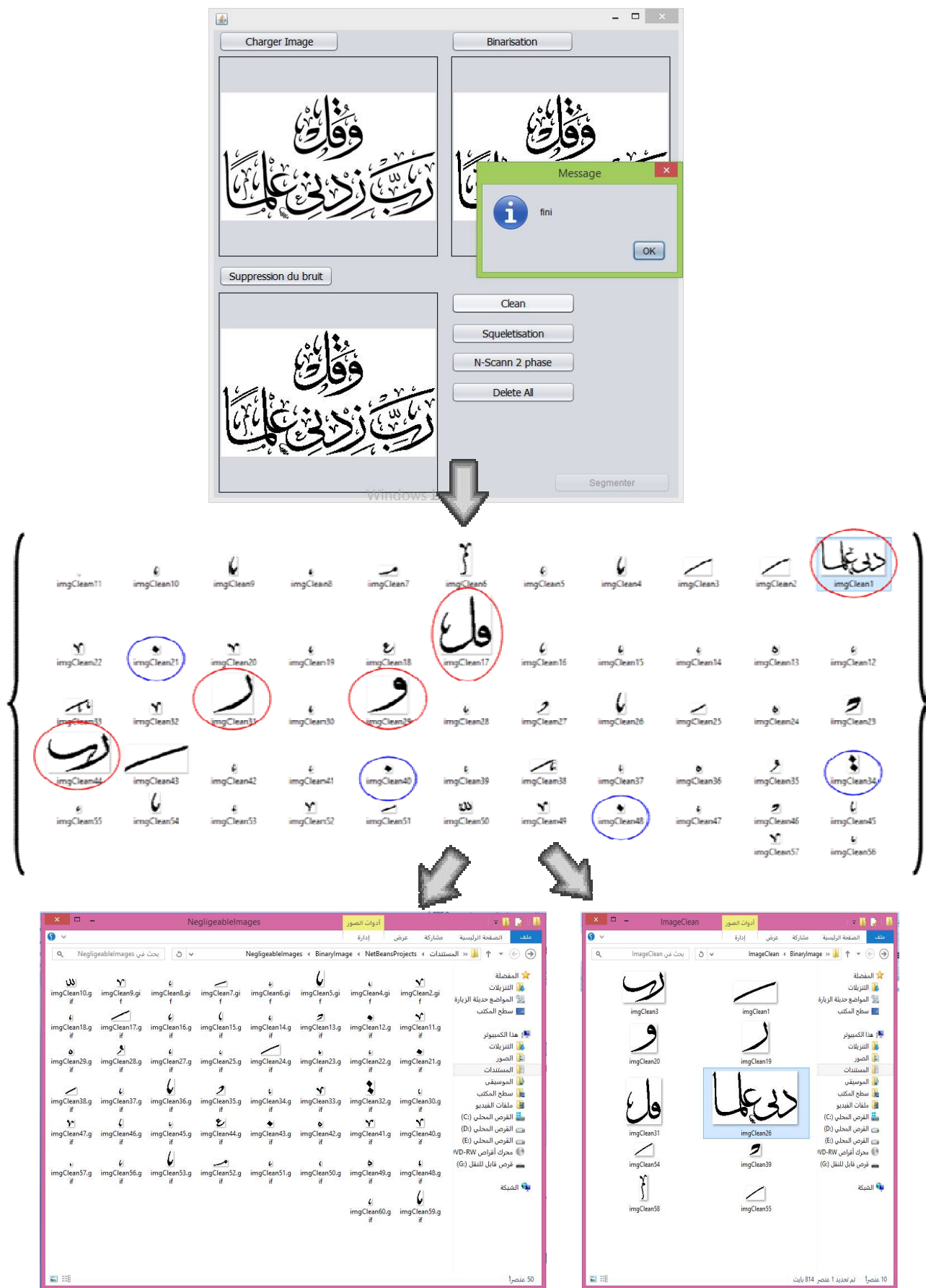


Figure 4.5: Illustration comment séparer les signes diacritiques à les Composants connexes

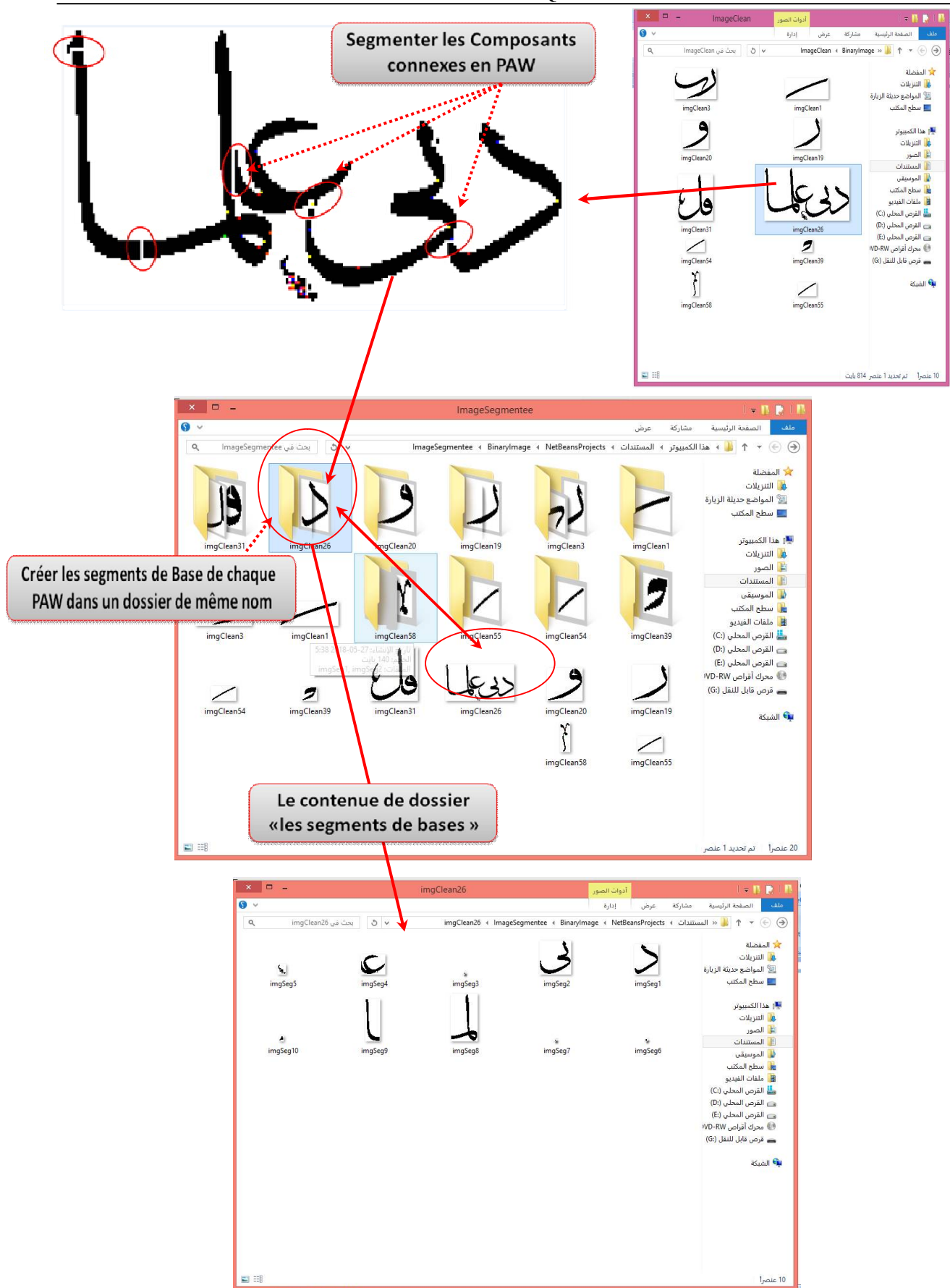
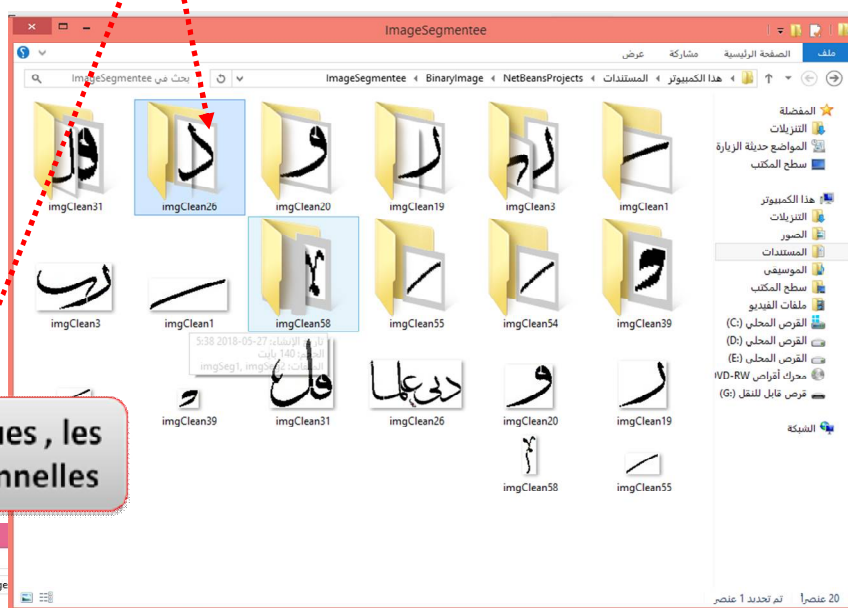


Figure 4.6: Illustration comment la méthode extraire les segments de base d'un composants connexe



Les Grands Composants connexes
segmentés en des Graphèmes
«des segments de base»



Les différents Signes diacritiques , les
points et les Marques Additionnelles



Figure 4.7: illustration comment le système génère les composants de base après la segmentation

4. Phase des tests & évaluation des résultats:

Après la présentation de l'architecture de notre système ainsi que les résultats obtenues dans chaque étape jusqu'à la fin, on obtenant l'ensemble des segments de base nécessaire du texte artistique traité. les résultats obtenues de cette méthode sont tout à fait satisfaisants; cependant, telle que nous appliquons notre méthode proposé sur **50** images (comme un base de teste) contenant des textes arabe artistique (écritent avec un style calligraphique) de différents niveaux de complexité , le taux de succès estimé de cette technique de segmentation Évalué à **92 %** des images ont été segmenté avec succès malgré les difficultés et les problèmes liés à la nature et au style d'écriture. Et **8 %** de défaillances dans les résultats résultant de l'existence des complications élevé dans la méthode d'écriture et des difficultés illimité dans la distinction entre les lettres et les mots.

Les 50 images testées sont divisées en quatre trois images une contenant des textes artistique simple , la deuxième contient des textes plus compliqués du premier et le troisième contenant des textes artistique très compliqués contenant des chevauchements et des intersections plus compliqués, les résultats obtenues pour chaque catégorie des texte artistiques selon leur niveau de complexité sont :

Les catégories	Niveau de complexité	Nbr des images	Les images segmentés avec succès	Les images segmentés avec des défaillances	Le Taux de succès du méthode proposée
Ensemble d'images 01	Textes Simples	10	10	0	100.00%
Ensemble d'images 02	Textes compliqués	33	30	3	90.91%
Ensemble d'images 03	Textes Très compliqués	7	6	1	85.71%
Nbr globale des images testés		50	46	4	92.21%

Tableau 4.1: Les taux de succès de notre système proposée selon le niveau de complexité du texte artistique

5. Conclusion:

Dans cette chapitre, nous avons présenté le principe de fonctionnement de notre système de segmentation en utilisant un algorithme structurel basé multi-scans pour surmonter le problème d'irrégularité, de chevauchements et des intersections dans l'écriture manuscrite artistique.

En conclusion, Notre contribution, fait saut quantique dans le domaine d'OCR, nous évoquons le fait que notre systèmes développée achèvent des résultats satisfaisants Malgré la difficulté de la qualité et la nature du processus d'écriture et le manque de méthodes spécialisées dans le traitement de ce type d'écriture.

Conclusion générale

L'objectif de nos travaux est l'amélioration du taux de segmentation d'un système hors-ligne de reconnaissance de l'écriture manuscrite artistique en intégrant un algorithme de segmentation basé multi-scans (N-Scans).

Un texte arabe manuscrit artistique repose beaucoup plus sur la beauté des formes des lettres et leurs liaisons, ainsi que la forme globale de point de vue esthétique sans préserver la forme standard du caractère. Cette écriture est composée d'un ensemble de caractères qui possèdent une représentation irrégulière, ce qui génère des chevauchements et des intersections entre les différents caractères. De plus, plusieurs formes diacritiques et de décoration sont ajoutées pour donner plus de beauté de la forme globale de l'écriture. Plus de ça, les espaces entre les pseudo-mot et même les caractères est irrégulier.

Nous avons dégagé de cette étude une contribution qui profiter de la structure de l'écriture en utilisant un algorithme structurel basé multi-scans pour surmonter le problème d'irrégularité, de chevauchements et des intersections dans l'écriture manuscrite artistique.

Le problème d'irrégularité, de chevauchements et des intersections ne peut être résolu qu'après avoir définir des points d'intérêt décrivant la structure de l'écriture artistique en détail. Ces points très utiles pour définir un ensemble de règles de segmentation. Ces vont être capable d'identifier les différents boucles, chevauchements et intersections, ainsi que les différents connecteurs des caractères. La méthode proposée est testée sur un ensemble d'image de teste de différents niveaux de complexité, et elle a donné un taux de segmentation très intéressant.

Le travail que nous avons réalisé durant ce mémoire, constitue une étape et un premier apport pour la segmentation structurelle l'écriture manuscrite arabe, cependant nous pensons qu'il peut être amélioré, et étendu par les points suivants :

- Ajouter d'autres règles pour gérer des écritures artistiques de complexité plus élevée que celle utilisée dans les tests.
- Améliorer l'algorithme pour qu'il puisse fonctionner sur des différents types d'écritures.

BIBLIOGRAPHIE

- [1] HAITAAMAR Shahrazed: Mémoire Magister en informatique Option Informatique Industrielle " Segmentation De Textes En Caracteres Pour La Reconnaissance Optique De L'écriture Arabe " , Université El-Hadj Lakhdhar Batna,2007
- [2] Ben Guessoum Imane: Mémoire MASTER ACADEMIQUE" La Reconnaissance En-Ligne Du Manuscrit Arabe Proposition d'un modèle de classificateur logique, application à la reconnaissance du texte arabe imprimé. " , Université D'el-Oued-,2014.
- [3] M. ZAIZ Faouzi, Thèse Doctorat en sciences Informatique: "Technique basée puzzle/SVM pour l'amélioration de la reconnaissance du texte arabe manuscrit", Université Mohamed Khider – BISKRA, 2017.
- [4] Safa ABA , Maroua CHIKH : Memoire De Master Informatique , Option : Système et Multimédia Intitulé " Reconnaissance Des Mots Arabes Manuscrises ". Université de Larbi Tébessi –Tébessa, 2016 . etat de l'art d'ocr
- [5] Redouane TLEMSANI, Thèse Doctorat en sciences Informatique Option Reconnaissance des Formes - Intelligence Artificielle: " LA RECONNAISSANCE EN-LIGNE DU MANUSCRIT ARABE ", UNIVERSITE DES SCIENCES ET DE LA TECHNOLOGIE D'ORAN -MOHAMED BOUDIAF-,2012
- [06] M. ZAIZ Faouzi : Mémoire de magister en Informatique "Les Supports Vecteurs Machines (SVM) pour la reconnaissance des caractères manuscrits arabes" ,Université Mohamed Khider – BISKRA,2010. تعديل
- [07] BOUKHAROUBA Abdelhak: These Doctorat En Sciences "Contribution à la segmentation et à la reconnaissance de l'écriture arabe manuscrite". Universite Mentouri -Constantine, 2011.
- [08] Matthieu Clin, Vincent Jeanselme : Projet image 2016 : Reconnaissance de documents — Ensiwiki ,
https://ensiwiki.ensimag.fr/index.php?title=Projet_image_2016:_Reconnaissance_de_documents
- [09] F. Menasri, N. Vincent, E. Augustin, M. Cheriet. Un système de reconnaissance de mots arabes manuscrits hors-ligne sans signes diacritiques. Antoine Tabbone et Thierry Paquet. Colloque International Francophone sur l'Ecrit et le Document, Oct 2008, France. Groupe de Recherche en Communication Ecrite, pp.121-126, 2008. HAL Id: hal-00334406
"https://hal.archives-ouvertes.fr/hal-00334406".
- [10] Mehennaoui Zahra. Mémoire Magister Option Intelligence Artificielle Distribuée (IAD) "Reconnaissance De L'écriture Arabe Manuscrite A Base Des Machines A Vecteurs De Supports", UNIVERSITE BADJI –MOKHTAR – ANNABA, 2006 .
- [11] MEGAR Ishak et GEUDDA Adel. Mémoire MASTER PROFESSIONNEL " Contribution à la reconnaissance d'écriture arabe manuscrite en utilisant la classification neuronale ", UNIVERSITE KASDI MERBAH OUARGLA, 2016 .
- [12] Najoua Ben Amara, Abdel Belaïd et Nouredine Ellouze : "Utilisation des modèles markoviens en reconnaissance de l'écriture arabe : Etat de l'art"
https://members.loria.fr/ABelaid/publis/cifed00_najoua.pdf.
- [13] KAMEL Mohamed: Mémoire Magister en informatique Option Intelligence Artificielle "reconnaissance de formes appliquée a l'écriture arabe manuscrite par des multi classifieurs" Université Mohamed Khider Biskra,2010.

- [14] H. MILED, C. OLIVIER, M. CHERIET*, K. ROMEO-PAKKER : "Une Méthode Rapide de Reconnaissance de l'Écriture Arabe Manuscrite" Seizième Colloque Grets — 15-19 Septembre 1997 — Grenoble .
- [15] Younes LASRI: " Contribution à la reconnaissance optique (OCR) du texte arabe imprimé " université sidi mohamed ben Abdellah, 2014
- [16] LAKHDARI Ahmed Toufik et ABBACI Salih: " La reconnaissance des caractères arabes manuscrits par les réseaux des neurones convolutionnels " , Université Kasdi Merbah Ouargla ,2017 .
- [17] Farès MENASRI: Thèse doctorat " Contributions à la reconnaissance de l'écriture arabe manuscrite" , l'Université Paris Descartes,2008 .
- [18] Leila CHERGUI , Thèse Docorat Spécialité : Informatique " Combinaison de classifieurs pour la reconnaissance de mots arabes manuscrits " , Université Mentouri Constantine,2013
- [19] Bernard GOSSELIN: Dissertation originale l'obtention du grade de Docteur en Sciences Appliquées" Application De Réseaux De Neurones Artificiels A La Reconnaissance Automatique De Caractères Manuscrits " , Faculté Polytechnique de Mons,1996.
- [20] Azizi Rebiai: Thèse De Magister " une approche hybride pour la reconnaissance d'écriture arabe manuscrite" , UNIVERSITE DE CONSTANTINE,2006لم يذكر في الفصل الأول
- [21] DERDOUR Khedidja: Mémoire de Magistère en Génie Industriel "Reconnaissance de formes du chiffre arabe imprimé : Application au code à barre d'un produit" , Université HADJLAKHDAR -BATNA,2009.
- [22] Redouane TLEMSANI: Thèse de Doctorat En Sciences En Informatique Option Reconnaissance des Formes - Intelligence Artificielle " La Reconnaissance En-Ligne Du Manuscrit Arabe " , Université Des Sciences Et De La Technologie D'oran -Mohamed Boudiaf-,2012.
- [22] Redouane TLEMSANI: Thèse de Doctorat En Sciences En Informatique Option Reconnaissance des Formes - Intelligence Artificielle " La Reconnaissance En-Ligne Du Manuscrit Arabe " , Université Des Sciences Et De La Technologie D'oran -Mohamed Boudiaf-,2012.
- [24] Hanene BOUKERMA: Mémoire MAGISTER en INFORMATIQUE Option : INTELLIGENCE ARTIFICIELLE "Combinaison de classifieurs flous pour la reconnaissance de l'écriture arabe manuscrite. " , UNIVERSITE 20 AOUT 1955 – SKIKDA-,2010.
- [25] Mohamed ELYAAKOUBI: Thèse de Doctorat en Informatique : " étude et la réalisation d'un système de justification du e-document scientifique multilingue " , UNIVERSITE CADI AYYAD – MARRAKECH ,2009.
- [26] Zoizou ABDELHAY: Projet fin d'étude Master Informatique : " Contribution Contribution à la segmentation du texte arabe imprimé " , UNIVERSITE Université Sidi Mohamed Ben Abdellah,2017.
- [27] Projet national de recherche : rapport final. "Traitement Automatique Des Chèques Postaux Algériens" ,Chef de projet : CHIBANI Youcef, Faculté d'électronique et d'informatique (Direction Générale de la Recherche Scientifique et du Développement Technologique du Ministère de l'Enseignement Supérieur et de la Recherche Scientifique) , Programme national de Recherche : 2011-2013.