

République Algérienne Démocratique et Populaire.
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique.

Université Saad Dahlab, Blida
USDB.

Faculté des sciences.
Département informatique.



**Mémoire pour l'obtention
d'un diplôme d'ingénieur d'état en informatique.**
Option : Système d'information (SI).

Sujet :

**Segmentation des mots arabes
imprimés en caractères**

Présenté par : ZAIMI Djamel
RABIA Said

Promoteur : D. MOKHTARI
Encadreur : A. SEHAD
L. MEZAI

Organisme d'accueil : Centre de développement des technologies avancées (CDTA).

Soutenue le: _____, devant le jury composé de :

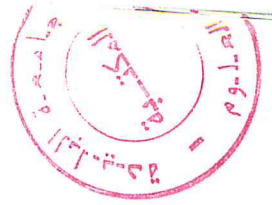
Président

Examinateur

Examinateur

- 2004/2005-

REMERCEMENTS



Toute notre gratitude, grâce et remerciement au bon Dieu, qui nous a donné la force, le courage et la volonté d'élaborer ce travail.

Nous tenons à exprimer nos vifs remerciements à nos encadreurs M^r A. *SEHAD* et M^{lle} L. *MEZAI* qui ont eu l'amabilité de nous proposer ce sujet, et ensuite pour leurs encouragements permanents, leur fructueux conseil, ainsi leur mise à notre disposition leur précieuse documentation.

Nous remercions aussi vivement notre promotrice M^{lle} : *D. MOKHTARI* chargé de cours à l'université de Blida, pour sa compréhension, ses précieux conseils, ainsi sa disponibilité et sa confiance en nous.

En fin, nous remercions l'ensemble des professeurs et tout le personnel, de l'institut d'informatique de l'université de BLIDA.



Dédicace

Je dédie ce modeste travail à :

- Ma très chère mère
- Mon très cher père

Qui m'ont encouragé et soutenu durant toute ma vie.

- Mes sœurs et mes frères en particuliers : **Abdelnacer, Latifa et Mohamed.**
- Toute ma famille.

- Mes remerciements vont aussi pour mes amis

Maamar, Houcine, Ahmed, Brahim, Amin, Salah, Farid, Mustapha, Youcef, Djamel, Elyas, Redouan, Chafik, Ismail, et Said

Et

Tous mes frères de la cité 1 et la cité 2



ZAIMI DJAMEL

Dédicaces

Je dédie ce modeste travail à :

Ma chère mère.

Mon cher grand frère, son épouse et leur petites filles : Nariman et Lamia.

Mes sœurs et frères.

Toute ma famille.

Tous mes amis et amies en particuliers : Abdelmalek, Sallah, Brahim, Ahmed, Kamel, AbdelKarim, Abdelwahab, Amar

Mon binôme Djamel.

Tous mes frères de la cité 1 et la cité 2.

Rabia Saïd.

Résumé

La reconnaissance de l'écriture a pris un grand intérêt de la part des chercheurs, mais malgré cela, la réalisation d'une machine capable de convertir des documents du format papier vers un format électronique reste encore un problème ouvert.

Avant de reconnaître les caractères, il faut d'abord les isoler. Pour les caractères imprimés latins, cette tâche est plutôt facile, puisqu'ils sont déjà séparés par des vides pour la plupart des polices. Ce n'est pas le cas en ce qui concerne les caractères arabes imprimés qui sont eux collés les uns aux autres et ont des formes cursives. La segmentation sert d'abord à isoler les lignes puis les lignes en mots, et ensuite les mots en caractères isolés avant d'être reconnus.

L'objectif de notre travail est de concevoir un système de segmentation des mots arabes imprimés quelque soit leur forme. Le système développé s'articule autour de deux modules distincts, un module de prétraitement, un module de segmentation.

Mots clés : OCR, reconnaissance optique des caractères, mot arabe imprimé, histogrammes, segmentation.

Abstract

The recognition of the writing take a big interest by the researchers, but in spite of it, the realization of a machine able to convert paper documents into electronic version remains an open problem.

Before recognizing the characters, first it is necessary to isolate them. For the Latin printed characters, this task is rather easy, since they are already separated by blanks for most polices. It is not the case with Arabic printed characters because they are joined and have cursive shapes. The segmentation serves to isolate the lines then the lines in words, and then the words in isolated characters before being recognized.

The objective of our work is to conceive a segmentation system of the Arabic printed words with various shapes. The developed system articulates around two distinct modules, pretreatment and segmentation.

Key words: OCR, optical characters recognition, printed Arabic word, histograms, segmentation.

Table des matières

Table des matières

| | |
|---|----------|
| INTRIDUCTION GENERALE..... | 1 |
| <i>CHAPITRE I</i> : généralités sur le traitement d'image..... | |
| I.1. Introduction | 3 |
| I.2. Notions de base | 3 |
| I.2.1 Définition d'une image..... | 3 |
| I.2.2. Pixel..... | 4 |
| I.2.3. Niveau de gris..... | 4 |
| I.2.4. Couleurs..... | 4 |
| I.2.5. Résolution..... | 5 |
| I.2.6. Contour et Texture..... | 5 |
| I.2.7. Voisinage..... | 5 |
| I.2.8. Histogramme..... | 6 |
| I.2.9. Bruit | 7 |
| I.3. Systèmes de traitement d'images..... | 7 |
| I.3.1. Acquisition d'images..... | 7 |
| I.3.2. Pré traitement d'images..... | 7 |
| I.3.3. Analyse..... | 11 |
| I.4. Conclusion..... | 12 |
| | |
| <i>CHAPITRE II</i> : Généralités sur les systèmes de reconnaissance de caractère arabe (OCR) | |
| II.1. Introduction..... | 13 |
| II.2. But d'OCR..... | 13 |
| II.3. Architecture d'un système OCR | 13 |
| II.3.1. Module d'acquisition..... | 14 |
| II.3.2. Module de prétraitement..... | 15 |
| II.3.3. Module d'analyse et de segmentation..... | 15 |
| II.3.4. Module d'extraction des primitives | 15 |
| II.3.5. Module d'apprentissage | 16 |
| II.3.6. Dictionnaire de prototype | 16 |
| II.3.7. Reconnaissance et décision | 16 |
| II.4. Domaine d'application de l'OCR | 17 |

| | |
|--|----|
| II.5. Différents aspects de l'OCR..... | 17 |
| II.6. Conclusion | 18 |

CHAPITRE III : Etat de l'art sur les méthodes de segmentation des mots arabes en caractère

| | |
|--|----|
| III.1. Introduction..... | 19 |
| III.2. Les outils des techniques de segmentation..... | 20 |
| III.2.1. Histogramme | 20 |
| III.2.2. Rectangulation | 21 |
| III.3. Processus de segmentation d'un document imprimé..... | 22 |
| III.3.1. Segmentation du texte en lignes..... | 22 |
| III.3.2. Segmentation de ligne en mots..... | 23 |
| III.3.3. Segmentation du mot en caractères isolés..... | 24 |
| III.4. Les différentes méthodes de segmentation des mots arabes en caractères..... | 25 |
| III.4.1. La méthode de Adnan Amin..... | 25 |
| III.4.2. La méthode de Adnan Amin améliorée..... | 27 |
| III.4.3. La proposition de Bouhlila..... | 29 |
| III.4.4. La méthode Over-segmentation..... | 30 |
| III.4.5. La proposition de la Hamami & Berkani..... | 32 |
| III.4.6. La méthode de Adnan Amin et Humoud B. Al-Sadoun..... | 35 |
| III.5. Tableau comparatif..... | 39 |
| III.6. Conclusion..... | 40 |

CHAPITRE IV : Etude conceptuelle.

| | |
|---|----|
| IV.1. Introduction..... | 41 |
| IV.2. Processus de segmentation..... | 41 |
| IV.2.1. Segmentation par la méthode de A. Amin améliorée..... | 42 |
| IV.2.1.1. Calcul des histogrammes..... | 42 |
| IV.2.1.2. Détection de la ligne de jonction..... | 43 |
| IV.2.1.3. Calcul du seuil de segmentation..... | 44 |
| IV.2.1.4. Critère de segmentation..... | 45 |
| IV.2.2. Contrôle de segmentation..... | 45 |
| IV.2.2.1. Les caractères commençant ou se terminant par des traits amincis..... | 45 |
| IV.2.2.2. Les lettres possédant des points | 46 |
| IV.2.2.3. La lettre 'Sin' | 51 |

| | |
|---|-----------|
| VI.3. Conclusion | 52 |
| CHAPITRE V : Présentation du logiciel | |
| V.1. Introduction..... | 54 |
| V.2. Environnement de travail..... | 54 |
| V.3. Description du logiciel..... | 56 |
| V.3.1. Menus..... | 59 |
| V.2.2. Barre des buttons..... | 60 |
| V.4. Résultats expérimentaux..... | 61 |
| V.5. Conclusion..... | 66 |
| CONCLUSION GENERALE..... | 68 |
| BIBLIOGRAPHIE..... | 70 |
| ANNEXE : Etude morphologique de l'écriture arabe | |

Liste des figures

| | |
|--|----|
| Fig.I.1. Le contour et la texture..... | 5 |
| Fig.I.2. Voisinage d'un pixel de coordonnées (x, y)..... | 6 |
| Fig.I.3. Histogramme des niveaux de gris d'une image..... | 6 |
| Fig.I.4. Système de traitement d'images..... | 7 |
| Fig.I.5 : (a) Image originale, (c) Image binarisée..... | 7 |
| Fig.I.6. Exemple de l'opération de filtrage médiane..... | 9 |
| Fig.I.7. Principe de fonctionnement de La dilatation..... | 9 |
| Fig.I.8. Exemple de l'opération de dilatation..... | 10 |
| Fig.I.9. Principe de fonctionnement de l'érosion..... | 10 |
| Fig.I.10. Exemple de l'opération d'érosion..... | 10 |
| Fig.II.1. Système de reconnaissance de caractère (OCR)..... | 14 |
| Fig.III.1. Type des histogrammes..... | 21 |
| Fig.III.2. Enveloppe connexe rectangulaire..... | 22 |
| Fig.III.3. Segmentation du texte en lignes par l'histogramme..... | 22 |
| Fig.III.4. Segmentation du texte en lignes par rectangulation..... | 23 |
| Fig.III.5. Segmentation de ligne en mots (et en sous mots) par l'histogramme..... | 24 |
| Fig.III.6. Segmentation de la ligne en mots (ou sous mots) par rectangulation..... | 24 |
| Fig.III.7. Segmentation d'un mot arabe par la méthode de Amin..... | 26 |
| Fig.III.8. La variation du tracé principale dans le même caractère..... | 27 |
| Fig.III.9. Anomalie de la méthode de A. Amin..... | 29 |
| Fig.III.10. Segmentation d'un mot arabe par la proposition de Bouhlila. | 29 |
| Fig.III.11. Segmentation d'un mot en caractère par la méthode Over-segmentation..... | 32 |
| Fig.III.12. Segmentation en caractères..... | 33 |
| Fig.III.13 Anomalie de la proposition..... | 34 |
| Fig.III.14. (a) Image originale, (b) Squelette du mot..... | 35 |
| Fig.III.15. Code de Freeman pour un voisinage 3x3..... | 35 |
| Fig.III.16. Exemple de calcul du code de Freeman..... | 35 |
| Fig.III.17. Caractères avec boucles et leurs primitives..... | 36 |
| Fig.III.18. Structure d'un nœud. | 36 |

| | |
|--|----|
| Fig.V.8. La fenêtre "About..." | 59 |
| Fig.V.9. Résultat de Binarisation..... | 62 |
| Fig.V.10. Résultat de Filtrage..... | 62 |
| Fig.V.11. Résultat de la Dilatation..... | 63 |
| Fig.V.12. Avantage de la méthode..... | 63 |
| Fig.V.13. Inconvénient de la méthode..... | 63 |
| Fig.V.14. Avantage de la méthode..... | 63 |
| Fig.V.15. Inconvénient de la méthode..... | 64 |
| Fig.V.16. Amélioration proposée..... | 64 |
| Fig.1. Exemples de styles d'écriture arabe..... | V |

Liste des algorithmes

| | |
|---|----|
| Algorithme 01 : Méthode de A. Amin..... | 26 |
| Algorithme 02 : Méthode de A. Amin améliorée..... | 28 |
| Algorithme 03 : Proposition de Bouhlila..... | 30 |
| Algorithme 04 : Proposition de Hamami & Berkani..... | 34 |
| Algorithme 05 : Histogramme horizontal..... | 42 |
| Algorithme 06 : Histogramme vertical..... | 42 |
| Algorithme 07 : Histogramme vertical Haut..... | 43 |
| Algorithme 08 : Ligne de jonction..... | 43 |
| Algorithme 09 : Seuil de segmentation..... | 44 |
| Algorithme 10 : Contrôle les caractères commencent ou terminent par un trait aminci..... | 46 |
| Algorithme 11 : Contrôle les lettres des classes 1 et 4..... | 48 |
| Algorithme 12 : Contrôle la lettre Sin..... | 52 |

Liste des tableaux

| | |
|---|-----|
| Tab1 : Tableau comparatif..... | 40 |
| Tab.2. Comparaison des résultats de segmentation..... | 66 |
| Tab.3. L'alphabet arabe..... | II |
| Tab.4. Les voyelles de la langue arabe..... | III |
| Tab.5. Exemple de ligature..... | III |
| Tab.6. Classification des lettres de l'alphabet arabe..... | IV |

Introduction générale

Introduction générale

Le problème de la reconnaissance de l'écriture imprimée remonte aux années 90, au cours desquelles, on inventa le scanner à balayage pour la télévision et les machines à lire. C'est un problème extrêmement complexe qui n'a pas de solution satisfaisante à ce jour. Ainsi, les chercheurs ont restreint leurs études à des problèmes particuliers qui sont en relation avec des applications bien définies. Parmi celles-ci, on cite les applications de la reconnaissance de l'écriture manuscrite : le tri automatique du courrier postal, le traitement automatique des dossiers administratifs, des formulaires d'enquêtes, ou encore l'enregistrement des chèques bancaires.

Contrairement aux écritures latines qui bénéficient de tentatives de réalisation de machines ou de logiciels de lecture optique, les travaux sur l'écriture arabe sont assez rares. A l'heure actuelle, il n'existe pas de système OCR performant pour l'écriture arabe, ce retard est dû essentiellement aux spécificités propres de l'écriture arabe.

Le mot constitue l'entité minimale et nécessaire à la compréhension d'un texte. Ainsi la reconnaissance d'un texte demande : soit de reconnaître chacun des mots qui le structurent, d'une manière globale, soit de les décomposer en lettres reconnues séparément.

La langue arabe est très riche en mots, ce qui rend la reconnaissance de mots d'une manière globale impossible, on opte alors pour la reconnaissance des caractères séparément. Ces entités de base n'étant pas initialement isolées, elles font partie d'un ensemble souvent structuré en mots, lignes et paragraphes, ce qui explique la nécessité d'un module de segmentation pour pouvoir effectuer la séparation des paragraphes en lignes, les lignes en mots et les mots en lettres. Plusieurs méthodes de segmentation ont été développées, néanmoins, la solution de A. Amin s'avère la meilleure. Malgré cela, elle marque toujours une insuffisance remarquable.

Dans le travail qu'on présente ici, on s'intéresse surtout à la partie segmentation de mots arabes en caractères qui est plutôt délicate à cause de la forme des caractères arabes. Nous devons proposer donc une méthode qui est basée sur le principe de la méthode améliorée de A. Amin en ajoutant quelques contrôles afin d'améliorer le résultat de segmentation pour certaines classes de caractères et en se basant sur les caractéristiques morphologiques de ces derniers.

Notre mémoire est organisé de la façon suivante :

Nous présentons dans le chapitre I quelques généralités sur le traitement d'images.

Le chapitre II est consacré à la description des systèmes O.C.R (Reconnaissance Optique des Caractères).

Dans le chapitre III, nous présentons un état de l'art sur les méthodes de segmentation d'un mot arabe en caractères.

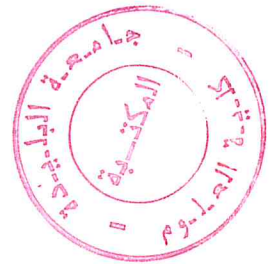
Le chapitre IV, est consacré à l'étude conceptuelle de notre application.

Dans le chapitre V, nous décrivons l'interface du logiciel ainsi que les résultats expérimentaux.

Enfin, nous terminerons avec une conclusion générale et des perspectives.

Chapitre I

Généralités sur le traitement d'images



I.1. Introduction

L'apparition de l'ordinateur a réduit considérablement le temps de calcul consacré aux applications complexes. L'intégration du monde graphique a été le premier pas vers la manipulation d'images sur un micro ordinateur, ceci a donné naissance à un nouveau domaine "le traitement d'images".

Le traitement d'images est un domaine qui est en pleine expansion, l'un de ses objectifs de base est de remplacer l'observateur humain par une machine, pour cela les images ont été numérisées pour pouvoir être traitées par l'ordinateur.

Le traitement d'images est un ensemble de techniques appliquées aux images, ces techniques permettent de :

- Faciliter la manipulation des images (stockage).
- Améliorer la qualité des images.
- Chercher à détecter la présence de certaines formes (contour, texture, ...).
- Réduire la quantité d'information par codage ou compression.

I.2. Notions de base

I.2.1. Définition d'une image

A l'échelle de l'observation visuelle, une image est une représentation bidimensionnelle d'objets tridimensionnels de nature diverse (pièces industrielles, vues aériennes, image médicales, scènes naturelles,...).

Elle contient en chaque point l'intensité lumineuse perçue par une caméra (ou tout autre système de mesure) en ce point. Elle rend compte de la variation du contraste dans la scène. L'intensité lumineuse varie de façon uniforme entre le blanc (beaucoup de réflexion de lumière, significative de forme non contrastée) et le noir (réflexion nulle de la lumière due à la présence d'un fort contraste). L'image est illustrée par une matrice de points. Chaque point est caractérisé par des informations tel que le niveau de gris et la position [Mar 89].

On distingue :

- *L'image analogique* : elle est fournie généralement par le capteur, et décrite sous la forme d'une fonction analogique $x_a(u,v)$.
Où u et v dénotent les coordonnées spatiales et $x_a(u,v)$ est l'intensité lumineuse au point (u,v) .

- *L'image numérique (digitale)* : pour pouvoir manipuler une image sur un système informatique, il est nécessaire de lui faire subir une transformation, qui la rendra lisible et manipulable par le système.

Le rôle de système d'acquisition est de transformer une image d'origine analogique, en une matrice de valeur numérique (digital) [Mar 89].

I.2.2. Pixel

Le pixel est l'abréviation de « Picture Element », c'est la plus petite entité qui compose l'image, il est caractérisé par deux attributs [Gui 96] :

- Une position dans la matrice d'image (numéro de ligne, numéro de colonne)
- Une valeur numérique indiquant sa couleur, elle peut être sur un, deux ou trois octets.

I.2.3. Niveau de gris

Le niveau de gris est la valeur d'intensité lumineuse en un point. La couleur du pixel peut prendre des valeurs allant du noir au blanc en passant par un nombre fini de niveaux intermédiaires. Ainsi, pour représenter les images à niveaux de gris, nous pouvons attribuer à chaque pixel de l'image une valeur correspondant à la quantité de lumière renvoyée. Cette valeur peut être comprise par exemple entre 0 et 255, dans ce cas le pixel est représenté sur un octet [Gon 77].

I.2.4. Couleurs

Toutes les couleurs que nous voyons dans la nature sont en effet composées des 3 principales couleurs : le ROUGE, le VERT, et le BLEU.

Dans une image couleur, un pixel de coordonnée (x, y) est représenté par trois composantes $R(x,y)$, $V(x,y)$, $B(x,y)$ donnant respectivement son intensité rouge, vert et bleu.

Un pixel d'une image couleur codé sur 24 bits peut par exemple, être représenté en mémoire sur 3 octets :

- un octet pour l'intensité du rouge.
- un octet pour l'intensité du vert.
- un octet pour l'intensité du bleu.

Ce qui donne la possibilité de coder $2^8 * 2^8 * 2^8 = 2^{24} = 16.777.216$ couleurs différentes [Tou 90].

I.2.5. Résolution

C'est le nombre de points utilisés pour représenter une image, c'est aussi le nombre de pixels disposés en ligne (longueur) multiplié par le nombre de pixels disposés en colonnes (largeur) dans une image. La résolution augmente ou diminue lorsqu'on veut agrandir ou réduire la taille de l'image [KUN 93]

Exemple :

Nous voyons la résolution dans notre ordinateur (de l'image à l'écran), on trouve par exemple 800 par 600 pixels : qui veut dire que l'image est représentée sur 800 colonnes et 600 lignes.

I.2.6. Contour et Texture

Une image numérique a été vue comme étant une suite de points dont chacun possède une valeur numérique. Une manière plus moderne et plus naturelle de représenter une image est de la modéliser en termes de contours et de textures. Les contours représentent la frontière entre les objets de l'image alors que les textures décrivent la structure de ceux-ci [KUN 93].

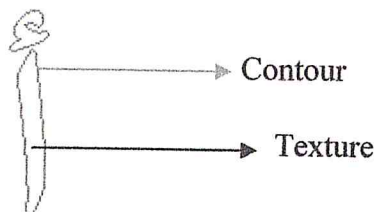


Fig.I.1. Le contour et la texture.

I.2.7. Voisinage

Le voisinage est défini par une matrice, ainsi le voisinage d'un pixel est composé de tous les pixels qui l'entourent, on utilise généralement deux types de voisinage [KUN 93]:

- **Voisinage à 4-connexité :**

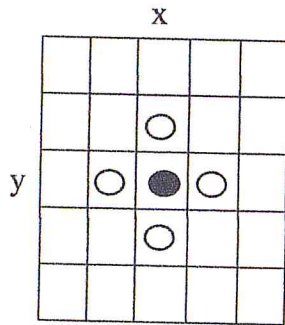
A partir d'un pixel de coordonnées (x, y) , nous pouvons se déplacer vers les quatre pixels voisins dont les coordonnées sont :

$$(x+1, y), (x, y-1), (x-1, y), (x, y+1)$$

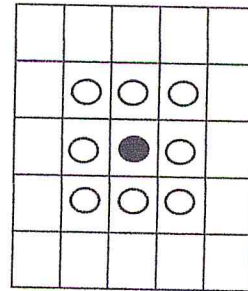
- **Voisinage à 8-connexité :**

A partir d'un pixel de coordonnées (x, y) , on peut se déplacer vers les huit pixels voisins dont les coordonnées sont :

$$(x-1, y-1), (x-1, y), (x-1, y+1), (x, y-1), (x, y+1), (x+1, y-1), (x+1, y), (x+1, y+1).$$



Voisinage 4-connexité



Voisinage 8-connexité

Fig.I.2. Voisinage d'un pixel de coordonnées (x, y) .

I.2.8. Histogramme

L'histogramme des niveaux de gris d'une image est la fonction qui associe à une valeur d'intensité 'i', le nombre de pixels dans l'image ayant cette valeur [KUN 93].

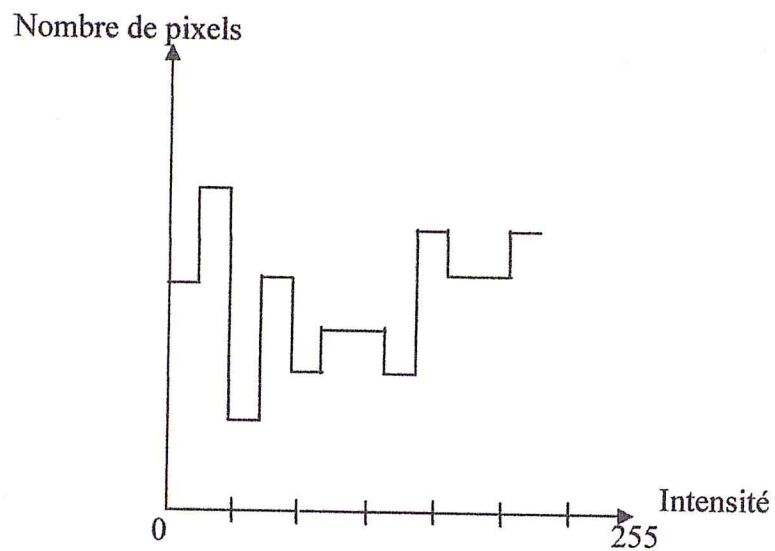


Fig.I.3. Histogramme des niveaux de gris d'une image.

L'histogramme permet de tirer des informations sur la distribution des niveaux de gris, il permet aussi d'avoir les bornes entre lesquels sont répartis la majorité des niveaux de gris dans le cas d'une image trop claire ou trop foncée [Ber 00].

I.2.9. Bruit

Un bruit (parasite) dans une image est considéré comme un phénomène de brusque variation de l'intensité d'un pixel par rapport à ses voisins. Il provient de l'éclairage et des dispositifs optiques [Tou 90].

I.3. Systèmes de traitement d'images

Un système de traitement d'images est composé de plusieurs modules comme le montre la figure ci-dessous :

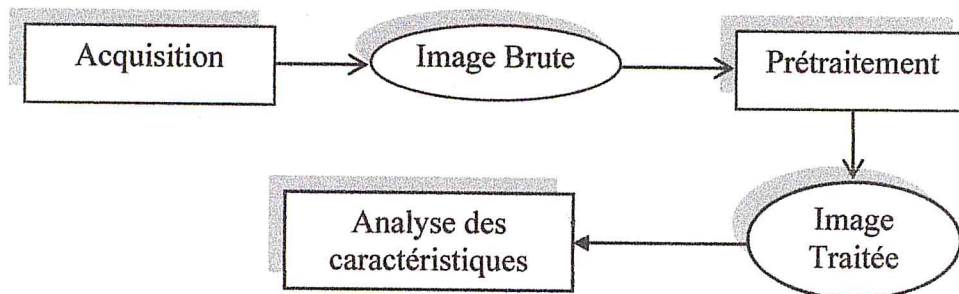


Fig.I.4. Système de traitement d'images.

I.3.1. Acquisition d'image

Pour pouvoir manipuler une image sur un système informatique, il est nécessaire de lui faire subir une transformation qui la rendra manipulable par ce système. Le passage de cet objet externe (l'image d'origine) à sa représentation interne (dans l'unité de traitement) se fait grâce à une procédure de numérisation. Ces systèmes de saisie, dénommés optiques, peuvent être classés en deux catégories principales : les caméras numériques et les scanners.

I.3.2. Pré traitement d'image

Le prétraitement est l'ensemble des opérations préparant l'image acquise à l'étape d'analyse, ces opérations consistent principalement à binariser l'image (seuillage) avant d'éliminer le bruit (filtrage) dû aux conditions d'acquisition [Bou 98].

❖ **Seuillage** : [Bou 98] [Sel 96]

Le seuillage est une méthode utilisée pour segmenter des images. Son principe est le suivant : la valeur d'un pixel à une position donnée est remplacée par '1' si elle est supérieure à un certain seuil donné, '0' sinon.

Le '1' représente le blanc et le '0' représente le noir.

$$\text{Valeur}(x, y) = \begin{cases} 1 & \text{si valeur}(x, y) > \text{seuil} \\ 0 & \text{sinon} \end{cases}$$

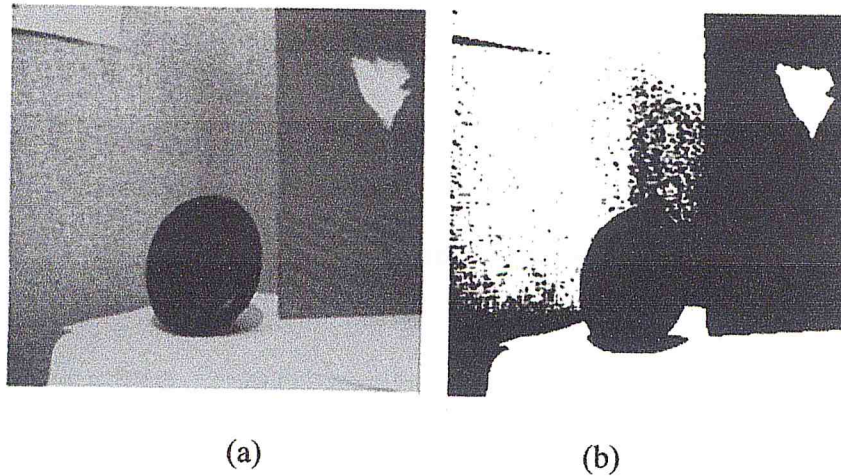


Fig.I.5 : (a) Image originale, (c) Image binarisée.

❖ **Filtrage**

Pour améliorer la qualité visuelle de l'image, on doit éliminer les effets de bruits (parasites) en lui faisant subir un traitement appelé *filtrage*.

Le filtrage consiste soit à atténuer les bruits soit à accentuer les contours pour mieux extraire les informations utiles.

Le filtrage assure plusieurs fonctions telles que :

- L'élimination du bruit.
- L'augmentation ou la diminution de la lumière et le contraste.
- L'agrandissement ou la réduction du signal.

On distingue :

- **Filtre médian** : Le principe est de prendre la valeur médiane sur le voisinage du pixel après avoir effectué un tri des niveaux de gris des pixels voisins.

En général, on effectue une dilatation afin de supprimer les points isolés.

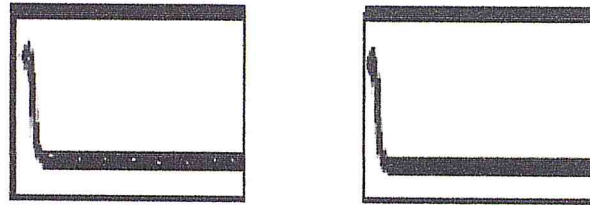


Image originale.

Image dilatée.

Fig.I.8. Exemple de l'opération de dilatation

➤ **Erosion** : elle permet d'éliminer les pixels noirs isolés. On effectue le OU logique des huit voisins du pixel considéré. L'érosion élimine les tâches noires dans les zones blanches.

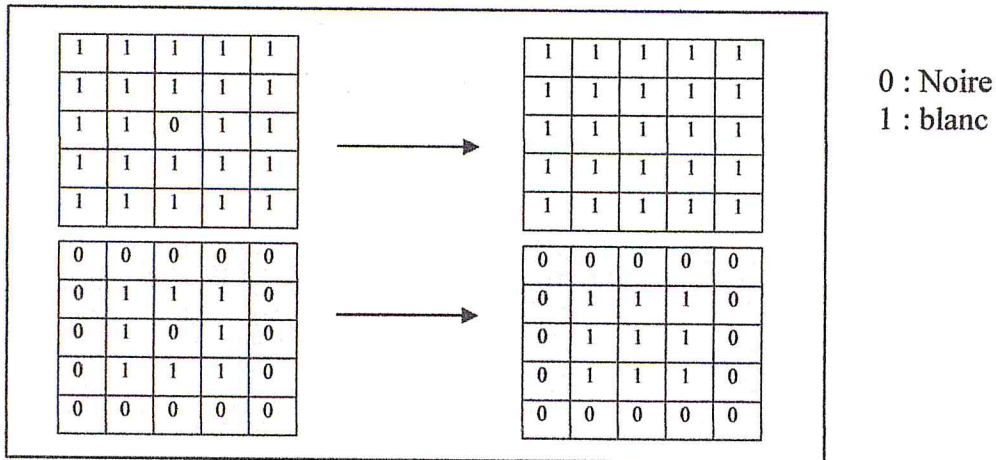


Fig.I.9. principe de fonctionnement de l'érosion.

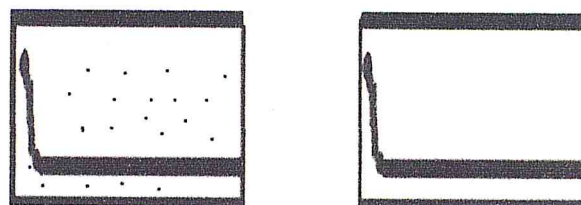


Image originale.

Image érosée.

Fig.I.10. Exemple de l'opération d'érosion.

I.3.3. Analyse

C'est la phase la plus importante dans un système de traitement d'images, car elle s'intéresse à l'analyse et à la compréhension de l'image. Dans l'analyse, on cherche à extraire les informations (contour, région) dans l'image à traiter en se basant sur la technique de segmentation [And 87].

Exemple :

$$\begin{pmatrix} 2 & 12 & 12 \\ 12 & 2 & 60 \\ 2 & 2 & 12 \end{pmatrix} \longrightarrow \begin{pmatrix} 2 & 12 & 12 \\ 12 & 12 & 60 \\ 2 & 2 & 12 \end{pmatrix}$$

Le tri des niveaux de gris : 2 2 2 2 12 12 12 12 60, la valeur médiane est 12 donc le pixel centrale prendra la valeur 12 au lieu de 2.

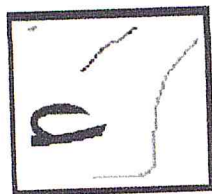


Image originale.

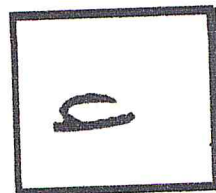
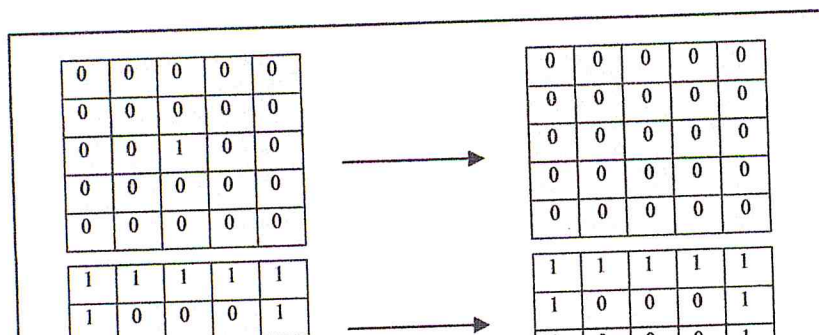


Image filtrée.

Fig.I.6. Exemple de l'opération de filtrage médiane

➤ **Dilatation** : elle permet d'éliminer les pixels blancs isolés. On effectue le ET logique des huit voisins du pixel considéré. La dilatation élimine les tâches (trous) blancs dans les zones noires et ajoute des pixels noirs aux contours des objets présents dans l'image [Khe 98].



0 : Noire
1 : blanc

b) Segmentation par extraction des contours

Mathématiquement, on définit un contour comme une brusque variation du niveau de gris dans l'image.

L'extraction des contours consiste à regrouper les pixels en une primitive pour avoir accès aux caractéristiques plus globales (direction du contour, radiométrie moyenne : l'intensité lumineuse moyenne.).

L'opération d'extraction de contours [Tou 90] peut être décomposée en plusieurs étapes distinctes qui sont :

1. **Mise en évidence des contours** : c'est une étape primordiale qui s'obtient par une dérivation de l'image.
2. **Réduction de l'épaisseur des contours** : l'épaisseur des contours doit être un pixel.
3. **Binarisation des contours** : la binarisation des contours permet d'obtenir une image 'contour' binaire. Cette étape clé permet de définir les contours que l'on considère comme importants et ceux que l'on rejette.
4. **Description des contours** : la description des contours permet d'organiser les contours en structures simples telles que les segments de droite, arcs de cercle, etc..., et de détecter les contours fermés afin d'isoler ceux-ci.

I.4. Conclusion

Le traitement d'image consiste à détecter la présence de certaines formes, certains contours ou certaines textures contenues dans l'image. Un autre aspect de traitement d'image concerne l'analyse de l'image. Il cherche à extraire les informations contenues dans les divers objets de la scène, les techniques de base sont l'extraction de formes (contours, textures) et la segmentation de l'image en segments présentant des caractéristiques spécifiques.

Actuellement, le traitement d'images est utilisé dans plusieurs domaines comme la robotique, la médecine ...etc. Car l'intégration d'une telle technique dans des systèmes informatiques a permis de faire de très grand pas vers l'avant dans le domaine scientifique et technologique.

Après avoir présenté quelques concepts fondamentaux concernant le traitement d'images, le chapitre suivant est consacré au système de reconnaissance des caractères (OCR).

Chapitre II

Généralités sur les systèmes de reconnaissance de caractères (OCR)

II.1. Introduction

La majorité des documents existants dans le monde sont des documents imprimés y compris les journaux, les articles techniques et les revues. Bien que la production des documents sous format électronique augmente, il existe encore une énorme quantité des documents imprimés qui manquent de leurs versions électroniques ce qui rend difficile la recherche d'information et le stockage des documents.

Les systèmes de reconnaissance optique de caractères consistent à récupérer automatiquement des textes à partir d'un document imprimé, en évitant la phase fastidieuse de la saisie, pour cette phase, on peut dire que pouvoir s'affranchir de l'opération de saisie de documents imprimés ou dactylographies grâce à un système automatique constitue un rêve des services informatiques.

D'autre part, les documents électroniques ont plusieurs avantages de point de vue stockage et recherche d'information. Par conséquent, il sera très intéressant d'avoir un système qui convertit les documents imprimés en des documents électroniques. [Nou 02].

II.2. But d'OCR

La transformation des documents en fichiers éditables a beaucoup d'intérêt :

- ❖ **Gagner l'espace de stockage** : un document sous format image occupe beaucoup d'espace mémoire. Pour remédier à ce problème, il faut transformer le document en un fichier texte éditable.
- ❖ **Faciliter la recherche** : La recherche dans une base de données contient des documents sous format image, est une opération très complexe puisque le traitement se fait sur des fichiers images. L'OCR remédie à ce problème, la lecture optique de caractères permet d'archiver vos documents sous forme de fichier texte facile à traiter.
- ❖ **Augmenter la productivité de saisie** : Le problème de la saisie manuelle est devenu l'un des problèmes majeur. Beaucoup d'organismes (entreprises, banques, ...etc.) utilisent les logiciels d'automatisation de l'opération de la saisie.

II.3. Architecture d'un système OCR

Grâce à sa complexité de traitement, aux nombreux types de caractères et chiffres à reconnaître, l'architecture suivante représente les différents modules constituant le système OCR.

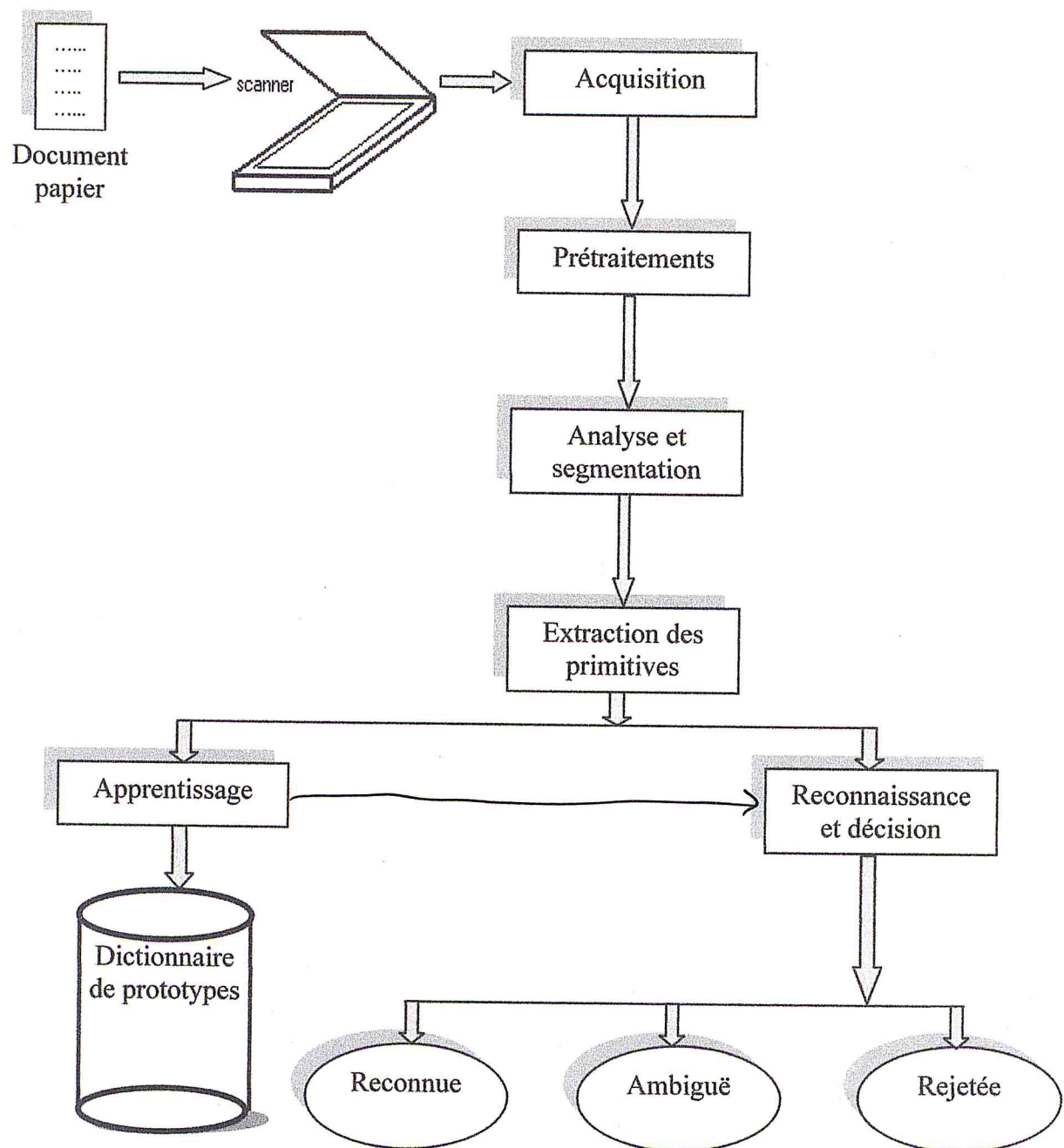


Fig.II.1. système de reconnaissance de caractère (OCR)

II.3.1. Module d'acquisition

L'opération d'acquisition permet le passage de l'information du monde réel (document papier) au monde numérique de l'ordinateur (forme d'image numérique) à travers un organe de captage (scanner, camera...). Ce passage nommé aussi numérisation ou digitalisation.

Le but de cette étape est de rendre l'image exploitable par l'ordinateur; l'acquisition est l'interface entre le système de traitement et l'environnement extérieur.

II.3.2. Module de prétraitement

L'information brute issue de l'opération d'acquisition peut engendrer des bruits dus à l'organe capteur lui-même, ce qui altère la qualité de l'image. Le prétraitement consiste à y remédier cet inconvénient en filtrant la forme saisie. Plusieurs opérations et techniques sont employées, elles dépendent totalement du mode d'acquisition.

II.3.3. Module d'analyse et de segmentation

Le document à traiter peut contenir outre que le texte, des graphismes, des images...etc. La première étape consiste à isoler les régions non textuelles, dans la deuxième étape le texte du document est divisé en parties, dans chaque partie, on sépare les lignes, qui seront divisées en mots et groupes de mots.

Le module d'analyse et de segmentation permet de localiser et d'identifier la structure physique du document, c'est-à-dire les blocs de texte, et d'affiner la segmentation en lignes, en mots puis en caractères.

II.3.4. Module d'extraction des primitives

Ce module permet l'analyse de la forme à reconnaître et délivre une information qui servira à son identification, cette phase d'analyse est étroitement liée à l'approche choisie dans la reconnaissance.

L'extraction de l'information est réalisée à l'aide de l'extraction des primitives (segment, droite, cercle, courbes,...etc.), qui sont des informations élémentaires.

➤ Catégories des primitives

On distingue trois catégories principales de primitives [Dar 94] : les primitives topologiques, structurelles et globales.

1. Primitives topologiques

Ce sont des propriétés topologiques de l'image. La topologie consiste, à appliquer directement sur l'image primitive, les mesures et les tests suivants :

- Compter le nombre de trous d'une forme.
- Mesurer les pentes et d'autres paramètres de courbes puis évaluer des orientations principales.
- Mesurer la longueur et l'épaisseur des traits.

- Détecter les croisements et les jonctions des traits.
- Mesurer les surfaces, le périmètre.

2. Primitives structurelles

Les primitives structurelles sont généralement extraites non pas de l'image brute, mais à partir d'une représentation de la forme (contour). Il s'agit principalement :

- Des segments de droite
- Des arcs, des boucles et des pentes.
- Des angularités, des points extrêmes et des points terminaux, jonction et croisement.

3. Primitives globales

Les primitives globales sont naturellement basées sur une transformation globale de l'image. La caractéristique d'une primitive globale dépend de la totalité des pixels d'une image.

II.3.5. Module d'apprentissage

Pour décider sur une forme donnée, on doit disposer d'un certain nombre de connaissances qui caractérisent clairement la forme injectée. Pour cela les concepteurs des systèmes de reconnaissance des formes ont décidé d'intégrer une base de connaissance appelée **apprentissage**. Le rôle de cette phase est de ranger les formes décrites par les paramètres en un certain nombre de catégories ou de classes. Le résultat de l'apprentissage est la réorganisation ou le renforcement des modèles existants en tenant compte de l'apport de la nouvelle forme, soit la création d'un nouveau modèle représentant la forme entrée.

II.3.6. Dictionnaire de prototype

Il est généré lors de la phase d'apprentissage et contient l'ensemble de prototype retenu, ainsi, il constitue la base de reconnaissance pour les étapes de décision (identification).

II.3.7. Reconnaissance et décision

C'est la dernière étape dans un système de reconnaissance, Elle est dite aussi la classification, cette étape constitue la phase de reconnaissance proprement dite, son rôle est d'identifier la forme ou de la comparer avec les prototypes (les références). La réponse peut être le nom de la forme en cas de bonne reconnaissance, plusieurs noms en cas d'ambiguïté,

ou bien le rejet de la forme en cas d'incompatibilité de la description avec les formes de références.

II.4. Domaine d'application de l'OCR [Yaz 92]

Parmi les domaines d'applications de l'OCR, on peut citer :

- **Les banques** : les lecteurs automatiques permettent la saisie automatique de chèques pour la vérification de la somme d'argent donnée en chiffres et en lettres, ainsi la vérification de signatures.
- **L'éducation** : la reconnaissance et l'apprentissage de l'écriture.
- **La poste** : la lecture des adresses et le tri automatique du courrier.
- **La télécommunication** : pour l'échange de fichiers informatisés à distance.
- **L'administration** : pour la reconnaissance des plans cartographiques et la lecture automatique de documents administratifs.
- **La bureautique** : pour l'indexation et l'archivage automatique des documents.
- **Les affaires et l'industrie** : pour la gestion des stocks et la reconnaissance des documents techniques.

II.5. Différents aspects de l'OCR [Yaz 92]

Avant de citer quelques caractéristiques des systèmes de reconnaissance optique des caractères, il est important de signaler qu'il n'existe pas un système universel d'OCR, mais plutôt des voies d'approches dépendant du type des données traitées et de l'application visée. Nous allons donner quelques caractéristiques des systèmes d'OCR.

➤ **Reconnaissance de l'imprimé ou du manuscrit**

Dans le cas de l'imprimé, les caractères sont bien alignés et souvent bien séparés verticalement ce qui simplifie la phase de lecture. De plus, le graphisme des caractères est conforme à un style calligraphique (fonte) qui constitue un module pour l'identification.

Dans le cas du manuscrit, les caractères sont souvent ligaturés et leur graphisme est inégalement proportionné. Cela nécessite l'emploi de techniques de délimitation très spécifiques et souvent des connaissances contextuelles pour guider la lecture.

➤ Reconnaissance en-ligne ou hors-ligne

La reconnaissance en-ligne (On-line) est une reconnaissance dynamique qui se déroule pendant l'écriture. Ce mode permet à l'utilisateur de corriger et de modifier son écriture d'une manière directe et instantanée. Dans la reconnaissance en-ligne, un stylo spécial permet de connaître l'ordre de tracé, comme dans le cas des tablettes tactile.

La reconnaissance hors-ligne (Off-line ou en différé) démarre après l'acquisition du document entier. Ce mode permet l'acquisition instantanée d'un nombre important de caractères, mais impose d'effectuer des prétraitements coûteux pour retrouver l'ordre de la lecture. Ce mode de reconnaissance concerne les documents scannés

➤ Reconnaissance monofonte, multifonte ou omnifonte :

Cette approche concerne les textes imprimés. On appelle fonte un style de caractères. Un système est dit monofonte s'il ne traite qu'une fonte à la fois. L'apprentissage est simple à réaliser puisque l'alphabet représenté est réduit.

Un système est dit multifonte s'il est capable de reconnaître un mélange de quelques fontes parmi un ensemble de fontes préalablement apprises. Dans ce cas, le prétraitement doit réduire les écarts entre les caractères (taille, épaisseur et inclinaison), l'apprentissage doit gérer les ambiguïtés dues au rapprochement des caractères des différentes fontes, la reconnaissance doit identifier les subtiles différences entre eux.

Enfin, un système est dit omni-fonte s'il est capable de reconnaître toute fonte sans l'avoir absolument appris.

II.6. Conclusion

Dans ce chapitre, nous avons effectué une brève synthèse sur la reconnaissance des caractères imprimés. Nous avons présenté une description générale d'un système de reconnaissance de caractères, et ses différents modules (les caractéristiques les plus importantes pour l'analyse du caractère, les différents procédés d'apprentissage et ainsi les différentes méthodes de reconnaissance).

Bien qu'il existe des systèmes d'OCR très puissants qui ont été commercialisés, mais il reste toujours quelques problèmes à résoudre surtout pour le manuscrit, auquel les systèmes existants n'ont pas donné des résultats très satisfaisants.

Dans le chapitre suivant nous abordons l'état de l'art des méthodes de segmentation des mots arabes en caractères.

Chapitre III

Etat de l'art sur les méthodes de Segmentation des mots arabes en caractères

III.1. Introduction

Dans ce chapitre, nous étudions le problème de localisation des éléments du texte arabe (lignes, interligne, espaces entre les mots et caractères isolés) et la séparation des caractères liés en caractères isolés. Ce problème est défini en traitement d'images par le terme "segmentation" car il s'agit de découper l'image en blocs (segment) contenant le texte.

L'objectif est de réaliser un module de segmentation de l'écriture arabe, qui prend en entrée un document imprimé (sous format image) et fournit en sortie une page de texte où toutes les lignes du texte sont identifiées et les caractères de chaque mot sont séparés les uns des autres. On utilise pour cela les définitions suivantes propre à la structure générale d'un texte arabe imprimé :

- Le texte imprimé est constitué d'un ensemble de lignes horizontales séparées entre elles par des espaces interlignes.
- Une ligne de texte est constituée d'un ensemble ordonné de mots séparés entre eux par des espaces.
- Un mot de la langue arabe est un ensemble d'un ou plusieurs sous mot séparés par des espaces, chaque sous mot est constitué soit d'un seul caractère isolé, soit de plusieurs caractères attachés.

La segmentation est réalisée en trois niveaux :

Niveau 1 : "niveau texte" qui permet de localiser les lignes et les interlignes et la séparation du texte en lignes horizontales.

Niveau 2 : "niveau ligne" qui permet de localiser les mots et la séparation des mots et les sous mots.

Niveau 3 : "niveau mot" qui permet de localiser le début et la fin de chaque caractère d'un mot et le séparer.

Il n'existe pas de méthode universelle qui permet de réaliser la séparation automatique du texte en lignes, les lignes en mots, et les mots en caractères, parmi les techniques principales utilisées pour réaliser la séparation de chaque niveau sont les histogrammes et la rectangulation.

III.2. Les outils des techniques de segmentation

Parmi les outils utilisées dans la détermination des informations contenues dans une image sont les histogrammes et la rectangulation.

III.2.1. Histogramme

L'histogramme est considéré comme l'un des principaux outils d'aide à l'analyse par représentation des informations contenues dans une image. L'image est considérée comme une matrice de pixels binaires (pixel allumé=1, pixel éteint=0), chaque ligne de texte est constituée d'un ensemble de lignes (ligne pixel) et de colonnes (colonne pixel).

L'histogramme est une représentation géométrique d'une fonction qui fournit pour chaque colonne pixel (respectivement ligne pixel), le nombre de pixels allumés dans la colonne pixel (respectivement ligne pixel), représentées sur l'axe des abscisses (respectivement coordonnées). Il permet de donner des informations sur la distribution des pixels de l'image, on peut trouver deux types d'histogrammes :

Histogramme **Horizontal (HH)** est définie de la manière suivante :

$$HH(i) = \sum_j \sum_l M(i, j)$$

Histogramme **Vertical (HV)** est définie de la manière suivante :

$$HV(j) = \sum_l \sum_i M(j, i)$$

Où

M : désigne l'image numérisée.

i correspond à l'indice des lignes et j correspond à l'indice de colonnes .

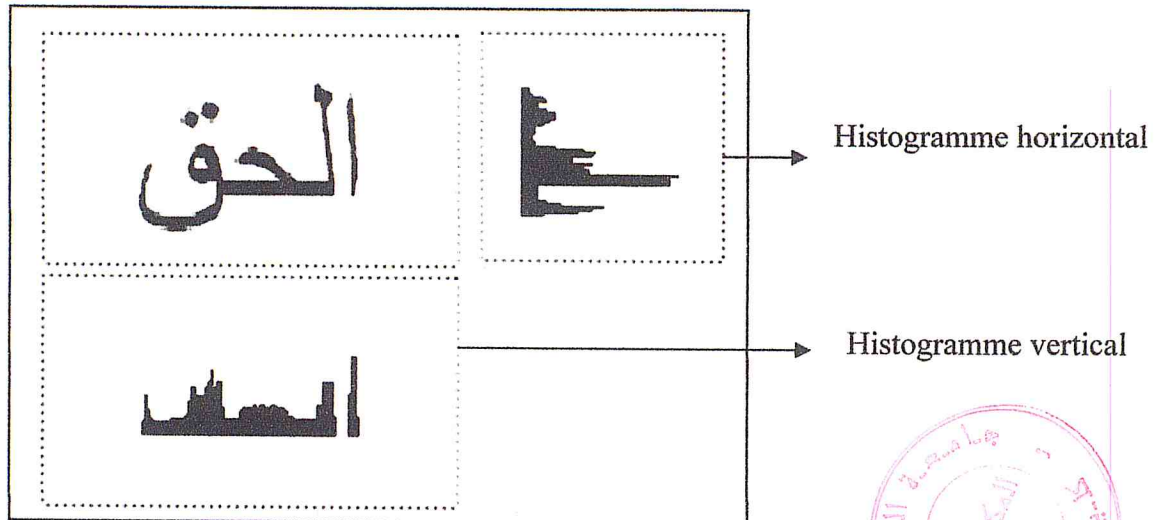


Fig.III.1. Type des histogrammes.

Ainsi, on peut détecter certaines régions particulières de l'image :

- *Région silencieuse* : correspond à une faible concentration de pixels, c'est-à-dire le nombre de pixels allumés est inférieur à un certain seuil.
- *Pics* : correspondent à une forte concentration de pixels allumés, c'est-à-dire le nombre de pixels allumés est supérieur à un certains seuil.

La détermination des seuils se fait expérimentalement selon la méthode utilisée.

III.2.2. Rectangulation

C'est une technique de segmentation en blocs rectangulaires de document imprimés en noir et blanc. Elle consiste à déterminer les rectangle vides maximaux, espaces séparateurs du texte, allongés verticalement ou horizontalement.

Une page d'un document apparaît alors comme un "pavage" d'entités géométriques sensiblement rectangulaires, séparé par des espaces vides. De même une ligne ou "bandes de signes" se compose de paquets de signe sensiblement rectangulaires séparés par des intervalles vides.

La segmentation en bloc est à priori l'opération la plus complexe, puisque l'on travaille sur l'image complète. Les espaces verticaux séparateurs sont les marges de gauche et de droite et les espaces intercolonnes de texte alors que les espaces horizontaux sont les marges supérieures et inférieures et les espaces horizontaux interblocs.

Les espaces carrés plus rares, séparent les titres des colonnes voisines, où les marges de titres sont centrées. Chaque élément de la structure physique (bloc, ligne, mot, parfois caractère) est localisé par les coordonnées de l'enveloppe rectangulaire connexe.

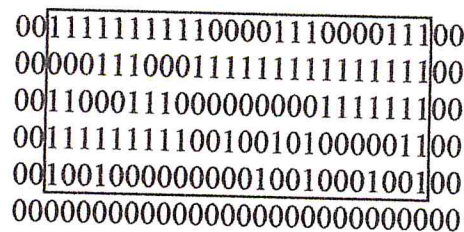


Fig.III.2. enveloppe connexe rectangulaire.

III.3. Processus de segmentation d'un document imprimée

III.3.1. Segmentation du texte en lignes

a. Par l'histogramme

La méthode appliquée pour séparer les lignes du texte consiste à balayer l'image du texte, et à l'analyser dans le sens horizontal ligne par ligne (du haut en bas) en construisant l'histogramme horizontal Hh.

Les zones silencieuses de l'histogramme horizontal sont considérées comme les délimiteurs de lignes correspondant approximativement aux interlignes du texte.

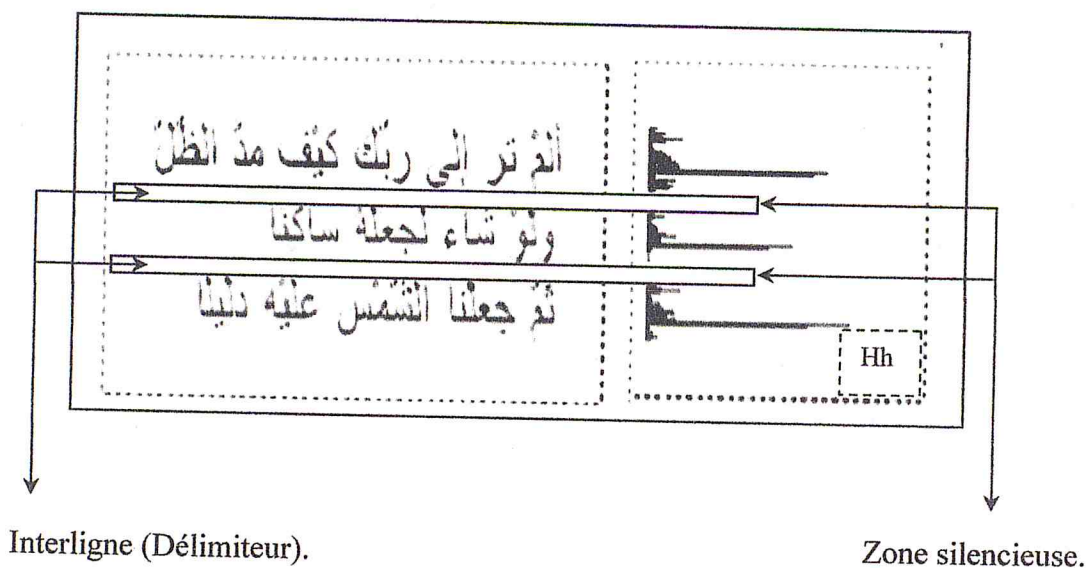


Fig.III.3. Segmentation du texte en lignes par l'histogramme.

b. Par rectangulation

Un bloc d'un document est composé d'une succession de bandes horizontales de signes (lignes) séparées par des bandes vides (interligne). On recherche les bandes vides par des rectangles nuls ayant un seuil horizontal maximal (largeur de la page) et un seuil vertical d'au moins d'un pixel.

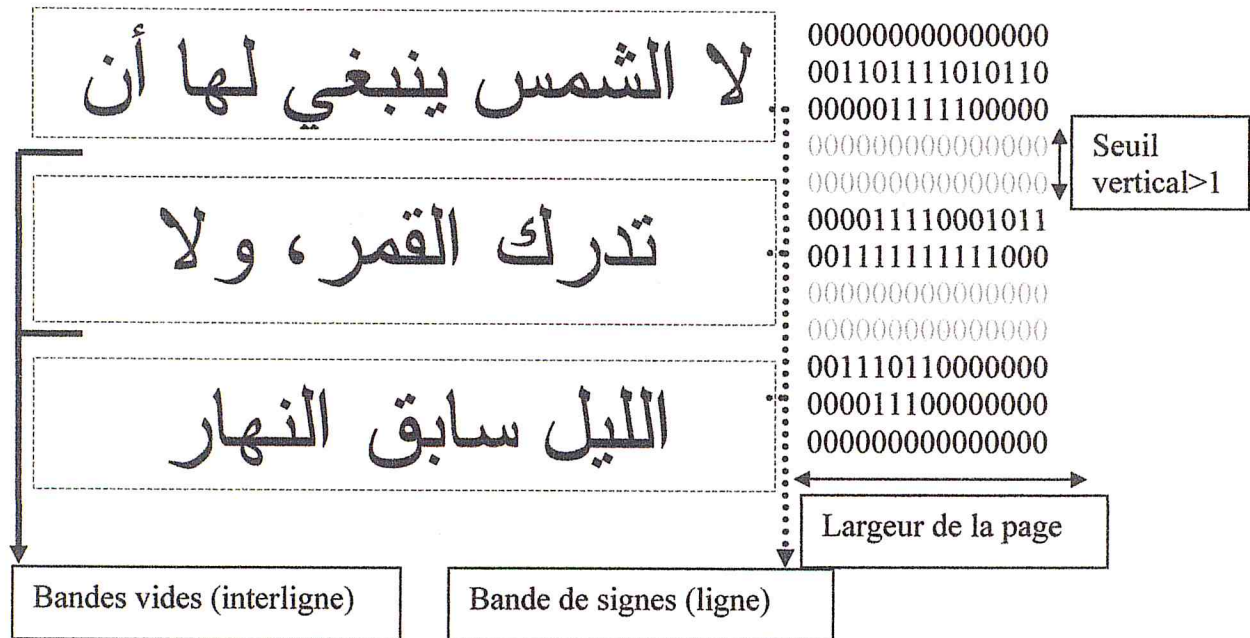


Fig.III.4. Segmentation du texte en lignes par rectangulation.

III.3.2. Segmentation de ligne en mots

b. Par l'histogrammes

La segmentation des mots et sous mots se fait de la manière suivante :

Chaque ligne du texte étant incluse dans un bloc composé de plusieurs lignes et plusieurs colonnes. On effectue alors un balayage vertical (colonne par colonne) du bloc en calculant l'histogramme vertical H_v délimité par les séparateurs de ligne. Les séparateurs des mots (et des sous mot) sont positionnés là où on rencontre les régions silencieuses de l'histogramme vertical.

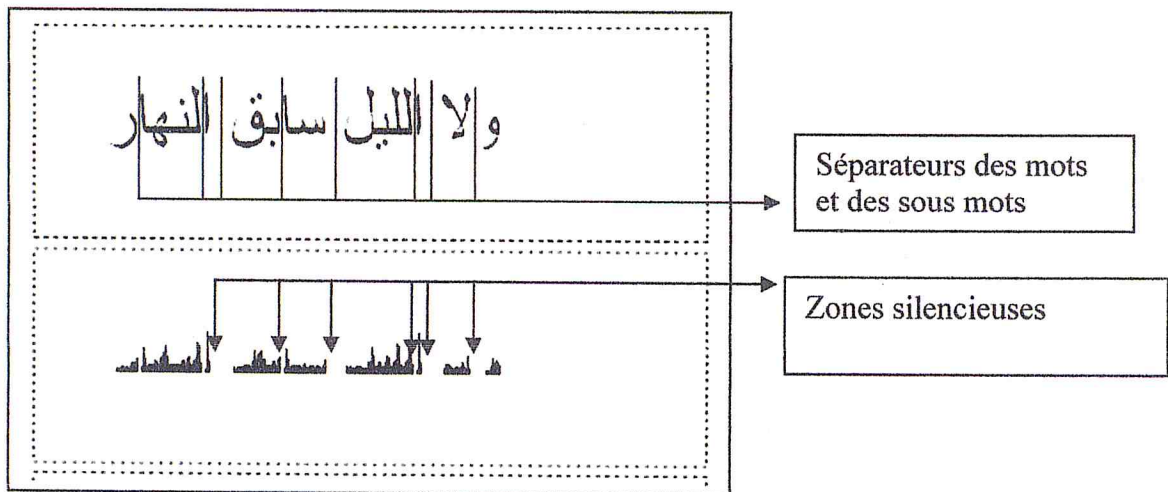


Fig.III.5. Segmentation de ligne en mots (et en sous mots) par l'histogramme.

b. Par rectangulation

Chaque ligne (bande de signes) se compose de paquets de signes (mots) séparés par des zones vides (séparateur vertical). La technique consiste à trouver ces zones qui sont des rectangles de hauteur maximale, et de largeur peut être choisi manuellement, il est d'ordre de 1/2.5 (40%) du seuil verticale [Azz 90].

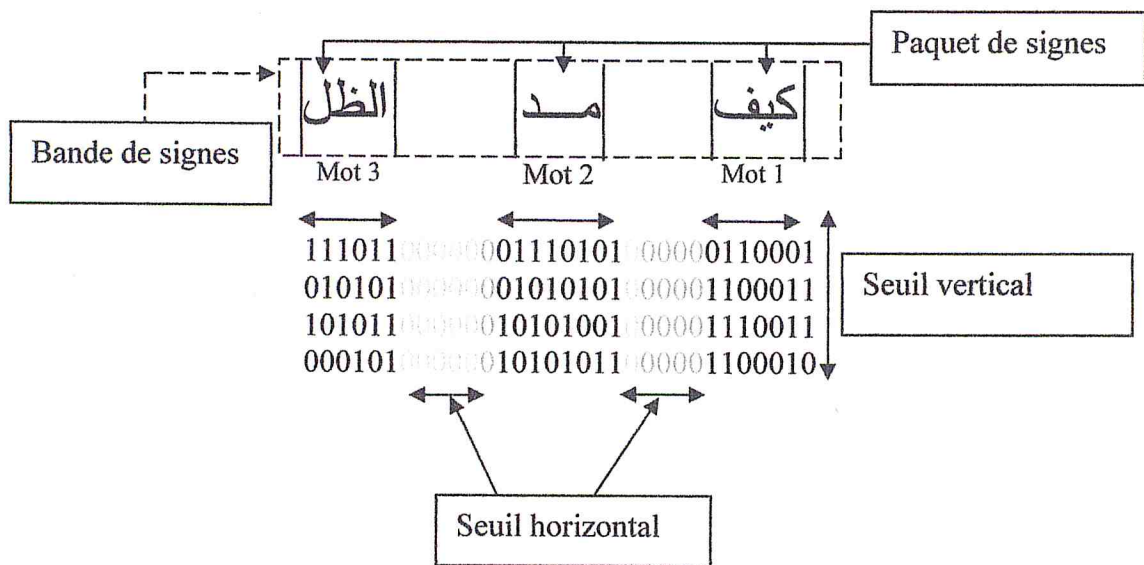


Fig.III.6. Segmentation de la ligne en mots (ou sous mots) par rectangulation

III.3.3. Segmentation du mot en caractères isolés

En ce qui concerne l'écriture arabe, plusieurs travaux ont été menés. Une approche a été développée, elle procède à une segmentation des mots en caractères isolés en se basant sur des propriétés propres à l'écriture arabe et en utilisant des histogrammes verticaux et horizontaux.

III.4. Les différentes méthodes de segmentation des mots arabes en caractères

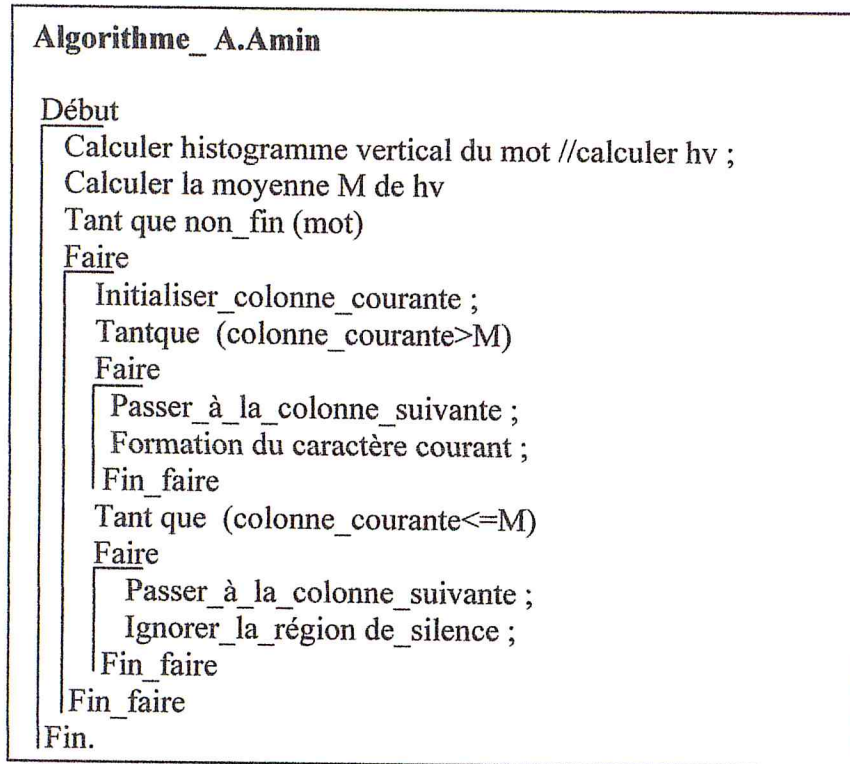
Le grand problème des méthodes de segmentation des mots arabes en caractères est la détermination des séparateurs délimitant le début et la fin de chaque caractère. Nous décrivons dans ce qui suit, quelques méthodes expérimentées permettant de détecter la liaison entre deux caractères consécutifs pour situer la fin d'un caractère et le début du caractère suivant.

III.4.1. La méthode de Adnan Amin [Ami86]

Pour distinguer les caractères du mot, Adnan Amin utilise une constatation observée au niveau de l'écriture arabe : un caractère présente, en général, un trait épais à sa naissance et un amincissement à sa fin.

Le principe de sa méthode est de détecter ces amincissements : ces derniers sont visibles au niveau de l'histogramme vertical du mot par des pics dont la valeur est inférieure à la moyennes des colonnes de l'histogramme du mot (ou du sous mot), et qui doivent être ignorés jusqu'à la rencontre d'une colonne dont la valeur est supérieure à la moyenne (début du caractère suivant).

- **Algorithme de la méthode de A.Amin :**



Algorithme 01 : Méthode de A. Amin.

- **Application de l' « Algorithme de A. Amin »**

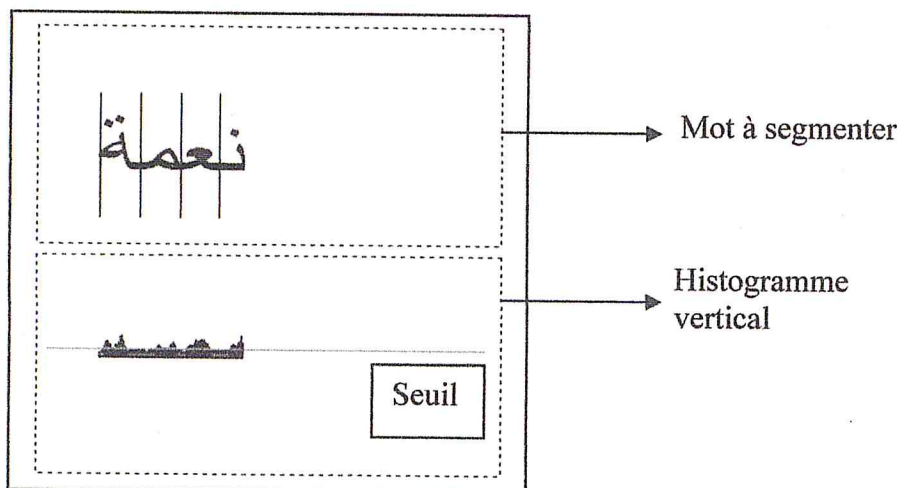


Fig.III.7. Segmentation d'un mot arabe par la méthode de Amin.

- **Avantages**
 - Simplicité de la mise en œuvre (facile à programmer).
 - Résultat satisfaisant pour les mots qui ont des caractères dont l'histogramme vertical est uniforme comme l'exemple précédent.
- **Inconvénients**
 - La détection des amincissements est basée sur une approximation grossière, qui est fonction des caractères constituant le mot, or l'étude morphologique de l'écriture arabe a démontré que ces zones sont uniformes le long d'un texte d'une même police.
 - Il n'est pas toujours vrai qu'un amincissement annonce la fin d'un caractère, certains caractères arabes prennent la forme de courbes dont l'épaisseur variable peut même chuter au dessous de la valeur moyenne obtenu pour un même caractère comme le caractère « Noune » (fig.III.8.)



Fig.III.8. La variation du tracé principale dans le même caractère.

III.4.2. La méthode de Adnan Amin améliorée [Azz 90]

Cette méthode prend en considération toutes les caractéristiques et en particulier de l'écriture arabe. L'opération de segmentation se base essentiellement sur deux points :

1. La constatation faite par A. Amin : « Le caractère arabe présente généralement une épaisseur plus importante au début qu'à la fin », cette constatation est vérifiée si on considère le spectre des caractères dans l'histogramme vertical.
2. La détection de la ligne de jonction : d'après l'étude morphologique de l'écriture arabe sur les critères de liaison des caractères entre eux. Deux résultats apparaissent :
 - La liaison entre deux caractères se fait toujours dans la région centrale de la ligne du texte, c'est le niveau de la ligne de jonction.
 - La colonne représentant les formes médianes des caractères montre que la partie du caractère servant à la liaison est faite généralement par des traits de plumes avec une

direction ou légèrement inclinée et une épaisseur constante qui représente l'épaisseur du tracé de la police.

Pour remédier les anomalies précédentes, Adnan Amin a introduit un autre type d'histogramme vertical qui représente la distance entre la médiane et le contour haut du caractère (histogramme haut). Cet histogramme permet de détecter la présence d'une partie du tracé des caractères qui sont au dessus de la ligne de jonction.

L'algorithme de segmentation devient :

Algorithme_Amin_amélioré

- 1) Calculer l'histogramme horizontal.
- 2) Détecter la ligne médiane : c'est la ligne qui a la plus forte concentration de pixels, qui est observée au niveau de l'histogramme horizontal.
- 3) Calculer l'histogramme vertical haut.
- 4) Calculer le seuil à partir de l'histogramme vertical haut qui est la valeur minimale la plus redondante.
- 5) Applique le critère de segmentation

$$\text{Hist_haut} [\text{colnne_courante}] \leq \text{seuil}.$$

Algorithme 02 : Méthode de A. Amin améliorée.

- **Inconvénients**

L'application de cette méthode, révèle des insuffisances au niveau d'une certaine classe de caractères, surtout pour les caractères situés à la fin des mots ou sous mots, ces anomalies sont dues essentiellement à la remarque suivante: « certains caractères situés à la fin des mots ou sous mots, vérifiant en quelques parties de leur tracé principal (à l'exception de la partie qui sert de jonction) le critère de segmentation (épaisseur < seuil calculé).

Les classes suivantes provoquent les anomalies :

Classe 1 : Les lettres de la famille de Bâ. (Voir annexe).

Classe 3 : Les lettres de la famille de Sin.

Classe 8 : Les autres lettres (Lam, Kaf , Mim ..).

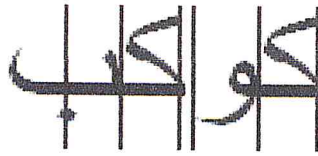


Fig.III.9. Anomalie de la méthode de A. Amin.

III.4.3. La proposition de Bouhlila [Tun 88]

Si Adnan Amin s'intéresse à distinguer la fin d'un caractère, Bouhlila essaye de détecter son début, qui se manifeste au niveau de l'histogramme vertical par une forte concentration de pixels.

Dans cette proposition, on commence par trouver la ligne de référence qui n'est autre que la ligne de jonction, cette ligne est retrouvée à partir de la valeur maximale de l'histogramme horizontal. Puis, on effectue la segmentation en caractère comme suit :

- Un histogramme vertical HV (j) est calculé dans la zone voisine à la ligne de référence ainsi que sa moyenne M.
- Une détection des colonnes est opérée conformément à l'équation $HV(j) > \text{Seuil}$ avec $\text{Seuil} = \text{Coefficient} * M$ et $\text{Coefficient} = 1.5$, généralement déterminé par des expérimentations.

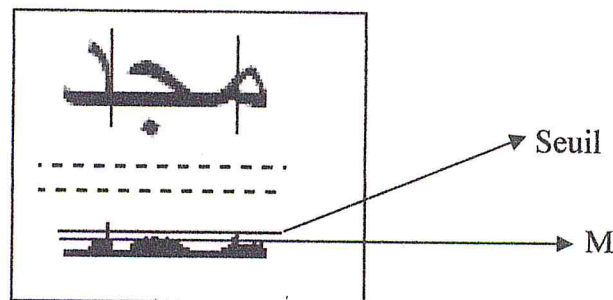


Fig.III.10. Segmentation d'un mot arabe par la proposition de Bouhlila.

- **Algorithme de la proposition de Bouhlila :**

```

Procédure segmentation_Bouhlila
Début
Calculer l'histogramme vertical du mot //calculer hv ;
Calculer la moyenne M de hv
Tant que non_fin (mot)
Faire
  Initialiser_colonne_courante ;
  Tant que (colonne_courante > M * coefficient)
  Faire
    Passer_à_la_colonne_suivante ;
    {Formation du caractère courant};
  Fin_faire
  Tant que (colonne_courante <= M * coefficient)
  Faire
    Passer_à_la_colonne_suivante ;
    {Ignorer la région de silence};
  Fin_faire
Fin_faire
Fin.

```

Algorithme 03 : Proposition de Bouhlila.

- **Avantages et Inconvénients**

Cette proposition est facile à implémenter et donne des résultats satisfaisants. Cependant elle nécessite l'ajustement de la valeur d'un coefficient en fonction de la police de caractères afin d'augmenter le taux de segmentation. Cette lacune ne lui permet pas en fait de s'adapter à plusieurs polices sauf si on détermine le coefficient expérimentalement pour chaque police de caractères et mettre un choix de fonte dans le système de reconnaissance. Ceci n'est pas commode dans un système omni-fonte qui doit détecter la fonte automatiquement. Dans cette proposition peut être recommandée pour le cas mono-fonte mais pas pour un système de reconnaissance multi-fontes.

III.4.4. La Méthode Over-segmentation [Gil 99]

Dans la méthode over-segmentation, on segmente le mot en plusieurs images 'atomiques' qui ne doivent pas avoir plus que la taille d'un caractère. En d'autres termes, chaque segment atomique doit appartenir à un seul caractère dans le cas idéal. Si le but est atteint, des combinaisons sont effectuées entre les différents segments (on combine deux à cinq segments

ensemble) en maintenant les relations d'espacement entre les segments atomiques du même groupe. On obtient au final $k < 5 * N$ segments, où N représente le nombre de segments atomiques. Dans le cas idéal, k est le nombre de caractères du mot. En passant les différentes combinaisons par un algorithme de reconnaissance de caractères, on obtient les caractères appropriés.

Les segments atomiques sont obtenus en utilisant deux méthodes heuristiques :

La première consiste à appliquer une fonction $f(x)$ de minimum local qui est décrite comme suit :

$$f(x) = \text{MAX}(B - p\text{Haut}(x), 0) + \text{MAX}(p\text{Bas}(x) - B, 0).$$

$p\text{Haut}(x)$ = coordonnée y du plus haut pixel de la colonne de x .

$p\text{Bas}(x)$ = coordonnée y du plus bas pixel de la colonne de x .

B = coordonnée y de la ligne de jonction (médiane).

La seconde méthode recherche un minimum local dans la partie supérieure du contour autour du composant. Cela permet de retrouver de nouveaux segments que la première méthode ne les trouve pas. Cependant, les deux méthodes retrouvent presque les mêmes points de segmentation. Pour y remédier, on enlève les points redondants.

Le critère de segmentation minimale est fixé à 4 pixels de largeur et à une matrice (segment image) contenant plus de 20 pixels en tout.

La figure suivante montre cette technique appliquée à un mot donné :

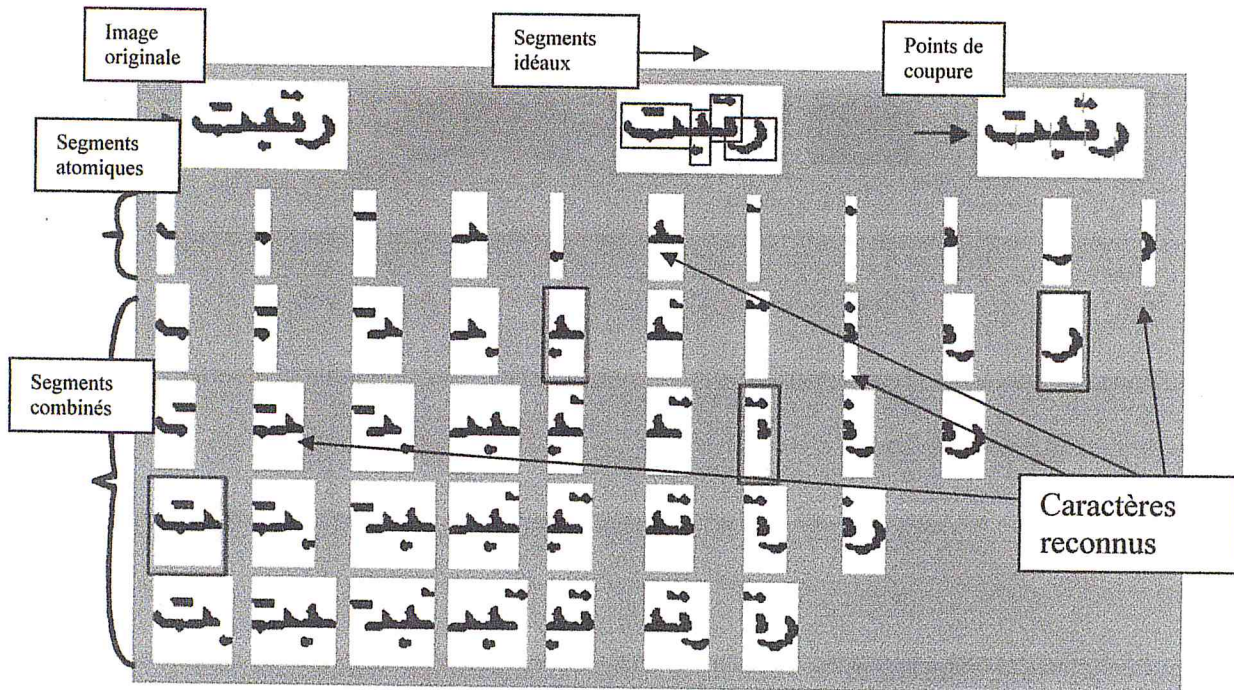


Fig.III.11. Segmentation d'un mot en caractère par la méthode Over-segmentation.

o **Avantage et inconvénients**

Le principal avantage de cette méthode est qu'elle peut être utilisée sur n'importe quelle fonte. Cependant, sa mise en œuvre est fastidieuse et dépend de l'algorithme de la reconnaissance qui doit être robuste.

III.4.5. La proposition de Hamami & Berkani [Ham 02]

Pour extraire les caractères du mot ou du sous mot, cette proposition a besoin de détecter :

- a- **La ligne de jonction** : Elle est représentée par la région qui a le plus grand nombre de pixels.
- b- **Les contours supérieurs du mot (ou sous mot)** : Par un balayage vertical, en descendant du haut vers le bas, les premiers pixels noirs rencontrés sont considérés comme le contour du mot ou du sous mot.
- c- **Le Contour inférieur du mot (ou sous mot)** : Par un balayage vertical, en montant du bas vers le haut, les premiers pixels noirs rencontrés sont considérés comme le contour bas du mot.
- d- **L'histogramme vertical du mot (ou sous mot)** : À ce stade, nous déterminons :
 - **Le seuil** qui est remarquable par la zone de la valeur stable, et correspond à la valeur la plus redondante dans l'histogramme vertical.
 - **Le nombre de transitions** verticales (0-1) ou (1-0) (blanc_noir ou noir_blanc) pour chaque colonne du mot.

ربنا إفتح بيننا و بين قومنا بالحق و أنت خير الفاتحين

Fig.III.12. Segmentation en caractères.

○ **La segmentation d'un caractère :**

Dans cette approche, quelques conditions sont introduites, elles permettent de détecter le début et la fin d'un caractère.

La colonne qui correspond au début d'un caractère est noté colonne initiale, si elle vérifie la condition suivante (condition (a)).

a- L'histogramme vertical de cette colonne doit être supérieur au seuil.

La dernière colonne d'un caractère correspond au début du caractère suivant, si elle vérifie les conditions suivantes (b).

b-

1- Le pixel du contour supérieur dans cette colonne doit être plus supérieur ou égal à la ligne de jonction.

2- Le pixel du contour du bas doit être inférieur ou égale à la ligne de jonction.

3- La différence entre le pixel du contour inférieur et le pixel du contour supérieur dans cette colonne doit être inférieure ou égale au seuil.

4- L'histogramme vertical de cette colonne doit être inférieur ou égale au seuil.

5- Le nombre de transitions verticales doit être égal à deux.

6- Le pixel du contour supérieur de cette colonne doit être au-dessus de la ligne du sommet de la colonne initiale.

L'algorithme suivant permet l'extraction des caractères d'un mot après avoir déterminé ses premières et dernières colonnes.

Algorithme Hamami_Berkani

Deb: Booléen variable (deb = vrai correspond au début du caractère)

Début

Deb := vrai;

Pour chaque mot

Faire

Tant que non fin de la partie connectée

Faire

Si deb est vrai alors

- Rechercher la première colonne qui vérifie la condition (a).
- Sauvegarder la colonne initial, dénoté j_1 ;
- Deb := faux.

Sinon

- $j_1 = j_2 + 1$;
- Aller à la prochaine colonne.
- Rechercher la colonne du début du prochain caractère, dénoté j'_1 ;

Si il existe alors

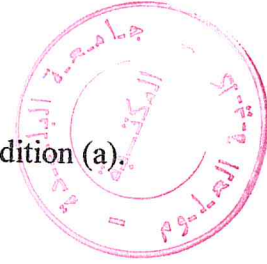
- Effectuer un balayage vertical de droite à gauche (de j'_1 à j_1)
- Déterminer la première colonne qui vérifie des conditions (b), dénoté j_2 .

Sinon Dans ce cas, nous avons atteint la fin de la partie connectée, dénoté j_2 ;

- j_1 et le j_2 représente respectivement la première et la dernière colonne du caractère.

Fait

Fin.

**Algorithme 04 : Proposition de Hamami & Berkani.**

○ **Avantage et inconvénient de la proposition de Hamami & Berkani**

Cette proposition donne de bons résultats pour tous les caractères non qui situés à la fin du mot. Par conséquent, il y a certains caractères situés à la fin du mot qui provoquent des fausses segmentations comme les caractères de la classe 1 et 3. [Voir Annexe].



Fig.III.13 Anomalie de la proposition.

III.4.6. La méthode de Adnan Amin et Humoud B. Al-Sadoun [Hum 92]

Dans cette technique, l'image originale du mot est prétraitée pour produire "le squelette de l'image" en utilisant l'algorithme d'amincissement parallèle [Guo 89] comme le montre la figure III.14.



Fig.III.14. (a) Image originale, (b) Squelette du mot.

L'objectif de cette méthode est de construire un arbre binaire contenant toutes les informations décrivant la structure de l'image. La technique utilisée nécessite après découpage, le codage de l'image amincie en commençant de la droite vers la gauche par l'utilisation d'une matrice 3x3 et en enregistrant la structure des parties tracés (code de Freeman [Fre 68])

Pour un voisinage 3x3, le code de Freeman est décrit comme suite :

| | | |
|---|-------|---|
| 3 | 2 | 1 |
| 4 | tracé | 0 |
| 5 | 6 | 7 |

Fig.III.15. Code de Freeman pour un voisinage 3x3.

L'exemple suivant illustre le calcul du code de Freeman pour un sous mot donné.

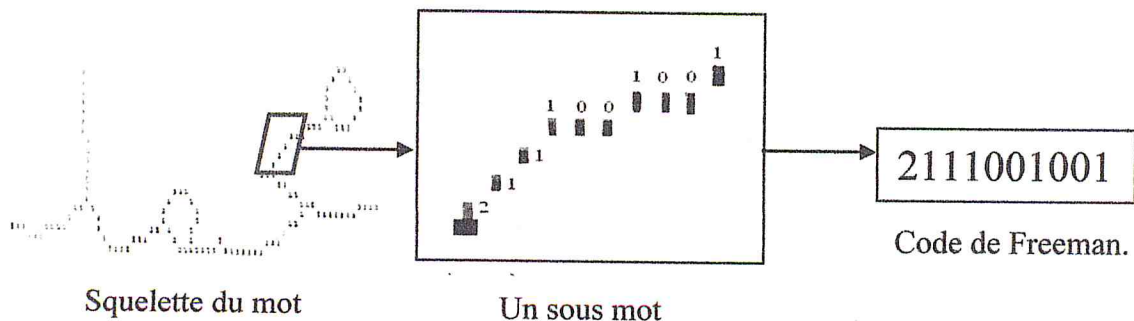


Fig.III.16. Exemple de calcul du code de Freeman.

- Dans le cas où un caractère présente une boucle (ا) ou une double boucle (ـ) dans le mot, on lui attribue en plus du code de Freeman, deux primitives supplémentaires qui sont : L (boucle) et LL (double boucle).



Fig.III.17. Caractères avec boucles et leurs primitives.

L'arbre binaire est constitué de plusieurs nœuds, chaque nœud décrit la forme du segment correspondant du sous-mot. La structure du nœud dans l'arbre binaire contient les champs suivants :

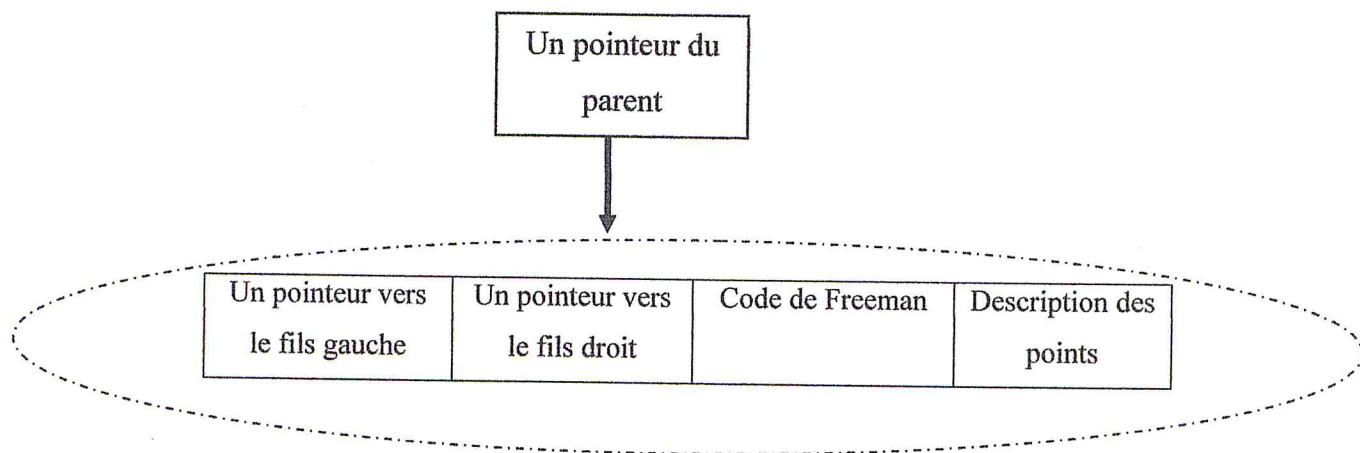


Fig.III.18. Structure d'un nœud.

Ainsi, l'arbre binaire de la figure III.14. est le suivant

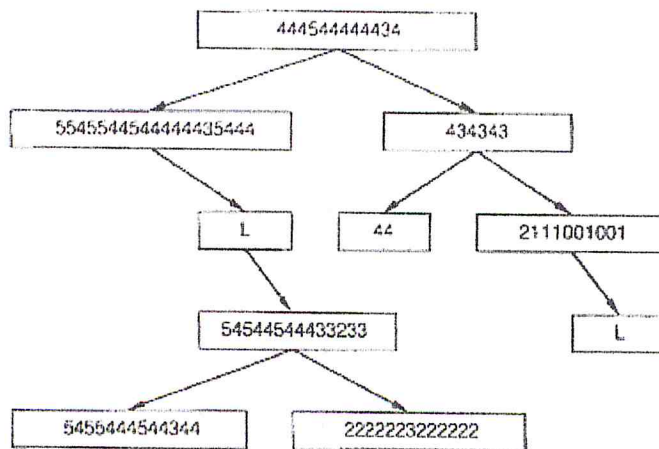


Fig.III.19. L'arbre binaire de l'image de la figure III.14.

sera examiné pour s'assurer si ce noeud est le début du prochain caractère ou la continuation du caractère courant. Les règles suivantes permettent d'assurer la fin d'un caractère courant :

- a. Le premier noeud consulté est le début du premier caractère.
- b. Un long segment horizontal indique la fin du caractère courant.
- c. L ou LL indique le début d'un nouveau caractère, et le noeud précédent indique la fin du caractère précédent.
- d. Un long segment vertical indique le début d'un nouveau caractère et le noeud précédant indique la fin du caractère précédent.
- e. Un petit segment vertical, en plus de quelques points indique le début d'un nouveau caractère et le noeud précédent indique la fin du caractère précédent.
- f. Un petit segment vertical sans points peut être considéré comme le début d'un nouveau caractère et le noeud précédent indique la fin du caractère précédent.

L'application de ces règles sur le mot de l'image de la figure III.14 donne le résultat représenté à la figure III.21 :

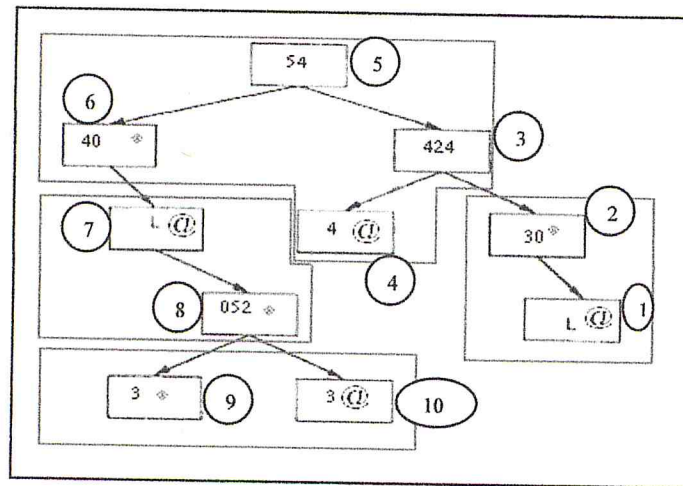


Fig.III.21. Segmentation de l'arbre binaire.

Le mot contient quatre caractères.

-Le noeud qui contient le symbole @, est considérée comme le début d'un caractère.

-le noeud qui contient le symbole * est considérée comme la fin d'un caractère.

Par conséquent, les quatre caractères du mot sont décrits dans les noeuds suivants :

- 1^{er} caractère: dans les noeuds 1, 2.
- 2^{ème} caractère: dans les noeuds 3, 4, 5, 6.
- 3^{ème} caractère: dans les noeuds 7, 8.
- 4^{ème} caractère: dans les noeuds 9, 10.

Après avoir créé l'arbre binaire selon les règles précitées. On construit l'arbre binaire lissé, le lissage de l'arbre binaire est conçu comme suit

- a. Minimiser le nombre des noeuds de l'arbre.
- b. Minimiser les codes de Freeman dans le champ d'information des noeuds.
- c. Éliminer ou minimiser tout bruit dans l'image amincie.

L'arbre binaire lissé de l'image de la figure III.14 est représenté dans la figure III.20.

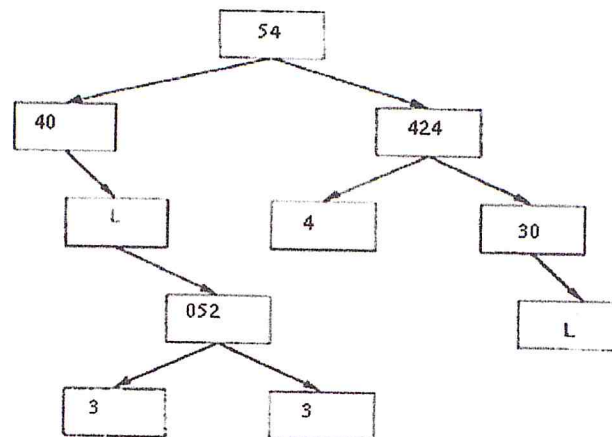


Fig.III.20. Arbre binaire lissé.

La segmentation nécessite le processus de division de l'arbre binaire en plusieurs sous-arbres tel que chaque sous-arbre représente un caractère. L'identification de la ligne de jonction n'est pas importante vu que le sous-mot est décrit par un arbre binaire.

Le processus de la segmentation est résumé dans les points suivants:

- a. L'arbre binaire peut être parcouru dans un ordre tel que le premier noeud traversé dans le processus de segmentation soit le début du premier caractère, ce noeud n'est pas nécessairement la racine.
- b. La description du caractère peut être étendue dans un ou plusieurs champs d'information des noeuds, néanmoins, ces noeuds sont placés l'un après l'autre selon la forme du caractère. D'où, le noeud qui est parcouru en premier correspond au début du caractère, et le noeud qui est parcouru en dernier correspond à la fin du caractère.
- c. Quelques caractères qui sont non connectés de la gauche peuvent être contenus dans un noeud en plus de la fin du caractère précédent. Dans ce cas, le noeud doit être le dernier noeud à traverser dans le sous-mot.

Le parcours de l'arbre permet d'identifier la fin du caractère en cours ou le début du prochain caractère. Chaque fois qu'un noeud est visité, le champ de l'information de ce noeud

Chapitre IV

Etude conceptuelle

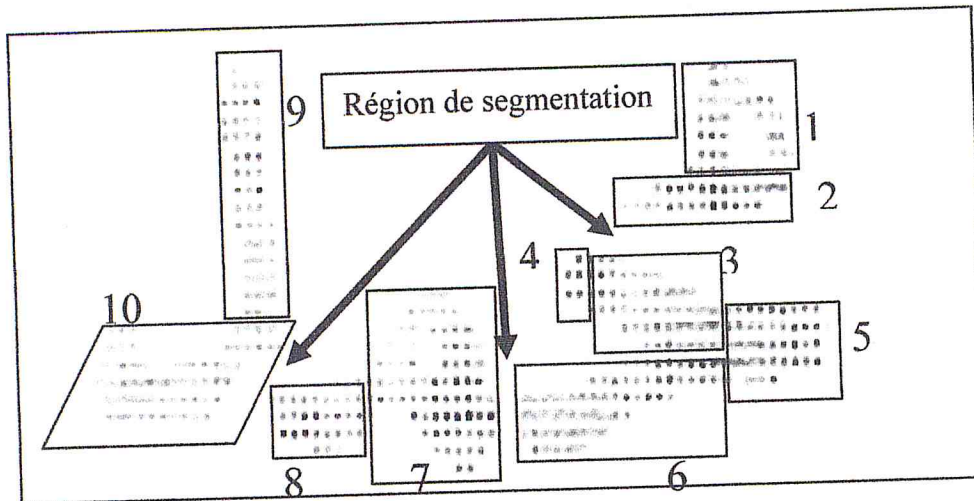


Fig.III.22. Segmentation de mot de la figure III.14.

○ **Avantages et inconvénients**

- **Avantage :**

- Résultats satisfaisants pour les mots qui ont des caractères ligaturés.

- **Inconvénients :**

- La description du caractère peut être étendue dans un ou plusieurs nœuds (sous arbre), et chaque nœud est représenté par cinq champs d'informations, ainsi il exige plus d'espace mémoire.
- La mise en œuvre de cette méthode est fastidieuse car elle applique la segmentation après la réalisation des étapes suivantes :
 - Construction de l'arbre binaire et l'arbre binaire laissé.
 - Construction de l'image squelette de l'image originale par l'algorithme d'amincissement parallèle.
 - Calcul du code de Freeman pour chaque sous-mot.

III.5. Tableau comparatif

Ce tableau comparatif, vient comme synthèse à tous ce qu'on a présenté.

| Méthode | Outils | Avantages | Inconvénients |
|-------------------------------|------------------------|---|---|
| Adnan Amin (Algorithme 01) | - Histogramme vertical | - Simple à programmer. - Bon résultat lorsque les caractères sont uniformes. | - Fausse segmentation pour les caractères sous forme de courbes (ن, و). |

| | | | |
|--------------------------------------|--|--|--|
| Adnan Amin améliorée (Algorithme 02) | - Histogrammes : vertical et horizontal | - Résoudre le problème des caractères qui se situent au dessous de la ligne de jonction. | - Fausse segmentation pour les caractères située à la fin des mots ou sous mots. |
| Bouhlila (Algorithme 03) | - Histogramme vertical | - Simple à programmer. | - Le seuil de segmentation est expérimental. |
| Over-segmentation | - Segments atomiques | - Résultat satisfaisant. | - Base de connaissance contient toutes les fontes de caractères. |
| Hamami & Berkani (Algorithme 04) | - Histogrammes : vertical et horizontal - Contours | - peu de structure de données. | - Fausse segmentation pour les caractères des classes (3, 1). (voir annexe_a) |
| Adnan Amin et Humoud B. Al-Sadoun | - Arbres binaires - Code de Freeman | - Bonne segmentation pour les caractères ligaturés. | - Complexe à programmer. |

Tableau 1 : Tableau comparatif.

III.6. Conclusion

Après avoir détaillé les différentes techniques de segmentation de mots arabes en caractères, on a remarqué qu'il n'existe pas de méthode parfaite qui permet de donner une bonne segmentation. Chacune d'entre-elles possède des avantages et des inconvénients. Cependant, la méthode d'Over-segmentation s'avère très robuste et donne de très bons résultats mais sa mise en œuvre se base essentiellement sur un système de reconnaissance de caractères. Quant à la méthode de Adnan Amin et Humoud B. Al-Sadoun donne des résultats satisfaisants pour les caractères qui sont ligaturés, mais très complexes à programmer.

Dans le prochain chapitre, nous présenterons une étude conceptuelle sur le processus de la segmentation que nous avons développée afin de bien cerner les limites de la méthode de A. Amin

IV.1. Introduction

Nous avons vu dans le chapitre précédent plusieurs techniques de segmentation de mots arabes en caractères. Parmi les quelles, la méthode développée par A. Amin (Algorithme N°1) qui donne une mauvaise segmentation pour certaines classes des caractères et la méthode améliorée par A. Amin (Algorithme N°2).

Notre travail consiste à réaliser un module de segmentation des mots arabes en caractères capable de remédier aux anomalies de la méthode de A. Amin améliorée (Algorithme N°2).

IV.2. Processus de segmentation

Le processus de segmentation des mots arabes en caractères que nous proposons est présenté dans la figure IV.1.

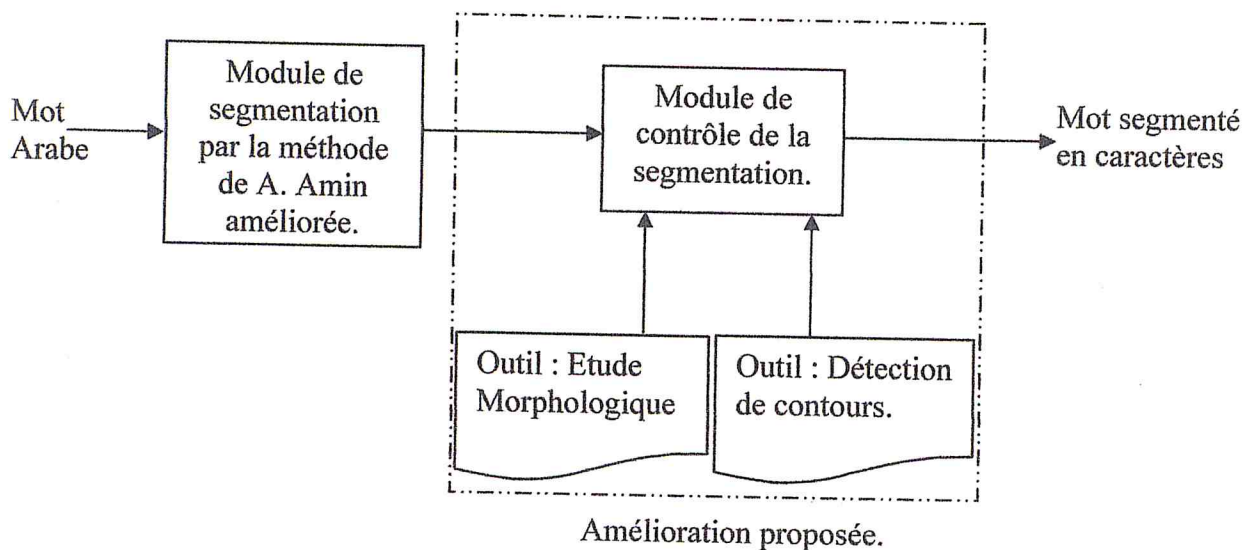


Fig.IV.1. Processus de la segmentation.

Ce processus intègre la méthode de segmentation améliorée de A. Amin, ainsi qu'un niveau de contrôle de la segmentation qui permet de pallier aux inconvénients de la deuxième méthode.

Il utilise différents outils tels que l'étude morphologique et la détection de contours

IV.2.1. Segmentation par la méthode de A. Amin améliorée

IV.2.1.1. Calcul des histogrammes

Le calcul des histogrammes peut se faire en une seule passe sur tous les pixels de l'image, l'algorithme fait comptabiliser le nombre de pixels allumé verticalement et horizontalement.

Rappelons que :

- Histogramme horizontal (HH) est définie de la manière suivante :

$$HH(i) = \sum_j \sum_i M(i, j)$$

- Histogramme vertical (HV) est définie comme suit :

$$HV(j) = \sum_i \sum_j M(j, i)$$

Où :

M : désigne l'image numérisée.

i correspond à l'indice des lignes et j correspond à l'indice de colonnes .

Les algorithmes de calcul des histogrammes sont :

Var histH Table[0..longueur] init 0; // représente l'histogramme horizontal

Var histV Table[0..largeur] init 0 ; // représente l'histogramme vertical.

Var histVH Table[0..largeur] init 0 ; // représente l'histogramme vertical haut.

Histogramme_horizontal ()

Début

Pour i=0 to largeur

Pour j=0 to longueur

Faire Si (P(i,j)allumé) alors histH[j]++ ;

Fin.

Histogramme_vertical ()

Début

Pour j allant de 0 à largeur

Pour i allant de 0 à longueur

Faire Si (P(i,j)allumé) alors histV[i]++ ;

Fin.

Algorithme 05 : Histogramme horizontal.

Algorithme 06 : Histogramme vertical.

Histogramme_vertical_Haut (ligne_jonction : entier)

Début

Pour i allant de 0 à largeur

Pour j allant de 0 à ligne_jonction

Faire histVH[i]= histV[i] ;

Fin.

Algorithme 07: Histogramme vertical Haut.

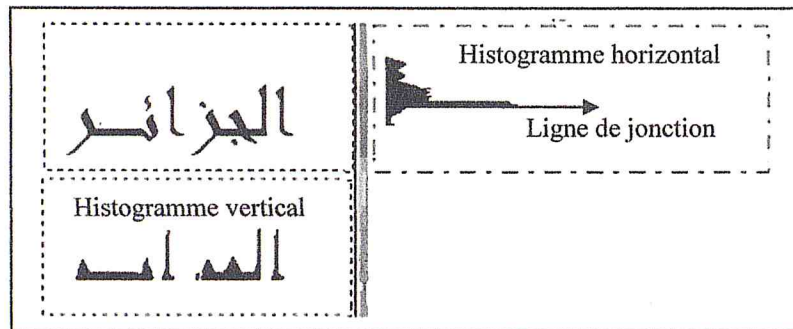


Fig.IV.2. Type des histogrammes.

IV.2.1.2. Détection de la ligne de jonction

La liaison entre deux caractères se fait toujours dans la région centrale de la ligne du texte, c'est le niveau de la ligne de jonction. Cette dernière est distinguée par la plus grande concentration des pixels dans l'histogramme horizontal.

Ligne_de_jonction()

Début

Var max ; Var ligne_jonction ;

Pour i allant de 0 à longueur

Si (max < histH[i]) alors

Début

max= histH[i];

k=i ;

Fin.

ligne_jonction =k ;//la ligne de jonction est définie par la fonction y=k ;

Fin.

Algorithme 08 : Ligne de jonction.

IV.2.1.3. Calcul du seuil de segmentation

La méthode de A. Amin se base sur les histogrammes horizontal et vertical haut qui ont été déjà calculés au préalable. A partir de l'histogramme vertical haut, le seuil correspond à la valeur minimale la plus redondante.

```

Seuil_de_segmentation ()
Début
  Var R[0.. longueur] init 0 ; Var max=0 ;
  //----calculer toutes les redondances
  Pour i allant de 0 à longueur
    Faire
      j=i +1;
      Tant que (j< longueur)
        Début
          Tant que (histVH [i]= histVH [j]) j++;
          R[i]++;
        Fin.
      Fin_faire
  //----calculer la valeur minimale la plus redondante
  Pour i allant de 0 à longueur
    Si (max< R[i]) alors
      Début
        max= R[i] ;
        k=i ;
      Fin.
  Seuil= histVH[k] ;
Fin.

```

Algorithme 09 : Seuil de segmentation.

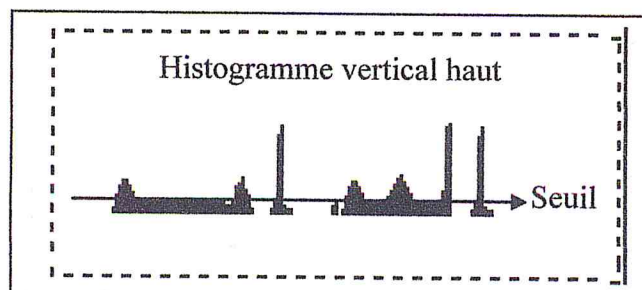


Fig.IV.3. Seuil de segmentation.

IV.2.1.4. Critère de segmentation

Une fois que le seuil est calculé, la segmentation est effectuée en utilisant le critère suivant :

$$\text{Hist_vertical_haut}[\text{colonne_courante}] \leq \text{seuil.}$$

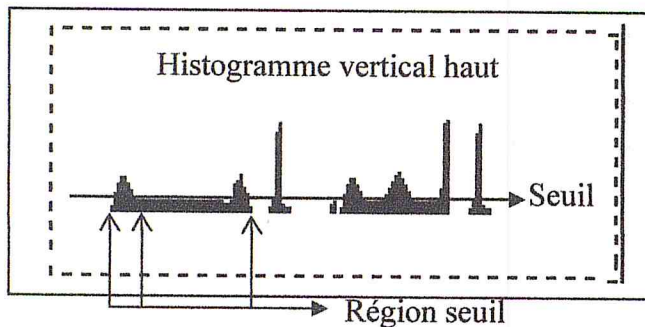


Fig.V1.4. Régions de segmentation

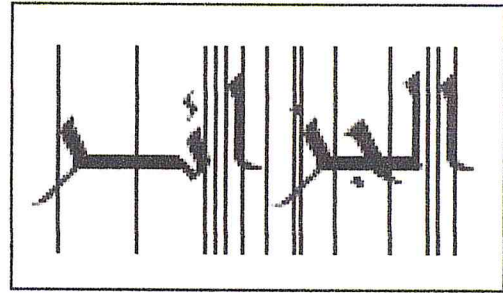


Fig.V1.5. Application du critère de segmentation.

La figure IV.5 montre que la segmentation n'a pas été réalisée correctement notamment :

- La segmentation de du caractère ج en trois segments.
- La segmentation de du caractère ل en deux segments.
- Le mal segmentation du caractère ر.

Cet exemple illustre clairement les anomalies que peut engendrer la méthode améliorée de A. Amin dans la segmentation des caractères de différentes classes. Ces anomalies seront corrigées par le contrôle de segmentation.

IV.2.2. Contrôle de segmentation

Cette phase permet de détecter les caractères mal segmentés qui sont :

IV.2.2.1. Les caractères commençant ou se terminant par des traits amincis

Les traits amincis sont observables au niveau de l'histogramme vertical haut par des régions inférieures ou égale au seuil, le processus de contrôle fait parcourir cette région.

Parmi les caractères qui possèdent des traits amincis au début ou à la fin nous avons:

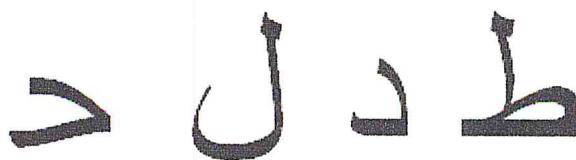


Fig.IV.6. Caractères commençant ou se terminant par un trait aminci.

L'algorithme de contrôle devient :

Procédure Amelio_proposé_trait_aminici

Début

- Initialiser le début du mot ou du sous mot, dénoté k_i .

Si $(HV [k_i] \leq \text{seuil})$ **Alors** éliminer la segmentation ;

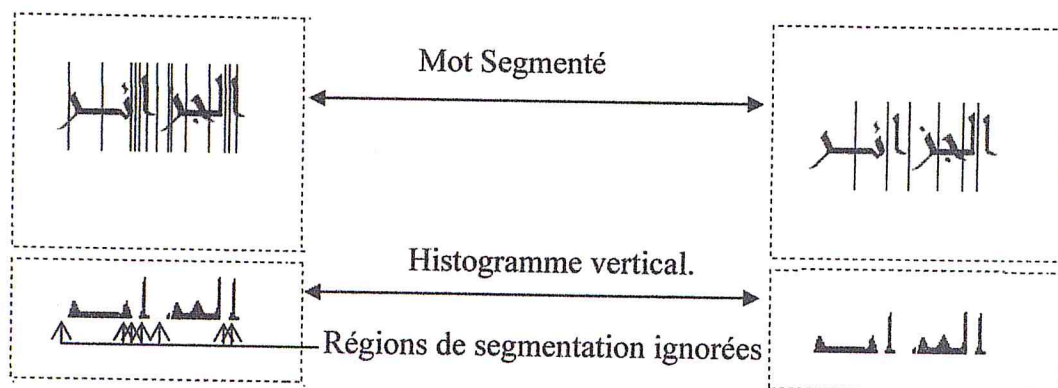
Tant que (non fin du mot) **faire** Aller la région suivant, k_j .

Si $(HV [k_j] \leq \text{seuil})$ **Alors** éliminer la segmentation ;

Fin.

Algorithme 10 : Contrôle les caractères commencent ou terminent par un trait aminci.

➤ **Application**



(a) Segmentation par A. Amin.

(b) Segmentation par l'amélioration proposée.

Fig.IV.7. Lettres possédants des traits amincis au début ou à la fin des caractères.

La figure **Fig.IV.7. (a)** montre que les caractères qui possèdent des traits amincis observable au niveau de l'histogramme vertical en position de début ou à la fin de mot ou sous mot ont été correctement segmentés alors que ceux de la figure **Fig.IV.7. (b)** ont été anormalement segmentés.

IV.2.2.2. Les lettres possédant des points

Les lettres qui possèdent des points et se situent à la fin du mot sont généralement des lettres des classes 1 et 4 (voir annexe), qui sont :

Classe 1 : ب، ت، ث، ي

Classe 4 : ق، ف

L'étude morphologique de l'écriture arabe sur les lettres des classes 1 et 4 en position de fin du mot, permet de déterminer quelques caractéristiques et qui sont :

- Elles contiennent une dent à la fin.
- On notera la présence de groupe de points et leurs positions jouent un rôle prépondérant pour la discrimination des caractères appartenant à une même famille de forme. Ces groupes de points (1 à 3 points) peuvent occuper deux positions possibles relativement à la forme : au dessus, ou en dessous de la ligne de jonction.

La méthode de A. Amin améliorée segmente les lettres de ces classes comme suit :

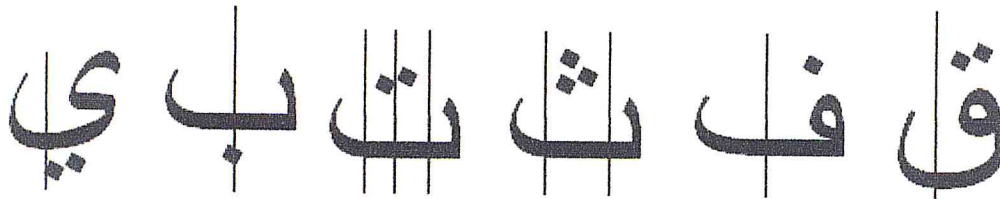


Fig. IV.8. Anomalie de la segmentation par la méthode de A. Amin Améliorée.

Analyse :

- Les lettres des classes précédentes sont caractérisées par des points communs qui sont :
 - a. Elles possèdent une dent à la fin.
 - b. Leur représentation par l'histogramme horizontal permet d'obtenir deux ensembles de rassemblement des pixels noirs séparés par une zone silencieuse, l'un des ces ensemble représente le corps de lettre, et l'autre représente le point (ou les point).

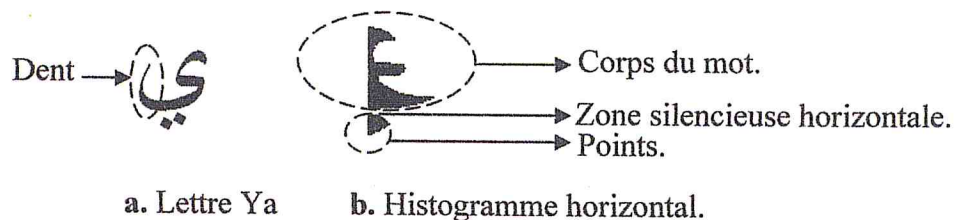


Fig. IV.9. Exemple des caractéristiques des lettres des classes 1 et 4.

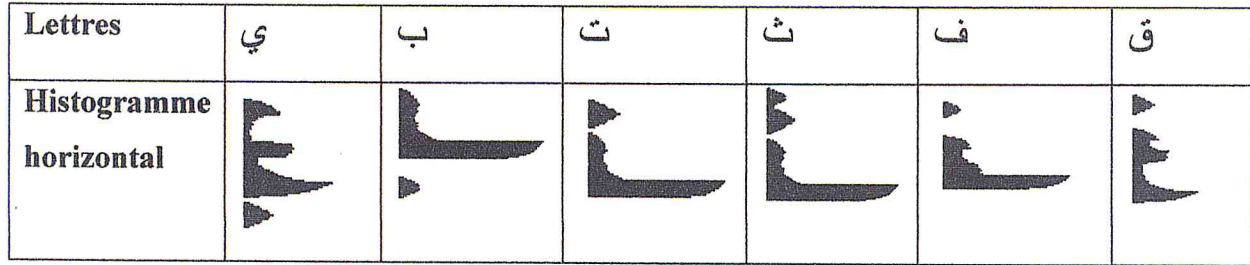


Fig. IV.10. Histogramme horizontal des lettres des classes 1 et 4.

Pour éviter cette anomalie, il suffit de réaliser les deux conditions suivantes:

- 1 La forme de la fin du mot est une forme de dent (ni forme de cercle comme (هـ), ni forme de segment vertical (ل)...).
2. Il existe une zone silencieuse dans l'histogramme horizontal

La procédure de contrôle devient :

Procédure Ameli_proposé_classe_1,4

Début

- Aller à la fin du mot.
- Déterminer la fin de la lettre considérée.
- Déterminer le début de cette lettre.
- Calculer l'histogramme horizontal au niveau de cette lettre.

Si ((la fin de la lettre considérée est une dent) && (\exists zone silencieuse horizontal))

Alors éliminer les segmentations autour de cette lettre.

Fin.

Algorithme 11 : Contrôle les lettres des classes 1 et 4.

Détermination du début d'une lettre

Pour déterminer le début d'une lettre, nous calculons l'intersection du contour de la lettre avec la droite (Δ) dont la fonction est $y=L-h/2$, tel que y est la coordonnée de la ligne de jonction et h est la hauteur de la dent (de la fin) d'une lettre. Si le nombre de point d'intersection de cette ligne ($y=L-h/2$) avec le contour de la lettre est égale à 4, ainsi le début de la lettre est détectée.

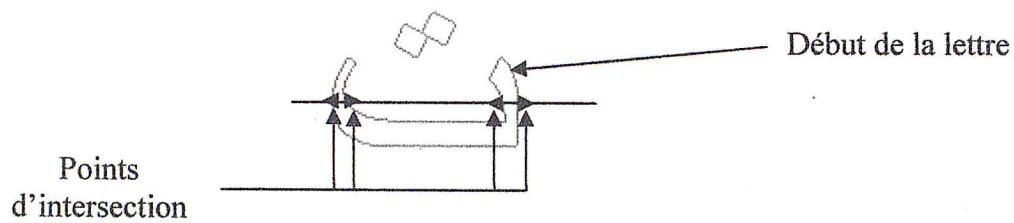


Fig.IV.11. Détermination du début d'une lettre.

Le calcul du contour du mot doit calculer les contours hauts, les contours bas, les contours gauches et les contours droits et, se fait comme suite :

➤ Calcul des contours gauche et droit

- Parcourir les lignes de l'image de la gauche vers la droite, les pixels noirs qui se situent directement après les pixels blancs sont considérés comme un contour gauche, et ceux qui se situent directement avant les pixels blancs sont considérés comme un contour droit.

➤ Calcul des contours haut et bas

- Parcourir les colonnes de l'image de haut vers le bas, les pixels noirs qui se situent directement après les pixels blancs sont considérés comme un contours haut, et ceux qui se situent directement avant les pixels blancs sont considérés comme un contour bas.

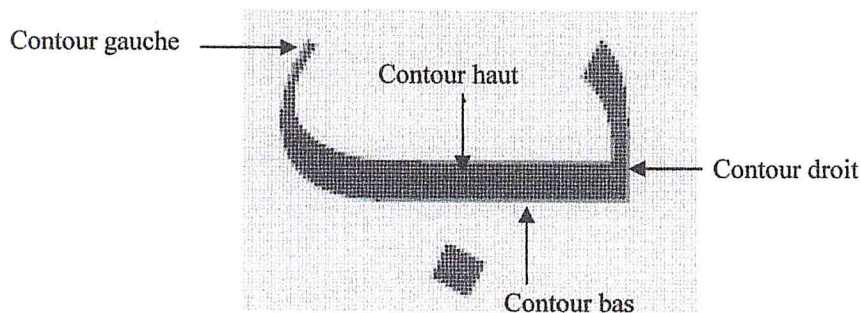


Fig.IV.12. Contour du mot.

➤ Application

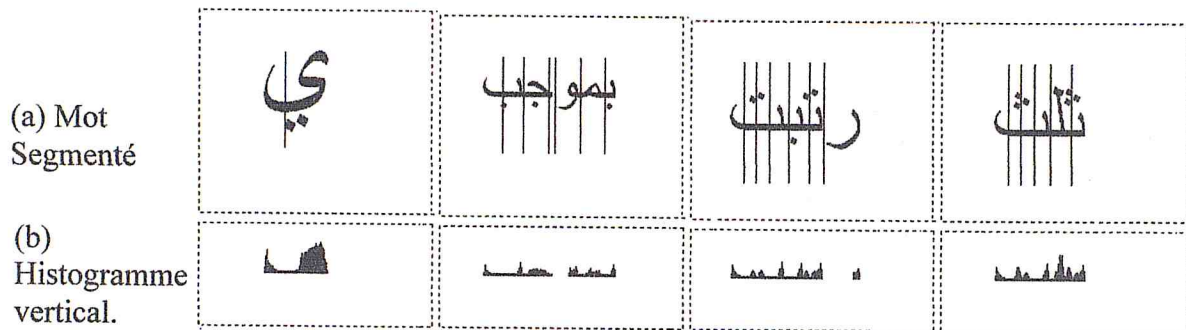


Fig.IV.13. Segmentation des lettres de la classe 1 par A. Amin améliorée (Application de l’algorithme 02).

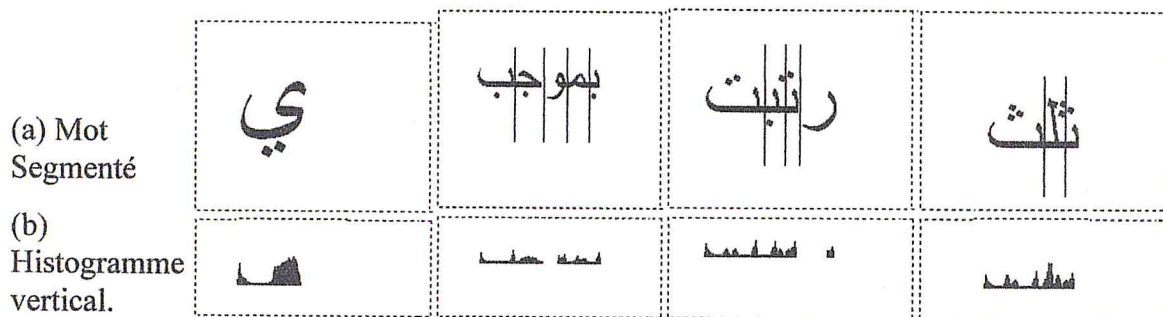


Fig.IV.14. Lettres des classes 1 contrôlées (Application de l’algorithme 11).

La figure Fig.IV.14 montre que les caractères a points et situés a fin du mot ou sous mot ont été correctement segmentés alors que ceux de la figure Fig.IV.13. (b) ont été anormalement segmentés.

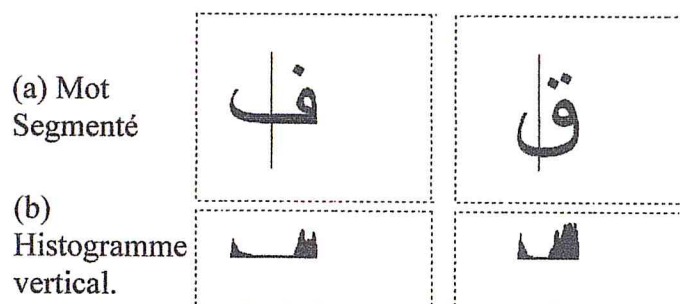


Fig.IV.15. Segmentation des lettres de la classe 4 par A. Amin améliorée. (Application de l’algorithme 2).

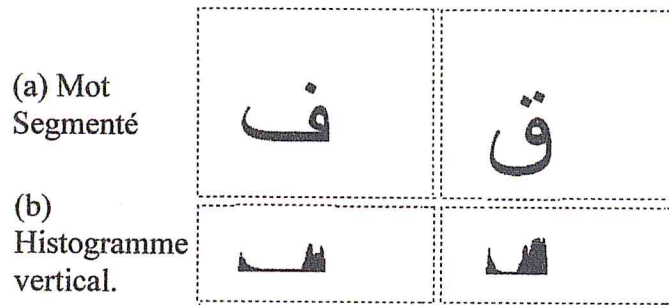


Fig.IV.16. Lettres des classes 4 contrôlées (application de l'algorithme 11).

La figure Fig.IV.16 montre que les caractères ف et ق ont été correctement segmentés alors que ceux de la figure Fig.IV.15 ont été anormalement segmentés.

IV.2.2.3. La lettre 'Sin'

La méthode de A. Amin segmente le caractère "sin" en trois parties comme c'est illustré dans la fig.IV.17.

Une solution à ce problème, basée sur l'observation du caractère 'Sin', nous permet de déterminer quelques caractéristiques topologiques de ce caractère.

Ce caractère est identifié par :

- Trois dents identiques observables au niveau l'histogramme vertical.
- Histogramme horizontal est un ensemble des rassemblements des pixels noirs non séparés par des zones silencieuses.

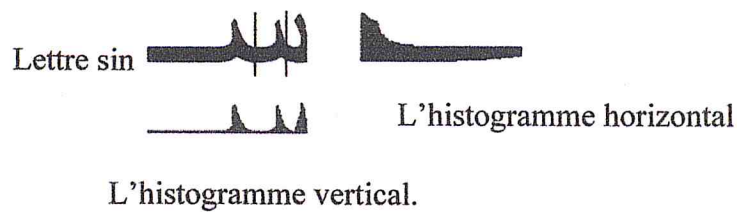
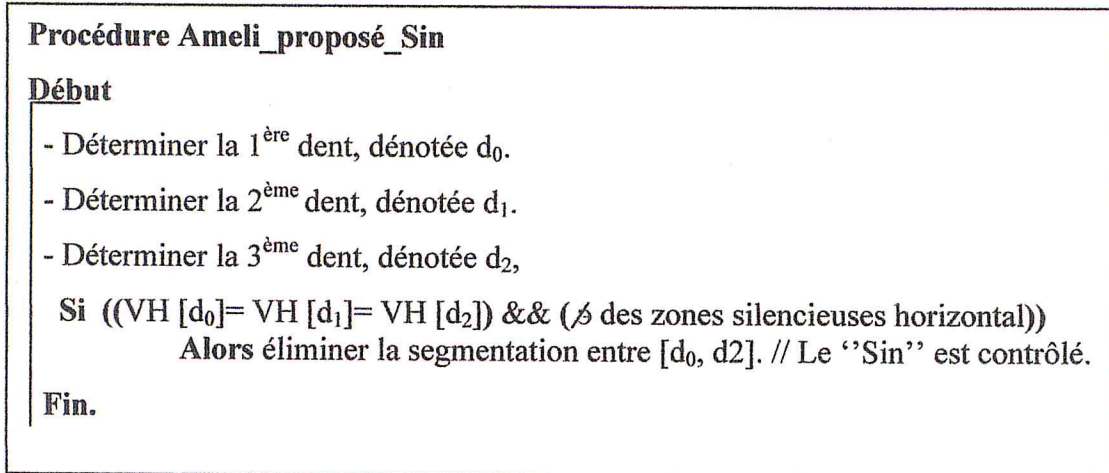


Fig.IV.17. Caractéristiques de la lettre " Sin".

Par conséquent, nous proposons l'algorithme de la segmentation suit :



Algorithme N° 12 : Contrôle la lettre sin.

➤ Application

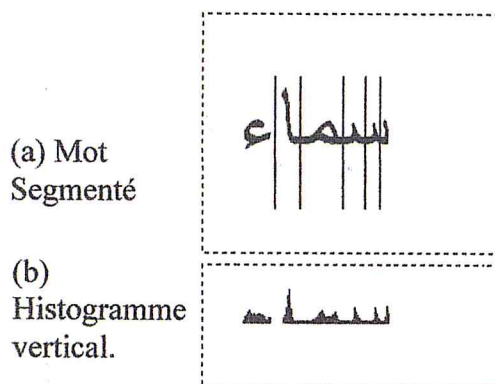


Fig.IV.18. Segmentation par A. Amin.

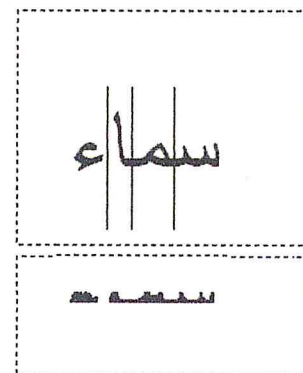


Fig.IV.19. lettre 'sin' contrôlé.

(Application de l'algorithme 2). (Application de l'algorithme 12).

La figure Fig.IV.19 montre que le caractère à dents س a été correctement segmenté par le module de contrôle (application de l'algorithme 12).

VI.3. Conclusion

Dans ce chapitre, nous avons montré une implémentation de la technique de segmentation des mots arabes en caractères selon la méthode de A. Amin améliorée, et nous avons proposés quelques améliorations dont le rôle est de contrôler les caractères qui sont mal segmentés surtout ceux qui commencent ou se terminent par un trait aminci, et quelques caractères possèdent des points dans la fin du mot. L'algorithme de contrôle est basé sur

l'usage des histogrammes, le contour des mots et les critères concernant la morphologie des caractères arabes. Ces critères utilisent le niveau structurel qui prend en considération toute les polices de caractères et toutes les tailles.

Chapitre V

Présentation du logiciel

V.1. Introduction

Le développement des logiciels a beaucoup évolué et dispose à présent d'outils performants. Notamment, depuis l'apparition et l'expansion de l'environnement Windows, qui offre à l'utilisateur une interface conviviale et facile à utiliser.

V.2. Environnement de travail

Parmi les langages et les logiciels de programmation existants, le langage c++ est celui qui s'adapte le mieux au traitement d'images. Aujourd'hui avec c++ Builder, on peut dire qu'il bénéficie à la fois de la puissance et de développement des applications Windows incontestablement plus rapide qu'avec les outils d'autres langages. Notre application a été développée sous l'environnement Windows, en se basant sur le logiciel de programmation Borland c++ Builder 6.

Ce logiciel est basé sur le concept de programmation orienté objet, il permet à un développeur non expérimenté de créer facilement une interface homme/machine d'aspect Windows.

- Le c++ offre la possibilité d'utiliser toutes les bibliothèques du système d'exploitation Windows.
- Le problème de la mémoire est résolu, l'utilisateur peut réserver autant d'espace mémoire qu'il veut (efficacité de c++ Builder dans la gestion de la mémoire).
- Le menu, les objets sont configurés dans le c++ Builder d'une façon standard, et l'utilisation peut changer leurs propriétés d'une façon très libre. [Dji 03].

Le développement d'une application Windows à l'aide de c++ Builder consiste en effet à :

- Concevoir l'interface utilisateur d'une façon visuelle et interactive
- Spécifier d'une manière tout aussi interactive les caractéristiques (fenêtres, boutons de commande, zone d'édition,... etc.).
- Spécifier le code à exécuter lorsque l'utilisateur effectue une action.

Borland apporte ainsi à travers le c++ Builder une grande puissance de programmation et une convivialité dans la mise en œuvre des outils suivants :

- Editeur multifenêtre
- Utilisation de la souris.

- Editeur du programme, compilateur, éditeur de liens, aide par hypertexte dans un environnement intégré.
- Possibilité de recouvrement de programme en mémoire (overlay).
- Gestion améliorée de projet.
- Débogueur intégré.
- Divers outil d'aide.

Notre environnement de programmation se présente comme suit :

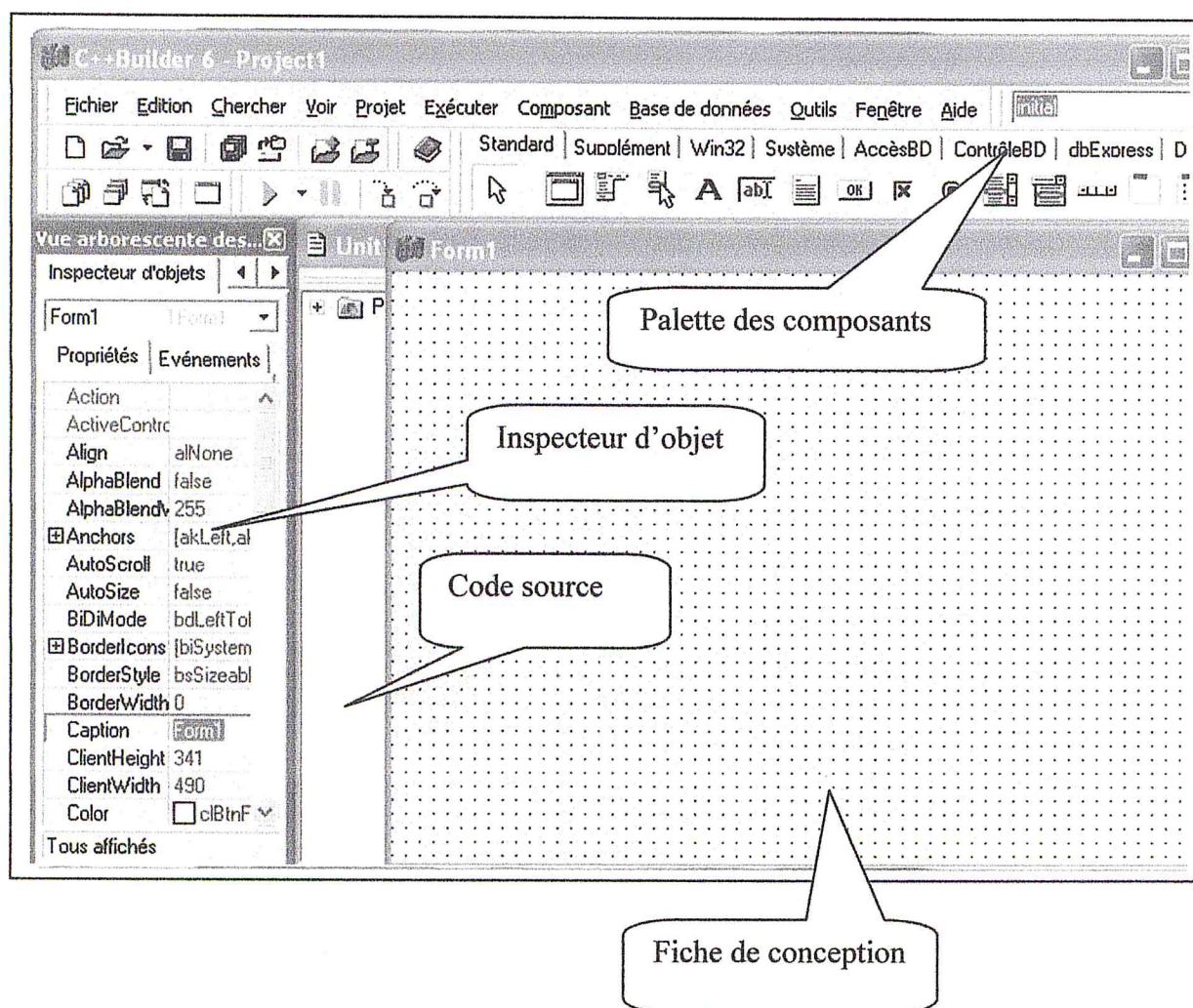


Fig.V.1. Interface d'environnement sous c++ Builder

En ouvrant un programme en c++ Builder, il nous serait présenté :

- Unit (unité) : est la partie où on écrit le code (c++).
- Form (forme) : est le code visuel où on place les composants.
- Palette des composants : des outils qui assurent différentes fonctions.

- **Projet (projet) :** une collection des fichiers et de codes qui gèrent et exécutent le programme.

V.3. Description du logiciel

L'interface du logiciel comporte une fenêtre principale offrant à l'utilisateur le module "Prétraitement", et le module "Segmentation", sont représentées chacun par une fenêtre.

La fenêtre principale est représentée comme suit :

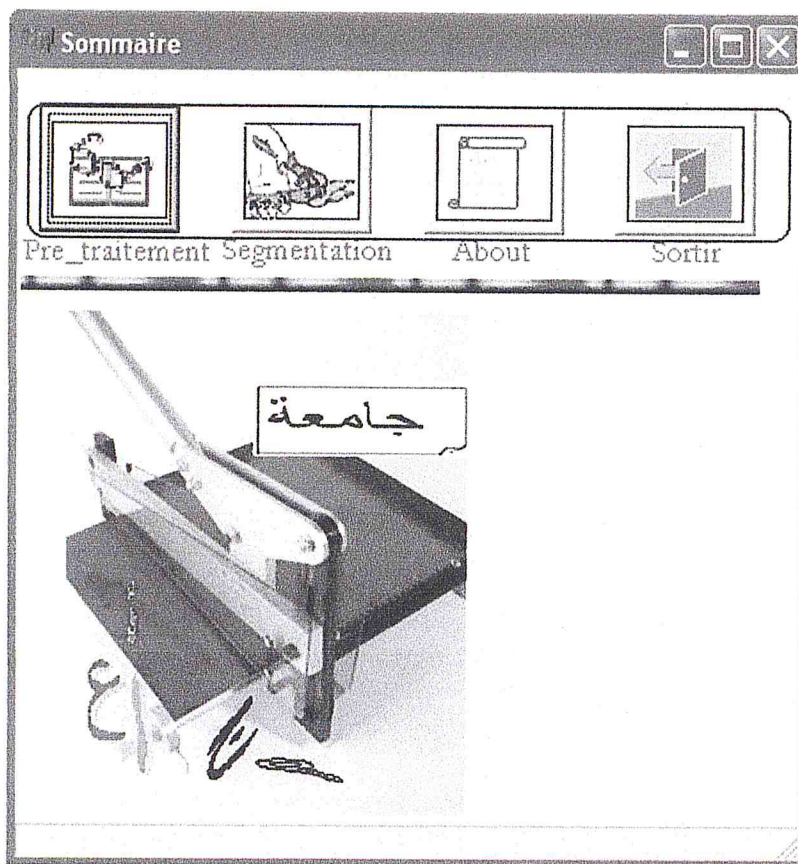


Fig.V.2. Fenêtre principale.

➤ **Prétraitement**

Ce menu permet d'accéder à trois commandes permettant le filtrage de l'image originale, la binarisation (avec le choix d'un seuil), le filtrage médiane, et la dilatation. Chacune de ces opérations permettent la suppression du bruit pour obtenir une image plus nette.

Le module de prétraitement est représenté par la fenêtre suivante :

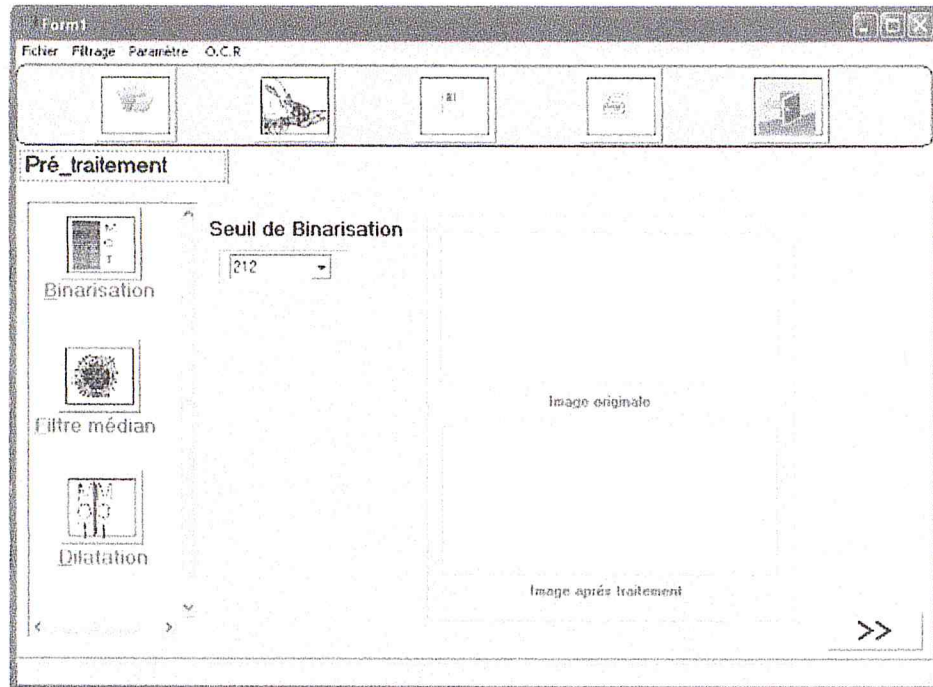


Fig.V.3. Fenêtre de module de prétraitement.

➤ Segmentation

Le menu de segmentation sert à représenter les différentes méthodes de segmentation, Il est compose de la fenêtre suivante :

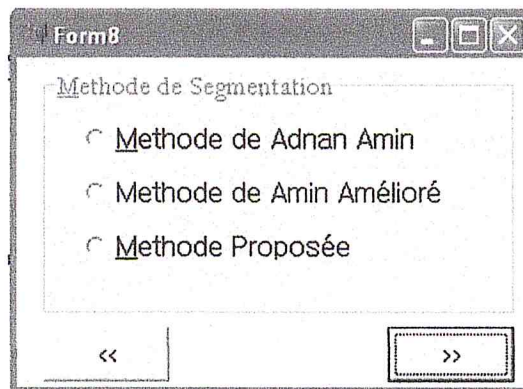


Fig.V.4. Fenêtre du choix de la méthode de segmentation.

Les trois méthodes : la méthode de A. Amin, la méthode de A. Amin améliorée, et la méthode proposée, sont représentées par les fenêtres suivantes:

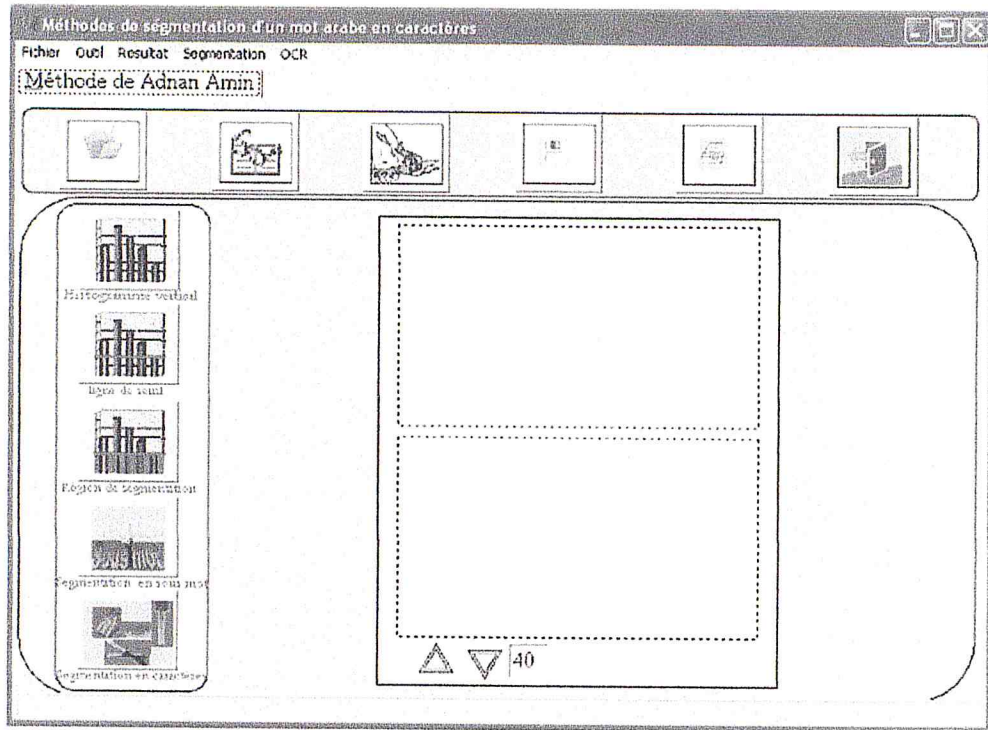


Fig.V.5. Fenêtre représente la segmentation par A. Amin.

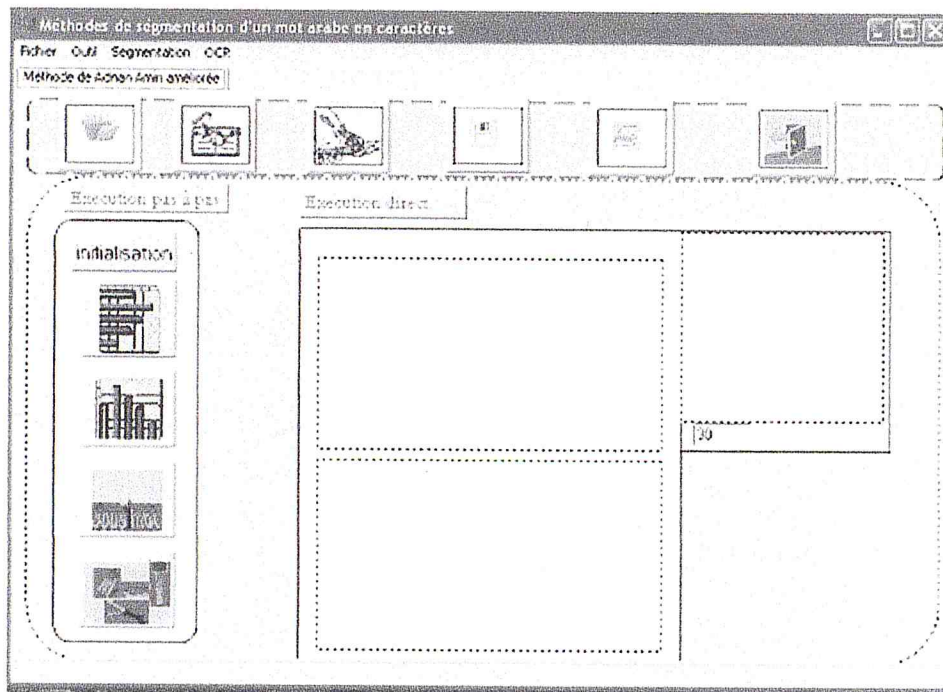


Fig.V.6. Fenêtre représente la segmentation par A. Amin améliorée.

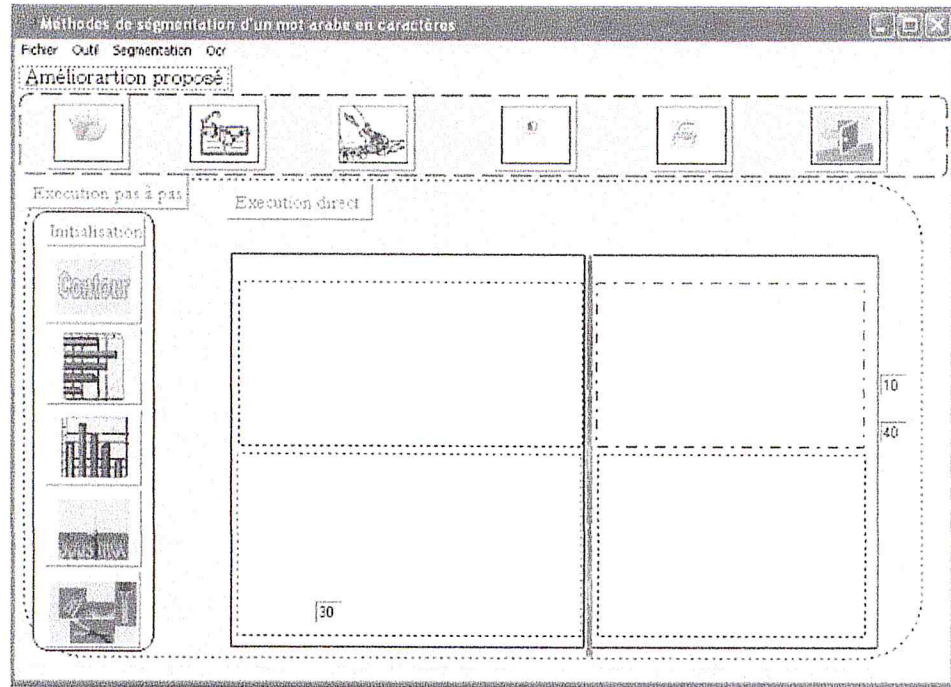


Fig.V.7. Fenêtre représente la segmentation par méthode proposée.

➤ About

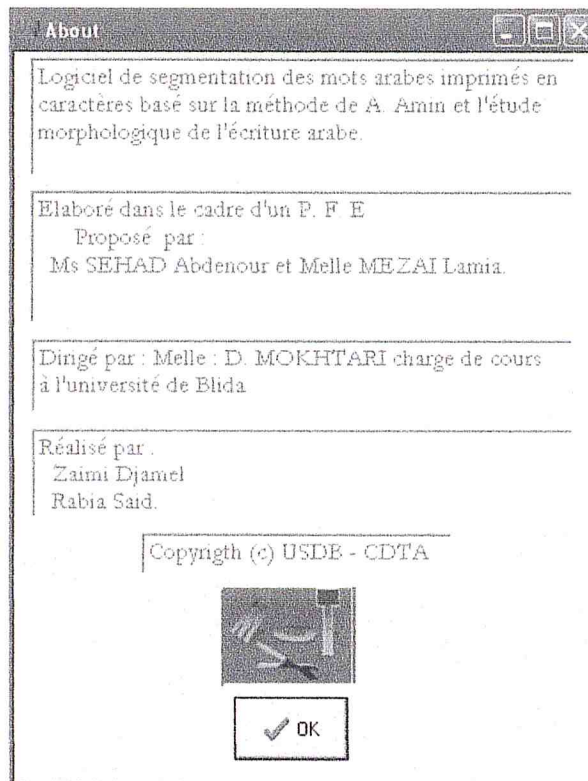


Fig.V.8. La fenêtre "About...".

V.3.1. Les menus

Les menus sont représentés dans ce qui suit

| Fichier | |
|----------------|--------|
| Nouveau projet | Ctrl+N |
| Ouvrir Image | Ctrl+O |
| Enregistrer | Ctrl+E |
| Imprimer | Ctrl+P |
| Quitter | Ctrl+X |

- **Nouveau projet** : Permet d'initialiser un nouveau projet.
- **Ouvrir Image** : Permet d'ouvrir une image d'extension BMP.
- **Enregistrer** : Permet de sauvegarder l'image courante.
- **Imprimer** : Permet d'imprimer l'image courante.
- **Quitter** : Permet de quitter la fenêtre en cours.

Filtrage

Binarisation
Filtre médian
Dilatation

- **Binraisation** : réaliser la binarisation de l'image en cours selon le choix d'un seuil.
- **Filtre médiane** : permet de diminuer le bruit.
- **Dilatation** : permet d'éliminer les pixels blancs isolés.


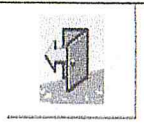





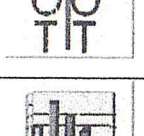
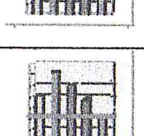
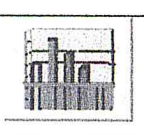



OCR



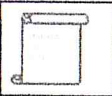
Aquisition
pré_traitement
Segmentation

- **Prétraitement** : permet d'ouvrir la fenêtre de prétraitement.
- **Segmentation** : permet de choisir une méthode de segmentation.

V.3.2. Barre des boutons

La barre des boutons contient les boutons de raccourcis vers les traitements les plus utilisés, ils sont représentés dans ce qui suit :

| Barre de bouton | Fonction |
|---|---|
|  | Permet d'ouvrir une image d'extension BMP. |
|  | Permet de fermer la fenêtre courante. |
|  | Ouvrir la fenêtre du module de Segmentation |
|  | Ouvrir la fenêtre de module de Prétraitement. |
|  | Sauvegarder l'image courante. |
|  | Effectuer la Binarisation. |
|  | Permet d'effectuer le Filtrage Médian. |
|  | Permet d'effectuer la Dilatation. |
|  | Calculer l'Histogramme Vertical. |
|  | Calculer le Seuil par la méthode de Amin. |
|  | Déterminer la Région De Segmentation. |
|  | Segmenter un mot arabe en sous mots. |
|  | Segmenter un mot arabe en caractères |

| | | |
|---|--|---|
|  | | Calculer l'Histogramme Horizontal. |
|  | | Calculer le Contour Du Mot. |
|  | | Visualiser la boîte de dialogue "About... " |

V.4. Résultats expérimentaux

Dans cette partie, nous présentons les résultats expérimentaux effectuée sur l'application développée.

➤ Module de prétraitement

❖ Binarisation

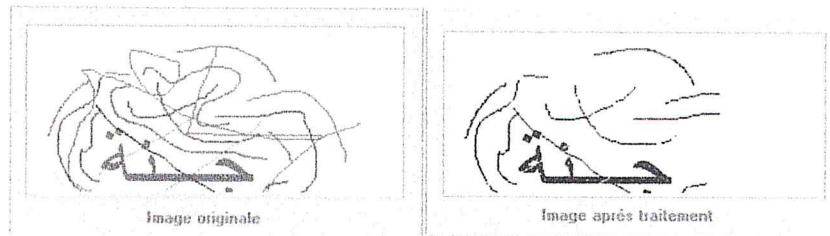


Fig.V.9. Résultat de Binarisation.
Seuil=212.

❖ Filtrage médian

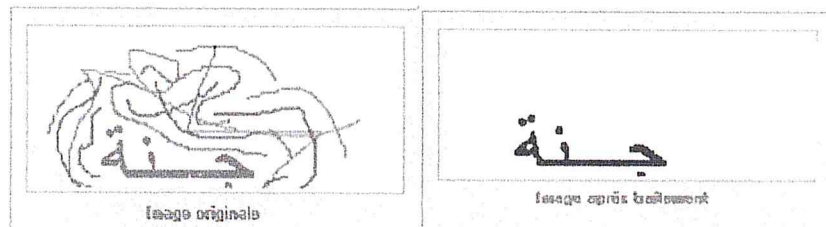


Fig.V.10. Résultat de Filtrage

❖ Dilatation

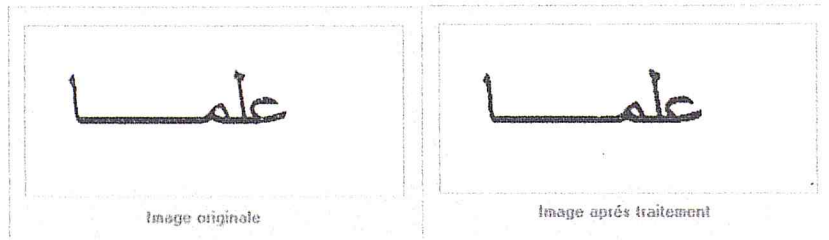


Fig.V.11. Résultat de la Dilatation.

➤ Segmentation

1. La méthode de Adnan Amin

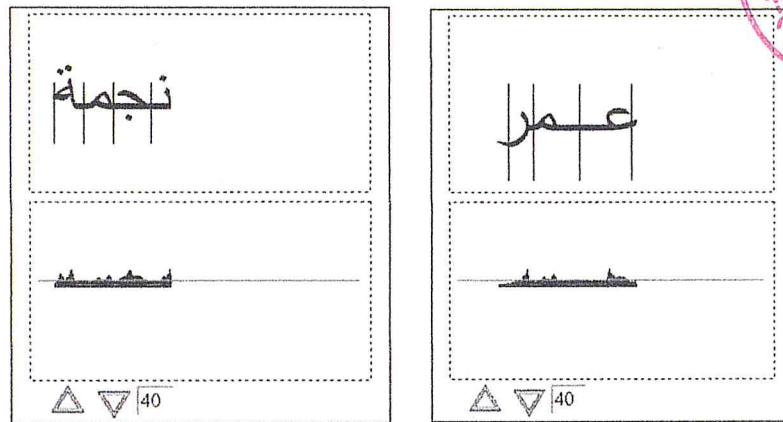


Fig.V.12. Avantage de la méthode. Fig.V.13. Inconvénient de la méthode.

2. Amélioration de Adnan Amin

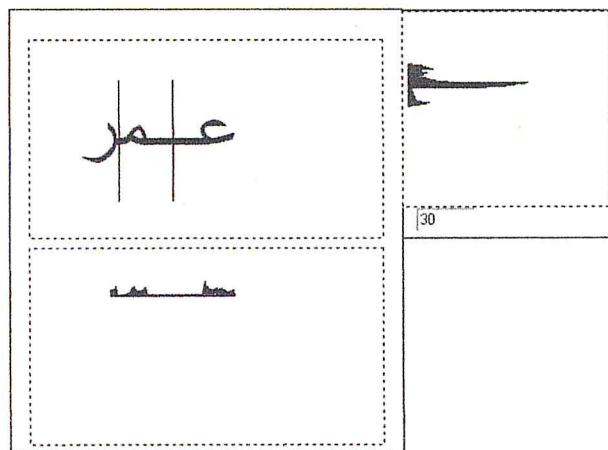


Fig.V.14. Avantage de la méthode.

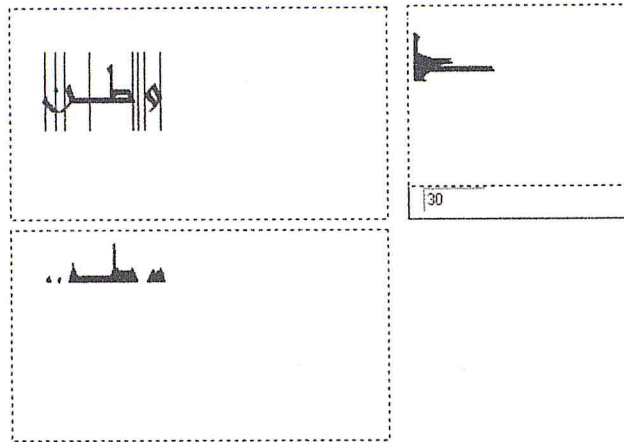


Fig.V.15. Inconvénient de la méthode.

3. Amélioration proposée

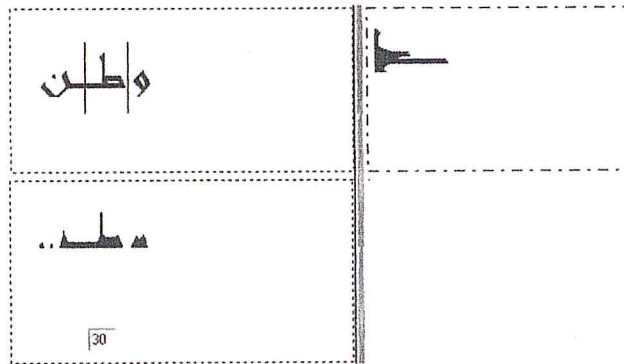


Fig.V.16. Amélioration proposée.

▪ Quelques exemples d'amélioration

Pour tester l'amélioration proposée, nous avons choisi plusieurs mots, et nous avons effectué la comparaison avec la méthode de A. Amin améliorée.

Le tableau suivant montre la différence entre les résultats de la segmentation.

| Echantillon | Segmentation par A.A améliorée | Segmentation par amélioration proposée |
|-------------|--------------------------------|--|
| عالم | عالم | عالم |
| النور | النور | النور |

| | | |
|--------|--------|--------|
| النور | النور | النور |
| النور | النور | النور |
| نجاح | نجاح | نجاح |
| كوكب | كوكب | كوكب |
| بيت | بيت | بيت |
| ثلاث | ثلاث | ثلاث |
| عدل | عدل | عدل |
| وطن | وطن | وطن |
| بلدي | بلدي | بلدي |
| الحق | الحق | الحق |
| لطيف | لطيف | لطيف |
| الصراط | الصراط | الصراط |

| | | |
|---------|---------|---------|
| نجم | نجم | نجم |
| التدريب | التدريب | التدريب |
| مارس | مارس | مارس |
| سماء | سمااء | سمااء |
| عسل | عسل | عسل |
| ي ت ث ب | ي ت ث ب | ي ت ث ب |
| غ ع ق ف | غ ع ق ف | غ ع ق ف |
| خ ج ح | خ ج ح | خ ج ح |
| شمس | شمس | شمس |
| الزكاة | الزكاة | الزكاة |

Tab.2. Comparaison des résultats de segmentation.

V.5. Conclusion

L'élaboration de ce logiciel a permis d'offrir à l'utilisateur un ensemble d'outils de traitement numérique d'image, de visualisation, et d'impression. Il est caractérisé par l'accessibilité à une bibliothèque d'algorithmes classiquement utilisé en traitement d'image

numériques, par la souplesse d'emploi à l'aide du menu interactif, par la diversité des méthodes de segmentations, et par la possibilités d'évolution.

Le développement futur de ce logiciel, portera essentiellement sur l'ajout de nouveau traitement, la mise au point d'algorithmes d'amélioration afin d'obtenir de meilleurs résultats.

Conclusion générale

Conclusion générale

La lecture automatique des documents est devenue indispensable dans plusieurs domaines, vu qu'elle facilite aux machines la compréhension des documents saisis par des outils, autre que le clavier, tel que la caméra ou le scanner.

Les systèmes OCR sont le fruit des recherches entrepris dans le domaine de la reconnaissance des textes saisis par un procédé optique. Ces systèmes sont structurés en architecture modulaire afin de simplifier les traitements.

L'étude que nous avons menée dans le cadre de ce projet s'est essentiellement basée sur la réalisation d'un module de segmentation des mots arabes imprimés en caractères.

En effet, un module d'acquisition est sollicité afin d'acquérir l'image du document, il offre tous les outils de manipulation, d'affichage et de sauvegarde.

Le module de prétraitement a été réalisé dans le but d'améliorer la qualité des images, présentant éventuellement du bruit, il regroupe les principales opérations dont il faut se prémunir avant d'entreprendre toute action d'analyse.

En ce qui concerne la segmentation, le but de ce module est de rechercher la structure physique du document imprimé, l'objectif consiste à extraire les entités physiques d'une image (ligne, mot, sous mot, caractère).

Le troisième niveau concerne la segmentation des mots en caractères qui est spécifique aux écritures cursives, en particulier l'arabe.

Plusieurs travaux ont été menés dans ce sens, en général elles se basent sur des propriétés propres à l'écriture arabe, parmi lesquelles, nous citons la méthode de A. Amin, la méthode de A. Amin Améliorée, la méthode Bouhlila, la méthode de Over_segmentation, ...

Le grand problème des méthodes de segmentation et la variation des seuils choisis pour déterminer les points de séparation potentiels, c'est-à-dire les séparateurs délimitant le début et la fin de chaque caractère.

Cette étude permet la segmentation des mots arabe imprimés en caractères par l'outil de l'histogramme et le contour des mots avec utilisation de l'étude morphologique de l'écriture arabe. Nous avons obtenu des résultats satisfaisants pour les caractères qui sont mal segmentées par les autres méthodes de Amin.

Dans la présente étude, nous avons développé un logiciel qui permet de segmenter des mots arabes en caractères, en se basant sur le progrès considérable que connaît le domaines de l'imagerie.

L'élaboration de ce logiciel a permis entre autres, de nous initier au vaste domaine du traitement numérique des images, de mieux maîtriser le langage c++ Builder 6.

Le système tel qu'il est développé, offre aux utilisateurs un ensemble d'outils de traitement, de visualisation et d'impression d'images. Il est caractérisé par l'utilisation de traitement des images numériques, par des méthodes de segmentation et par la possibilité d'évolution.

Nous espérons que ce modeste travail peut servir de base de départ à d'autres projets éventuels dans le domaine d'imagerie appliquée à la segmentation des mots arabes imprimés en caractères.

Nous estimons aussi que notre méthode de segmentation associée à d'autres méthodes de reconnaissance permet d'obtenir un système de reconnaissance de texte arabe performant.

Bibliographie

Références bibliographiques

- [Ami86] A. Amin, "Machine recognition of multi-fonts printed Arabic text" proc.8 th. International Conference on Pattern Recognition. Paris Oct. Pp 392-395(1986)
- [And 87] M. Andre « Introduction aux technique de traitement d'image », Paris, 1987.
- [Azz 90] K. Azzizi, S.Lemhannet, « Etude et implémentation d'un système de reconnaissance des mots arabes imprimés », Mémoire d'ingénieur, INI 1990.
- [Ben 94] A. Benhouhou, « Contribution à la conception et la réalisation d'un système de reconnaissance des caractères arabes imprimés multipolices » Projet master, INI 1994.
- [Ber 00] R. Berrou, F. Sahraoui, « Segmentation d'images par croissance de régions et contours actifs », Mémoire d'ingénieur INI 2000.
- [Bou 98] A. Bouderaa, S. Mazou, « Conception et réalisation d'un module de segmentation de textes arabes imprimés », Mémoire d'ingénieur USTHB 1998.
- [Dar 94] P. Dargenton, « Contribution à la segmentation et la reconnaissance de l'écriture manuscrite », Thèse de doctorat, Institut national des sciences appliquées de Lyon, 1994.
- [Dji 03] L. Djilali, B. Mohammed. «Extraction d'information et modélisation d'une carte de contour par la transforme de Hough », Mémoire d'ingénieur, Université Saad Dahlab, Blida, 2003.
- [Far 99-00] F. Hamaidi, M. Methia, « Recherche d'un algorithme de compression d'image reposant sur la théorie des Ondelettes avec conservation des propriétés relatives à la reconnaissance de formes », Mémoire d'ingénieur, INI .1999-2000.
- [Gil 99] A. Gillies, E. Erlandson, J. Trenkle, S. Schlosser, « Arabic Text Recognition System », NovoDynamics, Inc, 123 N. Ashly Street, Suite 120, Ann Arbor (USA), 1999
- [Gon 77] R.C Gonzalez, P. Wintz, « Digital Image Processing » Addison-Wesley 1977.
- [Guo 89] Z. Guo, R.W. Salle, "Amincissement Parallèle avec two-subiteration algorithmes", CACM Vol. 32, aucuns 3, pp.359-373(1989).
- [Gui 96] J. P. Guillois, "Technique de compression des images ", HERMES, 1996.
- [Ham 02] L. Hamami, D. Berkani, «Recognition System for printed multi-font and multi-size Arabic characters». The Arabian Journal for Science and Engineering, Volume 27, Number 1B. 2002.
- [Hum 92] A. Amin, H. B. Al-Sadoun, "A New Segmentation Technique of Arabic Test ", Proceedings., 11th IAPR International Conference on Pattern Recognition Methodology and Systems, 30 Aug.-3 Sept. 1992 Page(s):441 - 445.

- [Khe 98] M. Khechabtia, M. Iheddadene, « Contribution à la conception et à la réalisation d'un système de reconnaissance des mots arabes imprimés multi fontes », Mémoire d'ingénieur INI 1998.
- [KUN 93]. M. Kass, « Traitement numérique des images », Vol.2, 1993.
- [Mar 87] A. Marion, " Introduction aux techniques de traitement d'images", HERMES, 1987.
- [Nou 02] H. Nouredine, S. Meftah, « Système OCR », Mémoire d'ingénieur, INI 2002.
- [Sel 96] A. Sellab, H. Hocini, « Contribution à la conception et à la réalisation d'un système d'analyse des documents imprimés », Mémoire d'ingénieur USTHB 1996.
- [Tou 90] J . Toumazet, "Traitement de l'image sur micro-ordinateur", Paris, 1990.
- [Tun 88] « Regional conference on informatics and arabization », Tunis, 1988.
- [Yaz 92] M.YAZID, « Contribution à la conception et la réalisation d'une système de reconnaissance des caractères arabes imprimés multi-fontes », Mémoire d'ingénieur INI 1992.

Annexe

Etude morphologique de
l'écriture arabe

1. Introduction

Il nous est apparu nécessaire de faire une étude morphologique de l'écriture arabe qui présente des particularités importantes par rapport aux autres écritures connues (latin, chinoise, indu...etc.). Dans cette annexe, on présentera l'alphabet arabe et d'autres propriétés de l'écriture ligaturés, styles d'écritures...etc. Cette étude a pour but de spécifier les critères de segmentation de mots en caractères que nous verrons dans notre étude.

2. Particularités de l'écriture arabe

Dans l'écriture arabe, il n'existe pas de différence entre le manuscrit et l'imprimé. Au contraire l'imprimé a été inspiré du manuscrit et il n'y a ni majuscules ni minuscules. Seulement, chaque caractère peut être écrit en plusieurs formes suivant sa position dans le mot.

➤ Consonnes

L'alphabet arabe est composé de 28 caractères de base. La lettre «hamza ء» non considérée comme tel, joue toutefois un rôle important dans la phonétique, elle peut être indépendante comme elle peut se trouver sur des voyelles «Alif ا», «Ya ي», «Waw و», ou bien au dessous de la lettre «alif ا», ou aussi au dessus de la forme initiale et médiane de la lettre «Ya» pour donner la lettre Ya-hamza «أ». On considère aussi le caractère «Lam-Alif ل» comme étant la 29^{ème} lettre de l'alphabet. Chaque lettre peut s'écrire de 2 à 4 formes différentes selon son contexte dans le mot.

- Initial : en début de mot uniquement.
- Médiane : en milieu du mot.
- Finale : à la fin du mot mais lié avec le caractère précédent.
- Isolée : à la fin du mot mais non lié avec le caractère précédent.

7 caractères sur 28, on seulement deux formes distinctes (Waw, Ra, Za, Dhal, Alif) initial et finales et ils sont toujours suivis de formes initiales quelque soit leur position dans le mot.

On notera la présence de groupe de points et leurs positions jouent un rôle prépondérant pour la discrimination des caractères appartenant à une même famille de forme. Ces groupes de points (1 à 3 points) peuvent occuper trois positions possibles relativement à la forme : au-dessus, au milieu, ou en-dessous.

Le tableau suivant illustre l'alphabet arabe :

| Car | Isolée | Initiale | Médiane | Fin | Car | Isolée | Initiale | Médiane | Fin |
|------|--------|----------|---------|-----|-----|--------|----------|---------|-----|
| Alif | أ | أ | أ | أ | Ba | ب | ب | ب | ب |
| Ta | ت | ت | ت | ت | Tha | ث | ث | ث | ث |
| Djim | ج | ج | ج | ج | Ha | ح | ح | ح | ح |
| Kha | خ | خ | خ | خ | Dal | د | د | د | د |
| Dhal | ذ | ذ | ذ | ذ | Ra | ر | ر | ر | ر |
| Zin | ز | ز | ز | ز | Sin | س | س | س | س |
| Chin | ش | ش | ش | ش | Sad | ص | ص | ص | ص |
| Dhad | ض | ض | ض | ض | Taa | ط | ط | ط | ط |
| Dha | ظ | ظ | ظ | ظ | Ayn | ع | ع | ع | ع |
| Ghin | غ | غ | غ | غ | Fa | ف | ف | ف | ف |
| Qaf | ق | ق | ق | ق | Kaf | ك | ك | ك | ك |
| Lam | ل | ل | ل | ل | Mim | م | م | م | م |
| Noun | ن | ن | ن | ن | Ha | ه | ه | ه | ه |
| Waw | و | و | و | و | Ya | ي | ي | ي | ي |

Tab.3. L'alphabet arabe.

➤ Voyelles

Sont au nombre de 11 dont 3 de base dites brèves sont figurées par des signes diacritiques appelées *Fathah* (sa prononciation est 'a': un petit trait au dessus de la consonne), *Dammah* (sa prononciation est 'ou': un petit Waw au dessus de la consonne) et *Kasrah* (sa prononciation est 'i': un petit trait au dessous de la consonne). Il existe d'autres lettres qu'on peut considérer aussi comme des voyelles (Alif, Ya, Waw) suivant leur position dans le mot. Mis à part l'accent (Hamza).

Les journaux, les revues, les livres et autres imprimés administratifs n'utilisent pas des voyelles. Elles sont généralement utilisées dans les livres scolaires de primaire et dans le texte du livre sacré «Le Coran ». Les voyelles sont représentées dans la Tab.4

3. Classification de l'alphabet arabe [Ben 94]

Les caractères de l'alphabet arabe sont regroupés en classes suivant des critères géométriques. Voici un tableau qui illustre la classification des lettres de l'alphabet arabe.

| Classe | Lettre | | Classe | Lettre |
|--------|---------------|--|--------|------------------|
| 1 | ب، ت، ث، ن، ي | | 5 | ع، غ |
| 2 | ج، ح، خ | | 6 | ص، ض، ط، ظ |
| 3 | س، ش | | 7 | د، ذ، ر، ز، و، ا |
| 4 | ف، ق | | 8 | م، ك، ه، ل، ء |

Tab.6. Classification des lettres de l'alphabet arabe.

4. Styles d'écriture arabe [Ben 94]

Avant tout, il faut bien savoir que l'écriture arabe est l'une des écritures les plus artistiques au monde. Avec le temps, elle a connu plusieurs styles d'écritures. De nos jours, les styles les plus utilisés par les imprimeries sont le style **NESKHI**, **KOUFI** et le **REKKAH**.

❖ Style NESKHI

C'est un style d'origine coranique qui est très largement utilisé dans de nombreux domaines, son écriture est très arrondie et fortement cursive. Elle présente par contre d'importantes variations dans les liaisons entre les lettres qui ont entraîné la naissance de combinaisons de lettre pour former des ligatures. La majorité des écritures d'imprimeries actuelles sont une transcription directe de ce style.

❖ Style KOUFI

C'est le style le plus ancien, son nom vient de la ville de Koufa (en Irak). C'est une écriture à caractère rectangulaire d'une simplicité extrême, essentiellement utilisée comme écriture ornementale ou décorative dans les mosquées. De nos jours, elle est utilisée dans les titrages. Ce style se caractérise par des lettres avec des hauteurs centrales fortement prononcées et une épaisseur constante des traits.

❖ Style REKKAH

Ce style est originaire de Turquie, il se présente comme une écriture simple, sa mise en forme est basée sur les formes cursives. Ce style est le moins utilisé à cause des erreurs de lecture qu'il peut engendrer.

❖ Autres styles

Il existe beaucoup d'autres styles non apparentés à une classe déterminée, ces styles sont généralement utilisés dans les décorations et non en typographie.

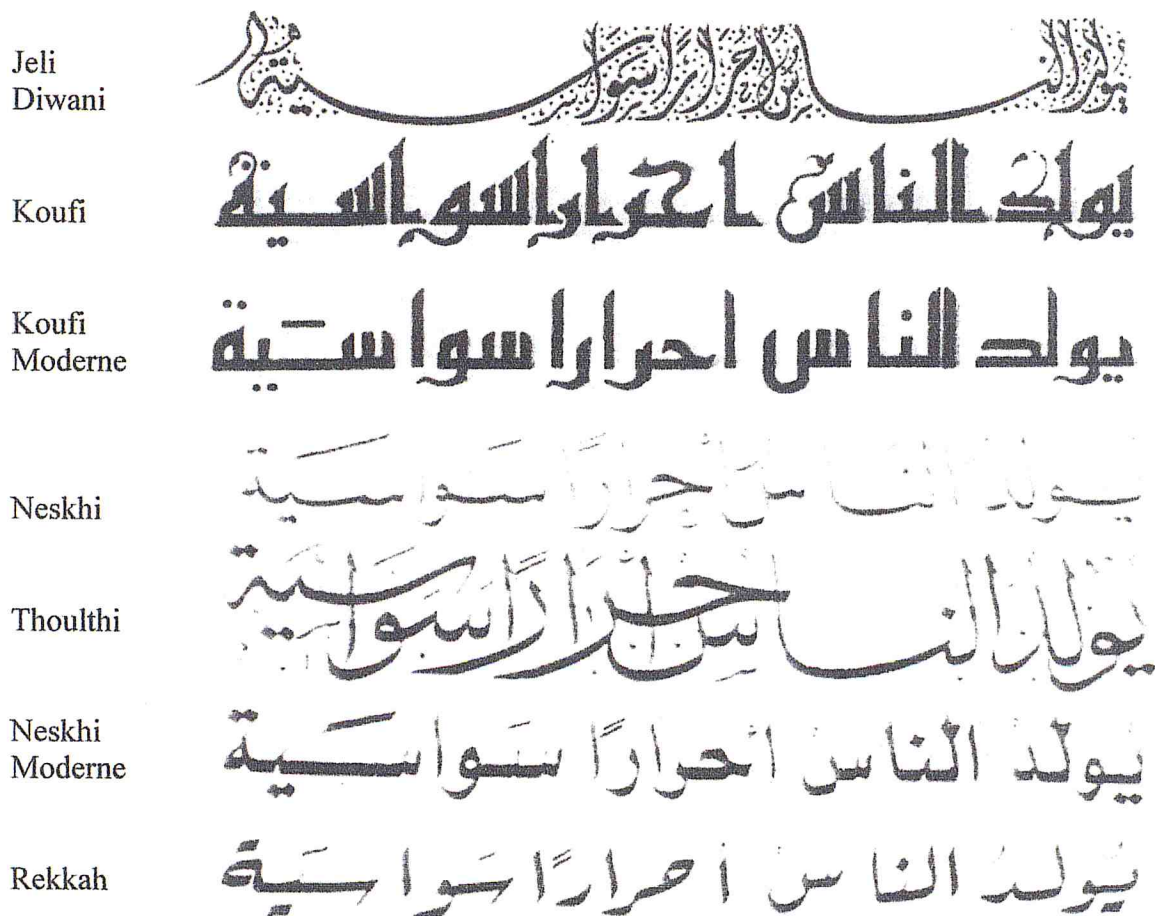


Fig.1. Exemples de styles d'écriture arabe.

5. Arabe imprimé et dactylographié [Yaz 92]

L'imprimerie s'est inspirée de l'écriture manuscrite pour fondre les caractères arabes. En effet, on ne fait pas de différence entre l'écriture manuscrite et imprimée. Il n'y a ni majuscules ni minuscules.

Un caractère arabe imprimé est caractérisé par la police (fonte) et le corps dans lequel il est assorti, la première présente la nomination de la fonte qui change selon les constructeurs des machines d'imprimerie (ou même le concepteur de la police de caractères), la deuxième représente les dimensions maximales de la matrice encadrant le caractère. Il existe actuellement plus de douze fontes officielles imprimables en arabe