UNIVERSITÉ Sidi Mohamed Ben Abdellah
Faculté Des Sciences Et Techniques Fès
Département d'Informatique

Projet de Fin d'Etudes

Master Sciences et Techniques
Systèmes Intelligents & Réseaux

Contribution à la reconnaissance optique (OCR) du texte arabe imprimé

Lieu de stage : **Laboratoire systèmes intelligents et applications – FSTF**
Istituto di Linguistica Computazionale – Pise, Italie

Réalisé par : Younes LASRI

Soutenu le : 18 juin 2014

Encadré par :

Mr : Arsalane ZARGHILI
Mme : Ilham CHAKER
Mr : Federico BOSCHETTI

Devant le jury composé de :

Pr. Arsalane ZARGHILI
Pr. Ilham CHAKER
Pr. Jamal KHARROUBI
Pr. Med Chaouki ABOUNAIMA
Mr. Federico BOSCHETTI

Année Universitaire 2013-2014

REMERCIEMENTS

A mes encadrants, monsieur ZARGHILI et madame CHAKER, pour leur aide et précieux conseils ainsi que pour le soutien accordé tout au long de ce projet.

A monsieur FEDERICO pour sa disponibilité et ses recommandations qui m'ont aidé à franchir les obstacles.

A tous les professeurs qui m'ont encadré durant ma formation en maîtrise et en Master à la FSTF.

A mes parents qui m'ont encouragé et m'ont appris à avoir confiance en moi.

A ma femme pour son soutien et ses sacrifices et mes deux enfants qui donnent du charme à ma vie.

Enfin, j'adresse mes plus sincères remerciements à tous mes proches et amis, qui m'ont toujours soutenu et encouragé au cours de la réalisation de ce projet.

RESUME

Le présent travail porte sur une étude concernant le domaine de reconnaissance optique de caractères arabes imprimés mono-fonte hors-ligne.

Une étude générale sur les systèmes de reconnaissance de l'écriture a été développée, en accordant un intérêt particulier à la phase de segmentation. Puis nous avons présenté la langue arabe, ses caractéristiques et ses données graphiques.

En suite nous avons présenté notre contribution à la reconnaissance de texte arabe imprimé mono-fonte hors-ligne.

Nous avons obtenu des résultats de réussite très performants pour la segmentation, et encourageants pour la reconnaissance grâce au post-traitement.

ABSTRACT

This work focuses on a study of the arabic optical characters recognition printed mono-font offline.

A general study of writing recognition systems was developed, with a particular interest to the segmentation phase. Then we introduced the Arabic language, its features and graphics data.

As a result we presented our contribution to arabic optical characters recognition printed mono-font offline.

We got successful results for segmentation, and encouraging ones as to the recognition through post-processing.

SOMMAIRE

LISTE DES FIGURES

LISTE DES TABLEAUX

INTRODUCTION GENERALE

La reconnaissance optique de caractères (OCR : Optical Character Recognition) fait objet de l'avenir de la communication homme-machine. Elle est utilisée dans plusieurs domaines où le texte est la base de travail, principalement en bureautique, pour des buts d'indexation et d'archivage automatique de documents, en publication assistée par ordinateur (PAO) pour faciliter la composition à partir d'une sélection de plusieurs documents, dans la poste pour le tri automatique du courrier, dans une banque pour faciliter la lecture des montants de chèques, …

La reconnaissance de l'écriture relève du domaine de la reconnaissance des formes qui s'intéresse aux formes de caractères. Le but est d'attribuer à une forme un identifiant des prototypes de référence déterminés préalablement.

Les travaux de recherche en reconnaissance optique de caractère arabes (AOCR), bien que moins avancés que pour d'autres langues, deviennent plus intensifs qu'avant, en effet la reconnaissance des caractères arabes reste encore aujourd'hui au niveau de la recherche et de l'expérimentation, le problème n'est pas encore résolu. Les travaux effectués dans ce domaine sont nombreux et leurs résultats du point de vue méthodologique et théorique sont très encourageants, toutefois les performances des systèmes prototypes développés en milieu académique sont loin d'égaler les performances exigées par la qualité de service des systèmes opérationnels.

Dans ce travail nous présentons un aperçu sur la reconnaissance des caractères, les étapes suivies pour la réalisation d'un système OCR puis nous nous intéressons à la phase de segmentation et plus particulièrement à la segmentation en caractères. Nous introduisons ensuite les caractéristiques et les données graphiques de l'écriture arabe, suivie par une étude détaillée de notre contribution à la reconnaissance de l'écriture arabe imprimée mono-fonte hors-ligne, finalement nous présentons la partie pratique de ce travail et les résultats obtenus dans la phase de la segmentation et de la reconnaissance ainsi qu'une évaluation de notre système AOCR en le comparant à d'autres systèmes.

Organisation de rapport :

Ce rapport est constitué de quatre chapitres organisés comme suit:

- Le premier chapitre est un rappel de certaines notions générales d'OCR. Ainsi que les étapes nécessaires pour la réalisation d'un système de reconnaissance de l'écriture.
- Le deuxième chapitre étudie l'OCR et la langue arabe. La première et la deuxième section rappellent les caractéristiques de la langue arabe et les données graphiques de l'alphabet arabe, et la troisième présente certains systèmes de reconnaissance de l'écriture arabe en précisant pour chacun le mode utilisé, l'approche de la reconnaissance, le type de segmentation, et les scores réalisés.
- Dans le troisième chapitre, nous présentons notre contribution à la reconnaissance de texte arabe imprimé mono-fonte hors ligne.
- Finalement nous présentons la partie pratique de ce travail dans le dernier chapitre, les résultats obtenus et des comparaisons avec d'autres systèmes AOCR.

CHAPITRE I : LA RECONNAISSANCE DE L'ECRITURE

I. Introduction

La reconnaissance de l'écriture est un traitement informatique, qui consiste à transformer un texte écrit sur papier (manuscrit, imprimé ou encore dactylographié) en un texte numérique, respectant un codage (comme le code ASCII pour les écritures latines et le code ASMO pour celles arabes)

II. Différents aspects de l'OCR

Il existe plusieurs aspects de l'OCR selon le mode d'acquisition (en ligne et hors ligne), ou selon que le traitement se fait sur la totalité du mot (reconnaissance globale) ou sur les caractères après segmentation (reconnaissance analytique).

1. Reconnaissance en ligne et hors-ligne

a. Reconnaissance en ligne :

La reconnaissance se fait en temps réel (pendant l'écriture), peut être utilisée seulement pour l'OCR manuscrit et nécessite un matériel spécifique (stylo numérique ou stylet pour écrire sur un agenda électronique ou tablette tactile). Ce mode présente l'avantage de la possibilité de la correction automatique et la saisie semi-automatique d'une façon interactive grâce à la réponse en continu de système.

b. Reconnaissance hors-ligne :

C'est le mode le plus général de l'OCR, il consiste à reconnaitre un texte préalablement écrit ou imprimé sur une image de texte scannée.

La reconnaissance de l'écriture hors-ligne est utilisée plus dans le domaine économique, tel que :

- Lecture des adresses postales : se traduit dans le tri automatique du courrier en lisant les codes postaux manuscrits.
- Domaine bancaire : avec des machines capables de lire les montants des chèques ce qui a rendu possible l'encaissement des chèques dans les guichets bancaires automatiques.
- Formulaires et bordereaux : des applications qui permettent de lire les formulaires de sondage, les bordereaux de commandes ou encore les réponses des examens écrits.

2. Reconnaissance globale et analytique

La reconnaissance de l'écriture utilise principalement deux approches : globale et analytique.

a. Approche globale :

Cette approche considère la présence de caractéristiques sur l'ensemble du mot évitant les difficultés liées aux ambiguïtés provenant de la segmentation, ce qui lui permet de bien absorber les petites variations locales, mais la limite à un vocabulaire distinct.

b. Approche analytique :

Contrairement à l'approche globale, l'approche analytique cherche à identifier les graphèmes issus de la segmentation (fragments de lettres, des lettres ou des regroupements de lettres) pour reconstituer les mots. Elle présente l'intérêt de pouvoir se généraliser à la reconnaissance d'un vocabulaire étendu ou limité dynamiquement

III. Les étapes d'un système de reconnaissance

Un système de reconnaissance est constitué de six étapes : Acquisition, prétraitements, segmentation, extraction des caractéristiques, classification, et post-traitement.

1. Acquisition

C'est la phase qui consiste à capter l'image d'un texte, et de la convertir en grandeurs numériques adéquates au système de traitement en utilisant des capteurs physiques (Numériseur, caméra..).
Cette étape est assez simple, mais elle influence les étapes suivantes du système.
L'acquisition est caractérisée par la résolution et le niveau d'éclairage.

a. Résolution

La résolution optimale d'une image dépend de l'épaisseur du trait d'écriture. Ainsi, pour les traitements ultérieurs puissent s'appliquer correctement, il faut que le trait d'écriture ait une épaisseur minimale de 3 pixels [2]
La résolution souvent utilisée de 300dpi

b. Niveau d'éclairage

Un éclairage élevé (du numériseur) réduit le bruit, mais fait disparaitre les traits minces [3]. Donc il faut choisir un niveau d'éclairage optimal selon la qualité du document physique.

2. Prétraitements

La phase de prétraitement consiste à préparer les données issues du capteur à la phase suivante.
Il s'agit essentiellement de réduire le bruit (dû aux conditions d'acquisition ou à la qualité du document d'origine) superposé aux données et essayer de ne garder que l'information significative de la forme représentée. [1]

Les opérations souvent utilisées sont : le redressement de l'écriture, le lissage, la normalisation et la squelettisation. (Figure 1)

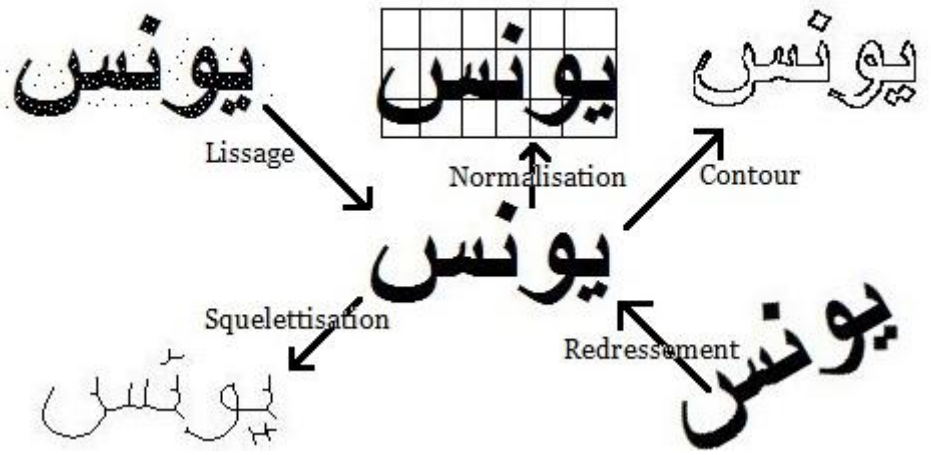


Figure 1 : Différentes opérations de prétraitement

a. Redressement de l'écriture

L'inclinaison des lignes peut causer des problèmes de segmentation du texte en lignes.

Pour remédier à ce problème, on redresse l'écriture en appliquant une rotation isométrique d'angle -α grâce à la transformation linéaire [4] :

$x' = x \cos\alpha + y \sin\alpha$

$y' = y \cos\alpha + x \sin\alpha$

Figure 2 : Texte incliné avec un angle

b. Lissage

L'image traitée peut être entachées de bruits dus aux artefacts de l'acquisition et à la qualité du document, conduisant soit à une absence de points ou à une surcharge de points. [3]

L'opération de nettoyage permet de supprimer les petites tâches et les excroissances de la forme.

Pour le bouchage il s'agit d'égaliser les contours et de boucher les trous internes à la forme du caractère en lui ajoutant des points noirs.

c. Normalisation

Pour faciliter les traitements ultérieurs, une normalisation de la taille est appliquée, en effet les images de tous les caractères se retrouvent définies dans une matrice de même taille.

Cette opération introduit généralement de légères déformations sur les images, cependant certains traits caractéristiques tels que la hampe dans les caractères (ل ظ ط ا par exemple) peuvent être éliminées à la suite de la normalisation, ce qui peut entraîner à des confusions entre certains caractères [4].

d. Squelettisation

Le but de cette technique est de simplifier l'image du caractère en une image à «ligne» plus facile à traiter en la réduisant au tracé du caractère.

Les algorithmes de squelettisation se basent sur des méthodes itératives, le processus s'effectue par passes successives pour déterminer si un tel ou tel pixel est essentiel pour le garder ou non dans le tracé [4].

i. Extraction du squelette [5]

L'opération de squelettisation consiste donc, dans le cas particulier de l'écriture, à éliminer l'épaisseur du trait ou plutôt à l'amincir jusqu'à l'épaisseur minimale d'un pixel.

Les critères retenus pour les méthodes de squelettisation sont les suivants :

- ✓ L'épaisseur de squelette doit être de 1 pixel ;

- ✓ Le squelette doit conserver les propriétés topologiques de la forme comme le nombre de trous et la connexité ;
- ✓ Le squelette doit respecter les propriétés métriques de la forme comme la longueur totale et la distance entre partie de la forme.

ii. L'amincissement

Les termes érosion ou éclaircissage ont un sens assez semblable à celui d'amincissement employé en morphologie mathématique et qui est contenu dans le terme anglais "thinning".

Cette technique consiste à appliquer un élément structurant d'une transformation dans laquelle le pixel central noir est remplacé par un pixel blanc en fonction de la configuration des pixels voisins.

Il existe trois stratégies sur la manière d'appliquer ce masque à l'ensemble des pixels de l'image.

Dans la première, un balayage horizontal et vertical de toute l'image est effectué en plusieurs passes jusqu'à ce qu'il ne reste plus de pixels à éroder, ou encore en une seule passe avec une érosion directe des pixels à chaque translation du masque. La troisième stratégie consiste en un suivi de contour appliqué successivement d'une manière analogue à une érosion "naturelle".

A l'intérieur de chacune de ces stratégies, on rencontre de nombreuses variantes qui conduisent à des améliorations portant sur les imperfections classiques du squelette.

iii. Les algorithmes à critères topologiques

Ils sont aussi appelés algorithmes de pelage. Ce sont des algorithmes itératifs supprimant à chaque étape, le long de la frontière de la forme, les points appelés inessentiels (c'est-à-dire n'appartenant pas au squelette). Ils utilisent souvent des fonctions booléennes opérant sur des voisinages de points, déterminant à chaque passage la validité des points frontières.

L'algorithme général

début
répéter
inessentiel ← faux
pour tout point dans image faire
si POINTINESSENTIEL(point) alors
inessentiel vrai

image (point) 0
fsi
fpour
jusqu'à inessentiel=faux
fin

L'algorithme général est de type séquentiel. Il consiste à balayer l'image ligne par ligne et à supprimer les points inessentiels au fur et à mesure qu'ils sont rencontrés. Il ne peut donc pas tenir compte de l'environnement dans lequel il opère et a tendance à enlever trop de points d'un seul coup. [3]

3. Segmentation (Etat de l'art)

A partir d'une image acquise il y'a d'abord séparation des blocs de texte et des blocs graphiques, puis à partir d'un bloc de texte il y a extraction des lignes, ensuite à partir de ces lignes sont extraits les mots puis les caractères (ou parties du caractère) [6].
La segmentation est une étape critique et décisive, en effet l'efficacité des systèmes de reconnaissance en dépend fortement.

a. Segmentation de la page

[1]Cette étape permet de localiser dans chaque page, les zones d'information conformément à leur apparence physique. Elle est associée généralement à l'étiquetage logique qui consiste à déterminer la nature du media représenté dans chaque zone (texte, graphique, photographie …).
Cette classification permet ensuite d'orienter la reconnaissance vers des systèmes spécialisés dans l'analyse de chaque type de media [7].
Une étude détaillée sur les techniques utilisées dans l'analyse de documents se trouve dans : ([8], [9], [10], [5], [11] et [12]).

b. Segmentation de texte en lignes

Pour localiser les lignes de texte il est possible de s'appuyer sur un modèle physique de disposition de l'écriture. La complexité de cette étape est très variable. En effet, on peut supposer en général que les lignes d'écriture sont plus ou moins parallèles et horizontales et mettre en œuvre alors des méthodes de détection relativement simples et satisfaisantes.

Toutefois les problèmes ne sont absolument pas résolus dans le cas de documents manuscrits présentant des dispositions variables, des ratures, des inclinaisons ou des chevauchements de lignes. Cependant, quelles que soient les approches envisagées, des problèmes subsistent notamment lorsque les lignes d'écriture se superposent partiellement en présence d'extensions hautes ou basses qui s'étendent jusque sur la ligne d'écriture inférieur ou supérieure. Le choix des points de coupure reste alors un problème difficile à résoudre sans faire appel au module de reconnaissance.

La méthode triviale de séparation de lignes fait appel à *la projection horizontale* qui n'est rien d'autre qu'une simple somme du nombre de points allumés par ligne.

On peut dire le début ou la fin d'une ligne de texte sont détectés, si la valeur de projection horizontale (figure 3) est inférieure à un seuil.

Le seuil est obtenu au minimum de l'histogramme. [3]

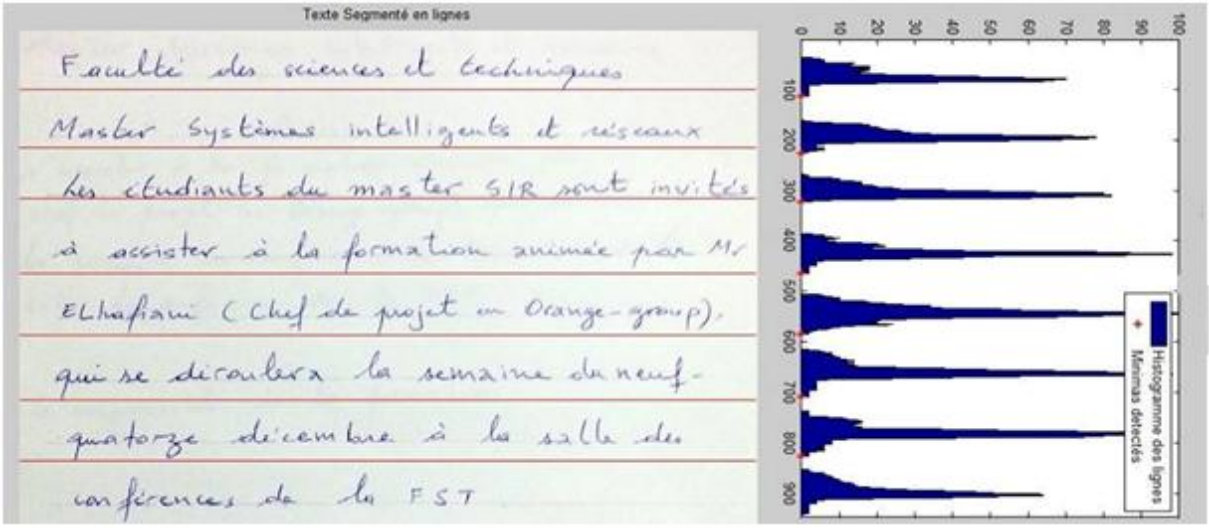


Figure 3 : Projection horizontale des lignes

c. Segmentation des lignes en mots

La localisation des mots dans les textes manuscrits est un problème qui relève du paradoxe de Sayre (1973) (Il faut segmenter le tracé pour reconnaître une lettre, mais il faut reconnaître une lettre pour segmenter le tracé). En effet, si théoriquement les règles de disposition de l'écriture imposent de marquer des espaces entre les mots plus importants (le cas d'imprimé), en pratique ces règles ne sont pas toujours vérifiées sur les écritures manuscrites non contraintes. Et on doit admettre que pour résoudre le problème, il est nécessaire de demander l'aide d'un système de reconnaissance.

De ce fait la localisation des mots dans une ligne de texte est un problème qui s'apparente à celui de la reconnaissance des caractères dans les mots.
On peut classer les approches proposées dans la littérature en deux grandes familles:

- La première concerne les approches qui recourent à une métrique spécifique adaptée au problème afin d'ordonner de la meilleur façon possible les espaces détectés dans les lignes pour qu'une simple technique de seuillage puisse permettre de séparer les espaces inter-mots des espaces inter-caractères.
- La seconde met en œuvre une étape de pré-connaissance pour attribuer les espaces détectés dans la phrase à l'une des deux classes, espace inter-mots, espace inter-lettres.

La méthode triviale de séparation de lignes fait appel à *la projection verticale* qui n'est rien d'autre qu'une simple somme du nombre de points allumés par colonne comme l'indique la figure 4 [3]

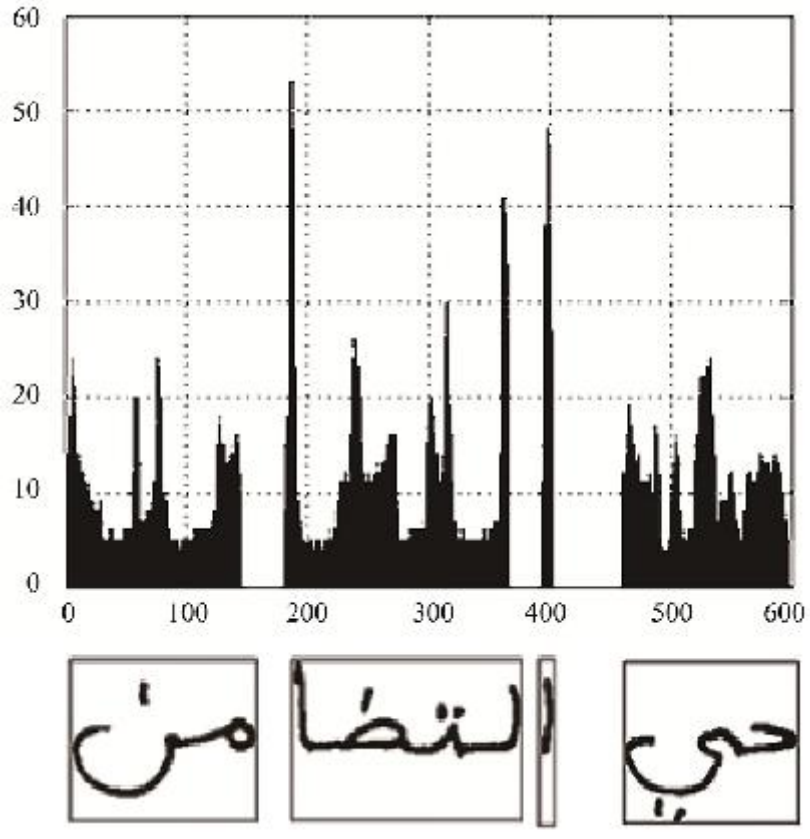


Figure 4 : Projection verticale d'une ligne

d. Segmentation des mots en caractères

La segmentation des mots en caractères est l'étape le plus délicate dans tout le processus d'un système de reconnaissance d'écriture.

C'est une opération qui tente de décomposer une image de séquence de caractères (mot) en sous-images de symboles individuels, ayant le but de décider si un motif isolé d'une image (caractère ou autre entité identifiable du mot) est correct ou non [13].

i. Organisation des méthodes

Certains auteurs tels que Tappet et Al dans [14] parlent de segmentation interne et externe, dépendant de si la segmentation se fait séparément ou simultanément avec la reconnaissance. D'autres auteurs utilisent les termes straight segmentation et segmentation recognition, pour exprimer le même sens que précédemment.
Selon le point de vue de Casey et Lecolinet dans [13] la classification des méthodes suivant l'utilisation ou non de la reconnaissance durant la phase de segmentation n'est pas une bonne classification, parce qu'on peut par exemple utiliser un correcteur d'orthographe comme post-processeur et dans ce cas il peut suggérer de substituer une lettre sortie par le classifieur par deux lettres, et cela est en fait une utilisation d'une segmentation de la sous image. Selon lui la distinction entre les méthodes est basée sur comment la segmentation et la classification interagissent dans tout le processus. Dans l'exemple précédent par exemple la segmentation intervient en deux temps, une fois avant la classification et une seconde fois après la classification. Après examen des méthodes, il les classifie en trois stratégies de segmentation. Plus d'autres méthodes hybrides à base des trois stratégies de base.
1) L'approche classique : dans laquelle les segments sont identifiés à base de propriétés de ressemblance de caractères. Elle utilise une technique de découpage de l'image en composants significatifs elle est appelée dissection.
2) Segmentation basée reconnaissance : dans laquelle le système cherche des composants qui correspondent à son alphabet dans l'image.
3) Méthodes holistiques : dans lesquelles le système essaye de reconnaître le mot comme un tout. Evitant ainsi le besoin de segmentation en caractères.
(Dans ce qui suit nous allons voir plus en détail ces stratégies)

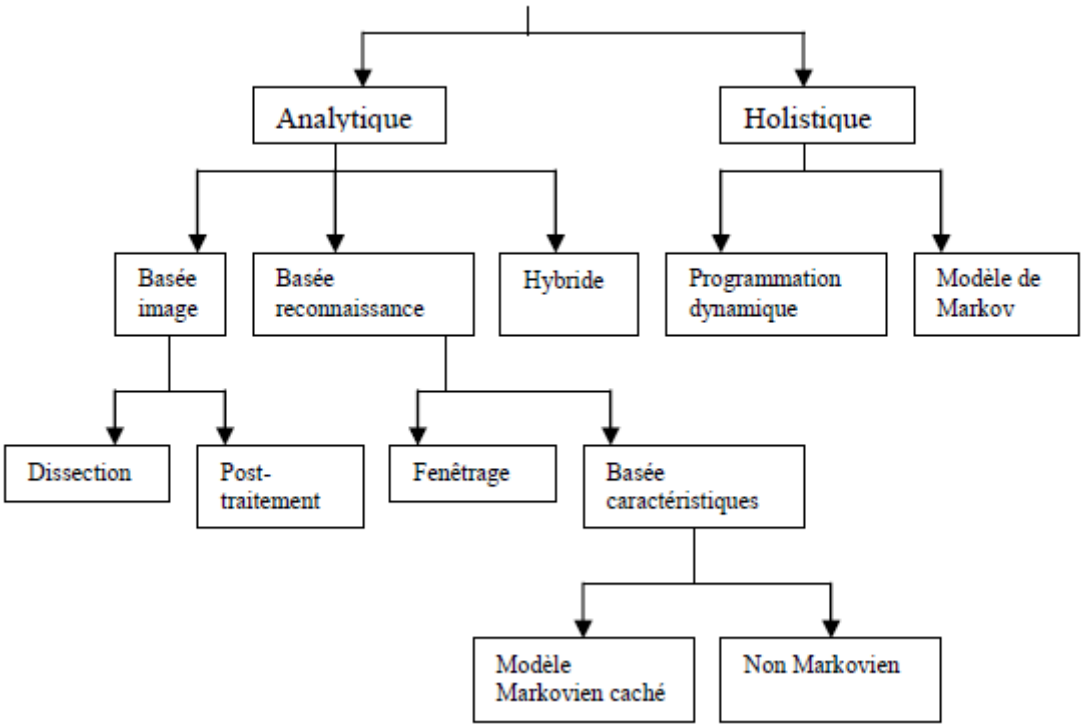


Figure 5 : Hiérarchie des méthodes de segmentation selon R.G.Casey.

ii. Technique de dissection pour segmentation

La dissection est le découpage de l'image en une séquence de sous-images en utilisant des caractéristiques générales. La dissection est un processus intelligent dans lequel on effectue une analyse de l'image sans invoquer la classification [13]. En effet dans certains documents décrivant les méthodes de segmentation où la classification n'intervient pas. La dissection est le processus entier de segmentation. Cependant dans les études actuelles la segmentation est un processus complexe alors la dissection devient un sous processus du processus de segmentation.

iii-2-1 Dissection directement en caractères :

iii-2-1-a Espace blanc et pitch :

Dans l'impression sur machine les espaces blancs servent souvent de séparateurs entre les caractères successifs. Cette propriété peut aussi être étendue à l'écriture manuelle, en fournissant des cases séparées dans lesquelles sont imprimés des symboles individuels. Ceci peut être applicable dans le cas d'applications telles que la facturation où la structure du document est spécialement conçue pour l'OCR. Dans des applications utilisant un ensemble limité de fontes, chaque caractère occupe un espace de largeur fixe. Le pitch (nombre de caractères par unité de distance

horizontale) fourni une base pour l'estimation des points de segmentation. La séquence de points de segmentation devrait approximativement être équidistante dans une distance correspondant au pitch. Cette technique est appropriée pour les textes imprimés où les caractères sont équidistants et où il y'a assez d'espace entre deux caractères adjacents [1].

iii-2-1-b Analyse des projections :

Les projections verticales (appelées aussi histogrammes verticales) d'une ligne imprimée consiste à compter les pixels noirs continus dans chaque colonne. Cela peut servir à détecter les espaces blancs entre les caractères successifs [15].
Cette technique a été utilisée comme base pour la segmentation de l'écriture non cursive. Mais elle a échoué devant les problèmes de chevauchement de caractères et d'écriture rapprochée, ou quand les traits des caractères n'ont pas la même épaisseur [16].
Cette technique a été rectifiée pour être utilisée dans la segmentation de l'écriture cursive. En effet plusieurs auteurs l'utilisent pour la segmentation de l'écriture arabe qui est de nature cursive.

iii-2-1-c Traitement des composantes connexes :

Les méthodes décrites précédemment ne sont généralement pas adéquates pour la segmentation de l'écriture manuscrite ou de fontes proportionnelles (la largeur des caractères est variable). Ce type d'écriture nécessite une analyse à deux dimensions.
Une approche banale est basée sur la détermination des régions noires connectées puis un traitement est utilisé pour combiner ou séparer ces composants en caractères.
Il existe deux types de méthodes utilisant cette technique.
La première est basée sur les boites de délimitation (bounding boxes).
La seconde est basée sur l'analyse détaillée de l'image des composants connectés [13].

Analyse les « bounding boxes » :

La distribution des boites informe beaucoup sur la segmentation des caractères non cursifs, en testant leur relation d'adjacence pour les fusionner, ou leur taille et aspect pour les séparer [13].
Cette méthode, bien que donnant des résultats éminents de point de vue vitesse de traitement et efficacité devant la méthode d'analyse des projections verticales, mais

elle reste limitée à la segmentation des caractères détachés (manuscrits ou imprimés).

Séparation des composants connectés :

Dans ce cas un traitement plus détaillé est nécessaire pour séparer les caractères joints de façon fiable. L'intersection de deux caractères peut donner une caractéristique spéciale de l'image. Par conséquent la méthode de dissection a été développée de façon à détecter ces caractéristiques et de les utiliser pour découper l'image d'une chaîne de caractères en sous-images [13].

L'algorithme de segmentation dans ce cas comporte deux modules.

Le premier s'appelle le module de pré-reconnaissance qui a pour but d'identifier les caractères connectés.

Le second module est celui de la segmentation. Il n'est activé que lors de l'identification de caractères connectés, et il a pour but de trouver des points de repère de l'image pour être considérés par la suite comme points de segmentation, rejetant ceux qui paraissent situés à l'intérieur du caractère et construire un chemin de découpage convenable [15].

iii-2-2 Dissection avec post-traitement contextuel (graphèmes) :

Dans ce cas la segmentation obtenue par dissection peut être ultérieurement soumise à une évaluation basée sur un contexte linguistique. Donc le système n'évalue pas directement les hypothèses de segmentation mais il essaye à peu prêt de corriger des segmentations incorrectes [17] [18] [19].

Cette méthode a surtout été utilisée pour la segmentation de l'écriture cursive.

Plusieurs techniques ont été proposées dont nous pouvons citer : la méthode proposée par K.M. Sayre dans [20] où une dissection en graphèmes basée sur la détection des zones caractéristiques de l'image est d'abord effectuée. Les classes reconnues par le classifieur n'étaient pas forcement des lettres, mais elles pouvaient correspondre à plus d'une lettre ou à un fragment de lettre.

Une autre méthode proposée par E. Lecolinet et J-P Crettez dans [21] où la dissection est basée sur la détection des ligatures. L'algorithme de segmentation comportait deux étapes :

1. la détection des zones de segmentation possibles

2. utilisation d'un algorithme de pré-reconnaissance, qui avait pour but non pas de reconnaître le caractère, mais d'évaluer si une sous-image définie par la pré-segmentation pouvait représenter un caractère valide.

Les zones de pré-segmentation étaient détectées en analysant le profile supérieur et inférieur et l'ouverture des concavités des mots.
Plusieurs autres méthodes sont exposées en détail dans [13].

iii. Segmentation basée reconnaissance

Les méthodes considérées ici segmentent aussi les mots en unités individuelles (généralement des lettres). Cependant le principe des opérations est complètement différent.
Ici l'algorithme de dissection n'est pas basé sur les caractéristiques, mais l'image est divisée systématiquement par le chevauchement d'un ensemble de morceaux sans tenir compte de ce qu'ils contiennent. Ils sont considérés comme essai de trouver un résultat de segmentation/reconnaissance cohérent.
Le principal avantage de ces méthodes est qu'elles évitent les problèmes de segmentation [13]. Ces méthodes sont aussi appelées «méthodes sans segmentation» (« segmentation-free methods ») [16].

iii-3-1 Les méthodes cherchant l'image :

La segmentation basée reconnaissance ici s'effectue en deux étapes.
a- génération des hypothèses de segmentation (étape de fenêtrage).
b- Choix de la meilleure hypothèse (étape de vérification).
La distinction entre les différentes méthodes réside dans la manière dont sont effectuées les deux étapes [13].
Une méthode utilisant une reconnaissance combinant la programmation dynamique et les réseaux de neurones était proposée dans [22]. Cette technique sélectionne la combinaison optimale de coupures à partir d'un ensemble prédéfini de fenêtres. A partir de cet ensemble toutes les segmentations possibles (légales) étaient construites en les combinant. Puis un graphe dont les nœuds représentaient les segments valides était créé et deux nœuds étaient reliés s'ils correspondaient à des nœuds voisins. Les chemins dans ce graphe représentaient toutes les segmentations valides du mot. A chaque nœud était assignée une distance. Le plus court chemin à travers le graphe correspondait à la meilleure reconnaissance et segmentation du mot.

iii-3-2 méthodes qui segmentent la représentation des caractéristiques de l'image :

iii-3-2-1 Modèle Markovien caché (H.M.M)

C'est une méthode probabiliste qui consiste en un ensemble d'états et les probabilités de transition entre ces états. En plus des observations faites par le système sur une image. Ces dernières sont représentées par des variables aléatoires, dont la distribution dépend de l'état. Elles constituent une représentation séquentielle des caractéristiques de l'image d'entrée. Ces caractéristiques peuvent représenter :
- la variation du langage lettre par lettre.
- La transition à l'intérieur du caractère état par état.
- La variation dans le mot, état par état dans l'ensemble de mots admissibles d'un lexique.

iii-3-2-2 Approches non Markoviennes :

Ces méthodes se sont inspirées des concepts utilisés dans la reconnaissance d'objets. Ici différentes caractéristiques et leurs positions d'occurrences sont enregistrées par image. Chaque caractéristique contribue à l'évidence de l'existence d'un ou plusieurs caractères dans une position d'occurrence [16]. Un calcul est effectué à une position donnée pour servir de score pour la classification, puis ces scores sont soumis à un traitement contextuel utilisant un lexique prédéfini, Dans le but de reconnaître les mots.
Cette méthode est utilisée pour la reconnaissance de textes imprimés dans une fonte connue. [13]
Elle est très répondue dans la reconnaissance des textes de nature connectés tels que : le Chinois, le Japonais, le Thaï ...
Une autre méthode qui reconnaît les graphes caractéristiques du mot est basée sur la comparaison des sous-graphes de caractéristiques avec des prototypes de caractères prédéfinis. La reconnaissance est effectuée en cherchant le chemin qui donne la meilleure interprétation des caractéristiques du mot. Les caractères sont classés par ordre de la qualité de comparaison.

iv. Stratégies mixtes (sur-segmentation)

Cette famille de méthodes utilise aussi la pré-segmentation, mais d'une façon pas aussi stricte que dans l'approche par graphèmes. Un algorithme de dissection est

appliqué à l'image, suffisamment pour que les limites de segmentation soient incluses dans la coupure (i.e. : chaque coupure représente un caractère ou une partie du caractère et pas plus). La segmentation optimale est définie par un sous-ensemble de coupures. Chaque sous-ensemble donne des hypothèses de segmentation qui sont évaluées lors de la classification, pour en choisir la segmentation la plus prometteuse [13].

Cette technique a été utilisée pour la segmentation des mots arabes.

Des références employant cette technique [23], [24], [25], [26] et [27].

v. Stratégies holistiques

Un processus holistique reconnaît un mot entier comme entité. Un inconvénient majeur de ce type de méthodes est que leur utilisation est toujours retreinte à un lexique limité et prédéfini. Ces méthodes conviennent mieux aux applications où le lexique est défini statiquement. Telles que la reconnaissance des chèques ou la reconnaissance en ligne des commandes d'ordinateurs pour des applications industrielles ou sur ordinateur personnel.

La plus part des algorithmes de méthodes holistiques suivent un schéma à deux étapes :

1- effectuer une extraction de caractéristiques.

2- Reconnaissance globale en comparant la représentation du mot avec les mots de référence stockés dans une bibliothèque.

4. Extraction des caractéristiques

C'est aussi l'une des étapes les plus importantes en OCR [1].

La reconnaissance d'un caractère passe d'abord par l'analyse de sa forme et l'extraction de ses traits caractéristiques (primitives) qui seront exploités pour son identification.

Les types de caractéristiques peuvent être classés en quatre groupes principaux [28] [6] :

- caractéristiques structurelles
- caractéristiques statistiques
- transformations globales
- superposition des modèles et corrélation

a. Caractéristiques structurelles :

Les caractéristiques structurelles décrivent une forme en terme de sa topologie et sa géométrie en donnant ses propriétés globales et locales.
Parmi ces caractéristiques on peut citer [28]:

- Les traits et les anses dans les différentes directions ainsi que leurs tailles.
- Les points terminaux.
- Les points d'intersections.
- Les boucles.
- Le nombre de points diacritiques et leur position par rapport à la ligne de base.
- Les voyellations et les zigzags (hamza).
- La hauteur et la largeur du caractère.
- La catégorie de la forme (partie primaire ou point diacritique, …).

Plusieurs autres caractéristiques peuvent être tirées, suivant qu'elles soient extraites d'une courbe, un trait ou un segment de contour.

b. Les caractéristiques statistiques :

Les caractéristiques statistiques décrivent une forme en terme d'un ensemble de mesures extraites à partir de cette forme. Les caractéristiques utilisées pour la reconnaissance de textes arabes sont : le zonage (zonning), les caractéristiques de lieu géométrique (Loci) et les moments [28].

- Le zonage consiste à superposer une grille $n \times m$ sur l'image du caractère et pour chacune des régions résultantes, calculer la moyenne ou le pourcentage de points en niveaux de gris, donnant ainsi un vecteur de taille $n \times m$ de caractéristiques.
- La méthode Loci est basée sur le calcul du nombre de segments blancs et de segments noirs le long d'une ligne verticale traversant la forme, ainsi que leurs longueurs [6].
- La méthode des moments : les moments d'une forme par rapport à son centre de gravité sont invariants par rapport à la translation et peuvent être invariants par rapport à la rotation [Al-Badr 94]. Ils sont aussi indépendants de l'échelle. Ces caractéristiques peuvent être facilement et rapidement extraites d'une image de texte, ils peuvent tolérer modérément les bruits et les variations. Une lecture détaillée sur les moments se trouve dans [Tsang 00].

c. Les transformations globales :

La transformation consiste à convertir la représentation en pixels en une représentation plus abstraite pour réduire la dimension des caractères, tout en conservant le maximum d'informations sur la forme à reconnaître.
Une des transformations les plus simples est celle qui représente le squelette ou le contour d'un caractère sous forme d'une chaîne de codes de directions [6]. La chaîne de code obtenue est souvent simplifiée pour réduire les redondances et les changements brusques de direction.

d. Superposition des modèles (template matching) et corrélation:

La méthode de «template matching » appliquée à une image binaire (en niveaux de gris ou squelettes), consiste à utiliser l'image de la forme comme vecteur de caractéristiques pour être comparé à un modèle (template) pixel par pixel dans la phase de reconnaissance, et une mesure de similarité est calculée [28].

5. Classification

La classification dans un système OCR regroupe deux tâches : l'apprentissage et la reconnaissance et décision.
A cette étape les caractéristiques de l'étape précédente sont utilisées pour identifier un segment de texte et l'attribuer à un modèle de référence [28].

a. L'apprentissage :

Il s'agit lors de cette étape d'apprendre au système les propriétés pertinentes du vocabulaire utilisé et de l'organiser en modèles de références.
L'idéal serait d'apprendre au système autant d'échantillons que de formes d'écritures différentes, mais cela est impossible à cause de la grande variabilité de l'écriture qui conduirait à une explosion combinatoire de modèles de représentation. La tendance consiste alors à remplacer le nombre par une meilleure qualité des traits caractéristiques [29], [6].
L'apprentissage consiste en deux concepts différents : l'entraînement et l'adaptation.

L'entraînement consiste à enseigner au système la description des caractères tandis que l'adaptation sert à améliorer les performances du système en profitant des expériences précédentes. Certains systèmes permettent à l'utilisateur d'identifier un caractère lorsqu'ils échouent à le reconnaître et ils utilisent l'entrée de l'utilisateur à chaque fois que le caractère est rencontré [6].

Les procédés d'apprentissage sont différents selon qu'il s'agisse de reconnaissance de caractères imprimés ou manuscrits ou de reconnaître des textes mono-fonte ou multi-fonte.

D'une manière générale, on distingue deux types de techniques d'apprentissage : supervisé et non supervisé.

- L'apprentissage est dit supervisé s'il est guidé par un superviseur appelé professeur. Il est réalisé lors d'une étape préliminaire de reconnaissance en introduisant un grand nombre d'échantillons de référence. Le professeur indique dans ce cas le nom de chaque échantillon. Le choix des caractères de référence est fait à la main en fonction de l'application. Le nombre d'échantillons peut varier de quelques unités à quelques dizaines, voir même quelques centaines par caractère [29], [28].
- L'apprentissage non supervisé ou sans professeur consiste à doter le système d'un mécanisme automatique qui s'appuie sur des règles précises de regroupement pour trouver les classes de référence avec une assistance minimale. Dans ce cas les échantillons sont introduits en un grand nombre par l'utilisateur sans indiquer leur classe [29].

b. Reconnaissance et décision :

La décision est la dernière étape de la reconnaissance. A partir de la description en paramètres du caractère traité, le module de reconnaissance cherche parmi les modèles de référence en présence, ceux qui lui sont les plus proches.

La reconnaissance peut conduire à un succès si la réponse est unique (un seul modèle répond à la description de la forme du caractère). Elle peut conduire à une confusion si la réponse est multiple (plusieurs modèles correspondent à la description). Enfin elle peut conduire à un rejet de la forme si aucun modèle ne correspond à sa description.

Dans les deux premiers cas, la décision peut être accompagnée d'une mesure de vraisemblance, appelée aussi score ou taux de reconnaissance [29].

Les approches de reconnaissance peuvent être regroupées en trois groupes principaux : l'approche statistique, l'approche structurelle, l'approche stochastique et l'approche hybride.

i. Approche statistique :

Elle est fondée sur l'étude statistique des mesures que l'on effectue sur les formes à reconnaître. L'étude de leur répartition dans un espace métrique et la caractérisation statistique des classes, permettent de prendre une décision de reconnaissance du type «plus forte probabilité d'appartenance à une classe» [29].

Les approches statistiques bénéficient des méthodes d'apprentissage automatique qui s'appuient sur des bases théoriques fondées, telles que la théorie de la décision bayésienne, les méthodes de classification non supervisées … En reconnaissance, le problème revient à affecter une forme inconnue à l'une des classes obtenues pendant l'apprentissage [6].

Nous pouvons citer trois méthodes statistiques parmi celles les plus couramment utilisées :

- L'approche bayésienne consiste à choisir parmi un ensemble de caractères, celui pour lequel la suite de primitives extraites a la plus forte probabilité à posteriori par rapport aux caractères préalablement appris [30].
- L'algorithme KNN (K Nearest Neighbors) affecte une forme inconnue à la classe de son plus proche voisin, en la comparant aux formes stockées dans une classe de références nommée prototypes. Il renvoie les K formes les plus proches de la forme à reconnaître suivant un critère de similarité. Une stratégie de décision permet d'affecter des valeurs de confiance à chacune des classes en compétition et d'attribuer la classe la plus vraisemblable (au sens de la métrique choisie) à la forme inconnue [29] et [31].
  Cette méthode présente l'avantage d'être facile à mettre en œuvre et fournit de bons résultats. Son principal inconvénient est lié à la faible vitesse de classification due au nombre important de distances à calculer.
- Un réseau de neurones est un graphe orienté pondéré. Les nœuds de ce graphe sont des automates simples appelés neurones formels. Les neurones sont dotés d'un état interne, l'activation, par lequel ils influencent les autres neurones du réseau. Cette activité se propage dans le graphe le long d'arcs pondérés appelés

liens synaptiques [32].

En OCR, les *primitives* extraites sur une image d'un caractère (ou de l'entité choisie) constituent les entrées du réseau. La sortie activée du réseau correspond au *caractère reconnu*. Le choix de l'architecture du réseau est un compromis entre la complexité des calculs et le taux de reconnaissance [33].

ii. Approche structurelle :

Les méthodes structurelles reposent sur la structure physique des caractères.
Elles cherchent à trouver des éléments simples ou primitifs, et à décrire leurs relations. Les primitives sont de type topologiques telles que : une boucle, un arc… et une relation peut être la position relative d'une primitive par rapport à une autre [30], [15].
Parmi les méthodes structurelles nous pouvons citer :

- Les méthodes de tests : elles consistent à appliquer sur chaque caractère traité des tests de plus en plus fins sur la présence ou l'absence de primitives, de manière à répartir les échantillons en classes. Le processus le plus habituel consiste à diviser à chaque test l'ensemble des choix en deux jusqu'à n'obtenir qu'une seule forme correspondant au caractère entré. Ce choix dichotomique est très rapide et très simple à mettre en œuvre, mais il est très sensible aux variations du tracé [29].
- La comparaison de chaînes : les caractères sont représentés par des chaînes de primitives. La comparaison du caractère traité avec le modèle de référence, consiste à mesurer la ressemblance entre les deux chaînes et à se prononcer sur celui-ci. La mesure de ressemblance peut se faire par calcul de distance ou par examen de l'inclusion de toute ou une partie d'une chaîne dans l'autre [29].
- L'approche syntaxique : en représentation syntaxique, chaque caractère est représenté par une phrase dans un langage où le vocabulaire est constitué de primitives. Les caractères d'une même famille sont représentés par une grammaire. La reconnaissance consiste à déterminer si la phrase de description du caractère peut être générée par la grammaire. L'inconvénient de cette méthode est l'absence d'algorithmes efficaces pour l'inférence grammaticale directe

[29].

iii. Approche stochastique

Contrairement aux méthodes précédemment décrites, l'approche stochastique utilise un modèle pour la reconnaissance, prenant en compte la grande variabilité de la forme. La distance communément utilisée dans les techniques de «comparaison dynamique» est remplacée par des probabilités calculées de manière plus fine par apprentissage. La forme est considérée comme un signal continu observable dans le temps à différents endroits constituant des états «d'observations». Le modèle décrit ces états à l'aide de probabilités de transitions d'états et de probabilités d'observation par état. La comparaison consiste à chercher dans ce graphe d'états, le chemin de probabilité forte correspondant à une suite d'éléments observés dans la chaîne d'entrée. [29].

Ces méthodes sont robustes et fiables du fait de l'existence d'algorithmes d'apprentissage efficaces [34]. Si l'apprentissage est lent, la reconnaissance est par contre très rapide car les modèles comprennent généralement peu d'états et le calcul est relativement immédiat. Les méthodes les plus répondues dans cette approche sont les méthodes utilisant les modèles de Markov cachés (HMM).

iv. Approche hybride

Pour améliorer les performances de reconnaissance, la tendance aujourd'hui est de construire des systèmes hybrides qui utilisent différents types de caractéristiques, et qui combinent plusieurs classifieurs en couches.

6. Post-traitement

L'objectif du post-traitement est l'amélioration du taux de reconnaissance des mots (par opposition au taux de reconnaissance du caractère). Cette phase est souvent implémentée comme un ensemble d'outils relatifs à la fréquence d'apparition des caractères dans une chaîne, aux lexiques et à d'autres informations contextuelles.

Comme la classification peut aboutir à plusieurs candidats possibles, le post-traitement a pour objectif d'opérer une sélection de la solution en utilisant des niveaux d'informations plus élevés (syntaxiques, lexicale, sémantiques…).

Le post-traitement se charge également de vérifier si la réponse est correcte (même si elle est unique) en se basant sur d'autres informations non disponibles au classifieur. [1]

CHAPITRE II : L'OCR ET LA LANGUE ARABE

I. Introduction

La reconnaissance de l'écriture arabe est un domaine de recherche relativement récent et qui a connu ces dernières années des progrès remarquables. Il présente un intérêt indéniable dans l'accomplissement de tâches considérées fastidieuses dans certains domaines comme le tri postal, la lecture de chèques bancaires, la lecture des bordereaux, … [35]

II. Caractéristiques de l'alphabet arabe

L'écriture arabe a vu le jour aux alentours du VIème siècle avant l'apparition de l'écriture cursive nabatéenne[1], et s'est progressivement répandue avec l'existence de l'Islam et la révélation coranique. [35]

L'arabe est une écriture consonantique qui utilise un alphabet de 28 lettres (Tableau 1) auquel il faut ajouter la Hamza « ء », qui est le plus souvent considérée comme signe complémentaire [36]. La hamza « ء » a une orthographe spéciale qui dépend de règles grammaticales, ce qui multiplie les formes nécessaires à sa représentation, puisqu'elle peut s'écrire seule ou sur le support de trois voyelles (alif, waw et ya) dont elle suit le code (Tableau 2).

De plus l'alphabet arabe comprend d'autres caractères additionnels tels que « ة» et « لا», de ce fait, certains auteurs considèrent que l'alphabet arabe comprend plutôt 31 lettres que 29.

Caractère	Nom Unicode	fin	milieu	début
ا	alef	ـا	ـا	ا
ب	beh	ـب	ـبـ	بـ

[1] Les Nabatéens (en arabe : الأنباط) sont un peuple commerçant de l'Antiquité vivant au sud de la Jordanie et de Canaan, et au nord de l'Arabie actuelle.

ت	teh	ـت	ـتـ	تـ
ث	theh	ـث	ـثـ	ثـ
ج	jeem	ـج	ـجـ	جـ
ح	hah	ـح	ـحـ	حـ
خ	khah	ـخ	ـخـ	خـ
د	dal	ـد	ـد	د
ذ	thal	ـذ	ـذ	ذ
ر	reh	ـر	ـر	ر
ز	zain	ـز	ـز	ز
س	seen	ـس	ـسـ	سـ
ش	sheen	ـش	ـشـ	شـ
ص	sad	ـص	ـصـ	صـ
ض	dad	ـض	ـضـ	ضـ
ط	Tah	ـط	ـطـ	طـ
ظ	zah	ـظ	ـظـ	ظـ
ع	ain	ـع	ـعـ	عـ
غ	ghain	ـغ	ـغـ	غـ
ف	feh	ـف	ـفـ	فـ
ق	qaf	ـق	ـقـ	قـ
ك	kaf	ـك	ـكـ	كـ
ل	lam	ـل	ـلـ	لـ
م	meem	ـم	ـمـ	مـ
ن	noon	ـن	ـنـ	نـ

ه	heh	ـه	ـهـ	هـ
و	waw	ـو	ـو	و
ي	yeh	ـي	ـيـ	يـ

Tableau 1: Les caractères arabes isolés, au début, au milieu et la fin du mot

Caractère	Initiale	Médiane	Finale	Isolé
Alif+~			ـآ	آ
Alif + ء			ـأ	أ
			ـإ	إ
Waw + ء			ـؤ	ؤ
Ya + ء	ـئـ	ـئـ	ـئ	ئ

Tableau 2 : Les positions qu'occupe Hamza en association avec Alif, Waw et Ya.

La considération du symbole « ~» qui s'écrit uniquement sur le support du caractère « ا», fait apparaître d'autres graphismes (Tableaux 2).

Les principales caractéristiques de la langue arabe sont :

Un trait caractéristique de l'écriture arabe est la présence d'une ligne de base horizontale dite encore lige de référence ou d'écriture. C'est le lieu des caractères d'une même chaîne (figure 6).

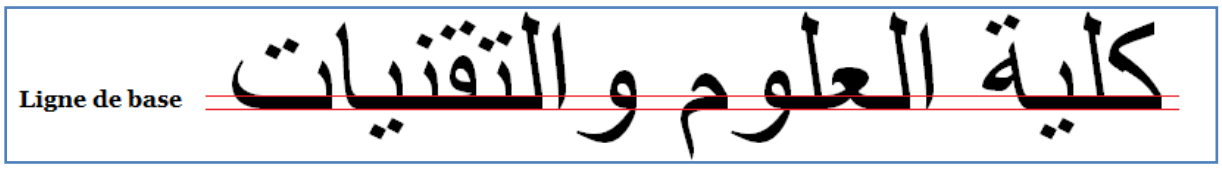


Figure 6 : Phrase arabe montrant la ligne de base

Les caractères arabes s'écrivent de façon cursive, de droite vers la gauche, aussi bien dans le cas de l'imprimé que du manuscrit.

Les dimensions des caractères (chasse et hauteur) sont variables, même s'il s'agit des différentes formes d'un caractère. (Tableau 1)

La forme d'une lettre écrite dépend de son contexte et le dessin du glyphe associé diffère selon que le caractère apparaît en position initiale, médiane ou

isolée dans une chaîne de caractères (Tableau 3). A chaque caractère peut correspondre jusqu'à quatre glyphes différents ce qui lève à environ 100 le nombre de formes à reconnaître. Les formes correspondantes à un même caractère, souvent appelées « formes internes », présentent parfois de sensibles différences ; dans certains cas, il est même difficile d'en déduire s'il s'agit d'une même lettre. Cependant le codage ASMO attribue un seul code pour les différentes formes d'un même caractère, contrairement au latin où le code ASCII prévoit deux codes différents pour la même lettre dans sa forme majuscule et minuscule [29].

Initiale	Médiane	Finale	Isolé
عـلوم	مـعهد	مـع	ودع
ه ـندام	مـهندس	عقلـه	يده

Tableau 3 : Les quatre formes des caractères «ع» et « ه »

Plus de la moitié des caractères arabes (16) incluent dans leur forme des points qui peuvent être au nombre de 1, 2 ou 3. Ces points peuvent se situer au dessus ou en dessous du corps du caractère, mais jamais en haut et en bas simultanément [6].

Certains caractères ne peuvent être rattachés à leur gauche et de ce fait ne peuvent se trouver qu'en position isolée ou finale, par conséquence il existe des mots composés de plusieurs parties qu'il est convenu d'appeler généralement PAW (peace of arabic word) ou encore pseudo-mot [6].
Un PAW correspond donc à une chaîne d'un ou de plusieurs caractères (Tableau **4**). L'écriture arabe est ainsi semi-cursive plutôt que totalement cursive.

5 PAWs	4 PAWs	3 PAWs	2 PAW	1 PAW
الفردوس، الأزهار	الورد، الدينار	معلوميات، أستاذ	تقنيات، فاس، علوم	يد، كلية

Tableau 4 : Exemples des mots composés de 1, 2, 3, 4 et 5 PAWs

Le mot arabe n'a pas de longueur fixe, il peut comprendre un ou plusieurs PAWs incluant chacun un nombre différent de caractères. De plus, différentes chasses possibles peuvent être associées à un même mot, en insérant un nombre variable de traits d'allongement.

Dans certaines fontes plusieurs caractères peuvent être écrits de façon combinée. Ces combinaisons ou ligatures, dont le nombre dépasse 1500, sont optionnelles contrairement aux ligatures horizontales qui sont obligatoires [29]. Les ligatures verticales sont utilisées pour des raisons d'esthétique. Elles dépendent du dessin de la police et du degré de qualité artistique du document. Elles peuvent être formées de deux, trois ou quatre caractères et peuvent prendre plusieurs significations selon l'emplacement des points. On parle souvent de ligature de niveau «n» où n désigne le nombre de caractères ligaturés. Les ligatures verticales, souvent composées de façon particulière, peuvent avoir lieu soit au début ou à la fin du PAW. La ligature classique de niveau 2 peut avoir lieu avec les couples de caractères donnés dans le tableau **5**.

لـ، مـ	مـ، جـ	فـ، جـ	قـ، جـ
لـ، جـ	مـ، حـ	فـ، حـ	قـ، حـ
لـ، حـ	مـ، خـ	فـ، خـ	قـ، خـ
لـ، خـ			

Tableau 5 : Caractères susceptibles d'être ligaturés verticalement selon [29].

Les caractères arabes peuvent être voyellées. Les voyelles appelées aussi diacritiques dans certains documents tels que [6] et [37] et courtes voyelles dans d'autres tels que [31], peuvent se placer au dessus ou en dessous du caractère. Les voyelles sont d'une invention postérieure aux consonnes. Dans l'arabe contemporain ordinaire, on écrit seulement les consonnes et les voyelles longues. Un même mot avec différentes voyelles courtes peut être compris comme verbe, nom ou adjectif ...

A titre d'exemple « علم » peut signifier « drapeau : عَلَمٌ » ou « savoir : عِلْمٌ » ou encore « enseigner :عَلَّمَ », selon sa voyellation.

Il existe 8 signes de voyellation qui peuvent se placer au dessus de la ligne d'écriture, tels que

fathah (◌َ)

dhammah (ُ)
soukoun (ْ)
chaddah (ّ) qui doit être accompagnée de l'une des vayellations fatha, Dammah ou kasrah, en dessous tels que Kasrah (ِ).
trois « tanwin » peuvent être formés à partir d'un double fatha (ً), d'un double dhammah (ٌ) ou d'un double kasrah (ٍ)

III. Données graphiques de l'alphabet arabe

L'alphabet arabe n'a qu'un système d'écriture dans lequel les lettres sont liées ou ne sont pas liées entre elles selon des règles précises. Il existe différents styles d'écriture, mais dans aucun d'eux il est possible de juxtaposer des lettres totalement isolées les unes des autres.
Il n'y a pas de lettres d'imprimerie en arabe, il n'y a que des caractères typographiques copiés de l'écriture manuscrite. Le caractère arabe est en effet dessiné non pas en fonction des contraintes géométriques des procédés de composition pour imprimerie, mais en fonction de la main et d'une esthétique visuelle héritée de la calligraphie. La fonctionnalité et la lisibilité sont sacrifiées à l'esthétique calligraphique qui substitue l'élégance à la clarté [29].

IV. Ocr arabe (AOCR)

La reconnaissance l'écriture arabe (AOCR : Arabic OCR) remonte aux années 70, depuis, plusieurs solutions ont été proposées. Elles sont aussi variées que celles utilisées dans le latin.
Dés les premiers travaux de reconnaissance de l'écriture arabe, les deux modes de reconnaissance, statique et dynamique ont été considérés [29]. L'intérêt a été d'autant porté sur les travaux dans le domaine de l'écriture manuscrite que l'écriture imprimée. Cependant les travaux en-ligne restent relativement peu nombreux.
Le tableau 6 (tiré de [6] et [29]), regroupe certains systèmes de reconnaissance de l'écriture arabe en précisant pour chacun le mode utilisé en-ligne ou hors-ligne,

l'approche de reconnaissance globale ou analytique, le type de segmentation, la représentation choisie ainsi que les scores réalisés.

Référence	Système	Approche	Segmen-tation	Primitives	Classification	Perform-ance
[38]	Hors-ligne, imprimé MF	Analytique	Externe	Structurelle / Statistique	Structurelle / Statistique / arabe de décision	RC 99%
[39]	Hors-ligne, Imprimé	Analytique	Externe	Dimension du graphème	Pré classification / mise en correspondance / reconstruction	RC 96%
[40]	Hors-ligne, manuscrit	Analytique	Externe	Structurelle	Transformation Off-line /on-line	-
[36]	Hors-ligne, imprimé	Globale	-	Structurelle	Mise en correspondance spatiale de modèles de primitives	RM 73.13-99.39%
[41]	En-ligne, PAWs	Analytique	Externe	Structurelle	Arabe de décision	RM 86-100%
[42]	Hors-ligne, manuscrit	Analytique	Externe	Structurelle	Structurelle	-
[43]	Hors-ligne, MF	Analytique	Externe	Statistiques	Réseaux de neurones	RC 64 – 100%
[44]	Hors-ligne, MF	Analytique	Externe	Statistiques	Distance quadratique	RC 87.87 – 95.24%
[45]	En-ligne, caractères Isolés	-	-	Chaînes de codes	Programmation dynamique	RC 95%
[46]	Hors-ligne, manuscrit Mot	Analytique	Externe	Statistiques	Syntaxique/ Distance	RC 91%
[47]	Hors-ligne, manuscrit.	Analytique	Externe	moments	Classifieur bayesien.	RC 99.5 %

[48]	Hors-ligne, manuscrit MS	Analytique	Externe	structurelles	Arbre de décision	SC 98.9 % RC 83 %
[49]	Hors-ligne, manuscrit.	Globale	-	structurelles	Dictionnaire	-
[50]	Hors-ligne, manuscrit.	Analytique	Externe	Structurelles/ statistiques	KNN	RC 82.5 %
[51]	Hors-ligne, MF	Analytique	Externe	Chaîne de codes	Arbre de décision	RC 90 %
[52]	Hors-ligne, caractères MS	-	-	Structurelles	Réseaux de neurones	RC 90-92 %
[53]	Hors-ligne, mots	Globale	-	Structurelles	Réseaux de neurones	RC 98%
[54]	Hors-ligne, Perse	Analytique	Externe	-	-	SC 93-98.9 %
[55]	Hors-ligne	Analytique	Externe	géométriques	-	SC 99-100 %
[56]	Hors-ligne, caractères isolés	-	-	Structurelles/ variables linguistiques	Logique floue	RC 100 %
[57]	Hors-ligne, caractères isolés	Analytique	Externe	Fuzzy linguistiques	Logique floue	RC 100%
[85]	Hors-ligne, imprimé	-	Interne-SWS	Moments	Table de correspondance	RC 94 %
[37]	Hors-ligne, imprimé	Analytique	Externe	-	Grammaire reguliere	RC 93.4 %
[59]	Hors-ligne, imprimé	Analytique	Externe	moments	Distance	RC 95-100 %
[60]	Hors-ligne, imprimé	Analytique	Externe	Descripteurs de Fourier	Classifieur Topologique	RC 99 %

[61]	En-ligne, caractères isolés	-	-	Structurelles	Arbre « handcrafted »	RC 99.6 %
[62]	Hors-ligne, MF	Analytique	Externe	Structurelles/ statistiques	Programmation Dynamique	RC 98 %
[63]	Hors-ligne, manuscrit	Analytique	Externe	Structurelles	Réseaux de neurones/ HMM	-
[64]	Hors-ligne, imprimé	Analytique	Externe	Structurelles	Réseaux de neurones	RC 89-93.1 %
[65]	Hors-ligne, imprimé	Analytique	Externe	Chaîne de codes	Sruct./ Mesure géom./ contexte	RC 95.87 %
[66]	Hors-ligne, imprimé	-	Externe	Structurelles	Syntaxique	RC 99 %
[67]	Hors-ligne, imprimé	Analytique	Externe	Structurelles	Réseaux de neurones	RC 99 %
[68]	Hors-ligne, manuscrit	Analytique	Externe	Structurelles	Dictionnaire	SC 95 %
[69]	Hors-ligne, imprimé MF	Analytique	Externe	Structurelles	Morphologie mathématique	RC 98 %
[70]	En-ligne, Caractères Isolés	-	-	Statistiques	HMMs	RC 98.1 %
[71]	En-ligne, MS	Globale	-	statistiques	DHMMs & NSHMMs	RC 90-93.5 % 1S RC 86-90 % MS
[72]	Hors-ligne, Manuscrit	Analytique	Externe	Structurelles	-	SC 98.52 %
[73]	Hors-ligne, Manuscrit	Analytique	Externe	Topologiques/ Statistiques	HMMs Distance	RC 79.5-82.5 %
[74]	Hors-ligne, manuscrit	Analytique	Externe	Structurelles	Mathématique	SC 81.88 %

[75]	Hors-ligne, Manuscrit Ms	Analytique	Externe	Chaîne de codes	-	SC 97.41 %
[33]	Hors-ligne, manuscrit	Analytique	Externe	Statistiques/ Structurelles	Réseaux de neurones	RC76.17-85.75%
[76]	Hors-ligne, imprimé	Analytique	Externe	Chaîne de codes	Rés. Neurones/ arbres	RC R.N 89.06% RC AR 90.68%
[77]	Hors-ligne, manuscrit	Analytique	Externe	Structurelles	Dictionnaire	RC 86 %
[78]	Hors-ligne, manuscrit	Analytique	Externe	Structurelles	Mise en correspondance	RC 87 %
[31]	Manuscrit	Globale	-	KNN, moments	-	RC 94%
[79]	Hors-ligne, Imprimé	Analytique	Externe	-	-	SC 97.01 %
[80]	Hors-ligne, Manuscrit	Analytique	Externe	Géométriques	Réseaux de neurones	RC 69.7 %
[81]	Imprimé, caract. isolés	Analytique	Externe	Morphologiques / Statistiques	Logique floue	-

Tableau 6 : Tableau récapitulatif précisant les caractéristiques et les performances de certains systèmes AOCR.

RC : Taux de reconnaissance caractère, RM : Taux de reconnaissance mot, SC : taux de segmentation de caractères, MF : Multifonte, MS : Multiscripteur

CHAPITRE III : CONTRIBUTION A LA RECONNAISSANCE D'ECRITURE ARABE

Dans ce chapitre, nous allons présenter un système de reconnaissance d'écriture arabe imprimé mono-fonte, qui permet de reconnaitre un texte à partir d'une image de texte.

D'abord l'image de texte traitée subit des prétraitements nécessaires, puis elle est segmentée en lignes, mots et en caractères.

Le système reconnait chaque caractère, en comparant l'image du caractère segmenté avec les images de la base de données (avec un indice de dissimilarité), ainsi en concaténant les caractères retournés nous obtenons le texte reconnu.

Une étape finalement de post traitement augmente le taux de la reconnaissance du système.

I. Corpus

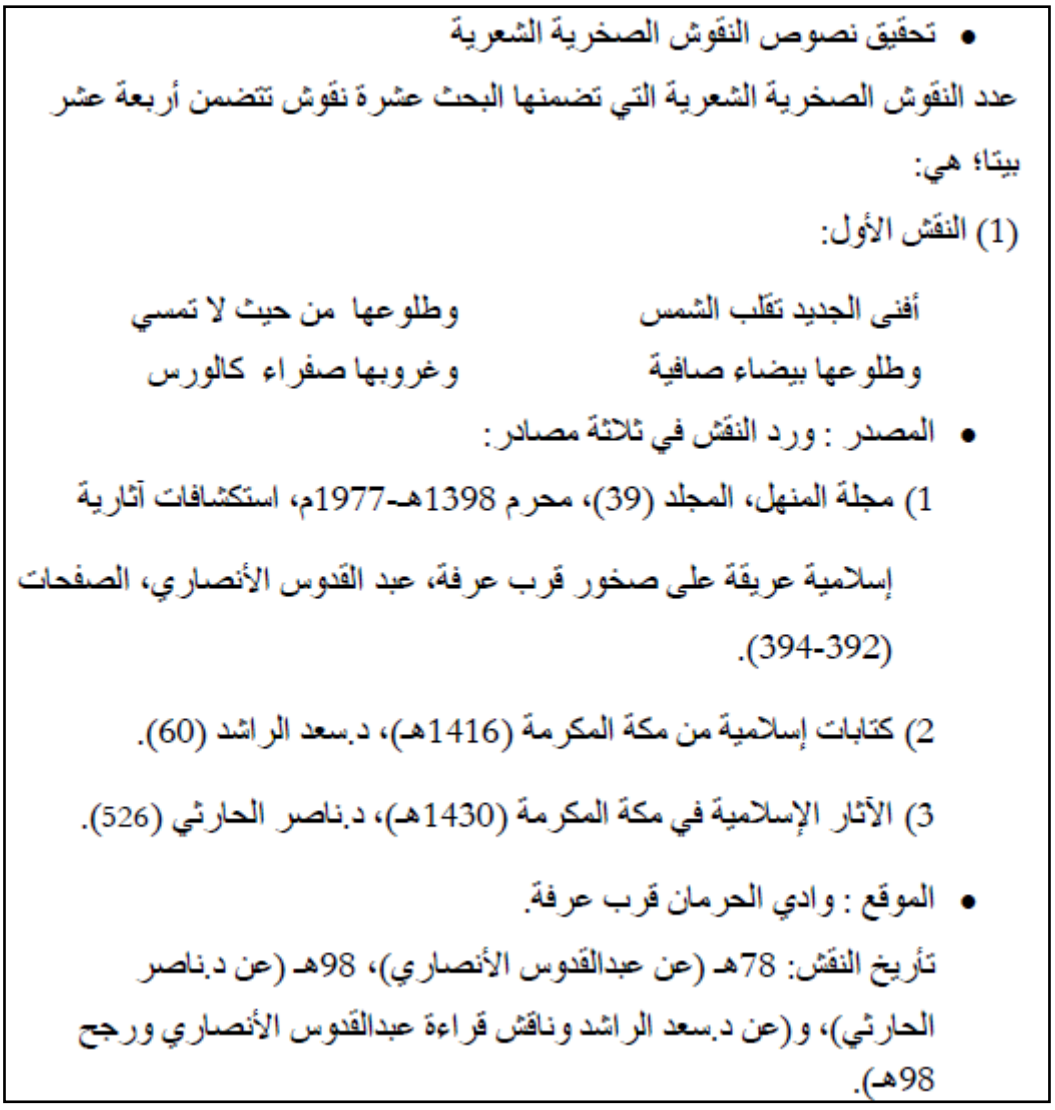
Nous avons travaillé sur un livre en format PDF (Figure 7), intitulé « **النقوش الشعرية الصخرية في المملكة العربية السعودية وقيمتها الأدبية** » qu'on peut traduire en français à : Gravures rupestres poétiques dans le Royaume d'Arabie Saoudite et leur valeur littéraire.

Figure 7 : Couverture du livre «Gravures rupestres poétiques»

Caractéristiques du texte :

- ✓ Police : Uthman Taha Naskh

- ✓ Taille de la police : 16
- ✓ Couleur : noir
- ✓ Ne contient pas de signes de voyellation
- ✓ Ne contient pas de ligatures verticales
- ✓ Contient des chiffres, points de ponctuation, puces de numérotation et des caractères spéciaux.



- تحقيق نصوص النقوش الصخرية الشعرية

عدد النقوش الصخرية الشعرية التي تضمنها البحث عشرة نقوش تتضمن أربعة عشر بيتا؛ هي:

(1) النقش الأول:

أفنى الجديد تقلب الشمس وطلوعها من حيث لا تمسي

وطلوعها بيضاء صافية وغروبها صفراء كالورس

- المصدر : ورد النقش في ثلاثة مصادر:

1) مجلة المنهل، المجلد (39)، محرم 1398هـ-1977م، استكشافات آثارية إسلامية عريقة على صخور قرب عرفة، عبد القدوس الأنصاري، الصفحات (392-394).

2) كتابات إسلامية من مكة المكرمة (1416هـ)، د.سعد الراشد (60).

3) الآثار الإسلامية في مكة المكرمة (1430هـ)، د.ناصر الحارثي (526).

- الموقع : وادي الحرمان قرب عرفة.

تأريخ النقش: 78هـ (عن عبدالقدوس الأنصاري)، 98هـ (عن د.ناصر الحارثي)، و(عن د.سعد الراشد ونافش قراءة عبدالقدوس الأنصاري ورجح 98هـ).

Figure 8 : Exemple d'une page du document

II. Prétraitement de l'image

L'entrée du système est une image capturée du document PDF, elle peut être constituée d'un seul caractère, d'un mot, plusieurs mots ou même de plusieurs lignes.

1. Chargement de l'image

L'image d'entrée est chargée puis elle est convertie en image niveau de gris.

2. Binarisation

Nous avons fait une simple binarisation à seuillage, puisque dans notre cas l'image n'est pas bruitée et qu'elle ne contient que du blanc et les niveaux de gris.

3. Segmentation en lignes

Dans la segmentation en lignes, nous avons implémenté la méthode de la projection horizontale modifiée, en effet nous avons utilisé un tableau dont la taille est égale à la hauteur de l'image. Si la ligne i de l'image contient au moins un point noir, alors t[i]=1 sinon t[i]=0 ;
Ainsi, si nous trouvons un nombre de zéros successifs qui dépasse un seuil (3 dans notre cas) alors il s'agit d'un saut de ligne. (Figure 9)

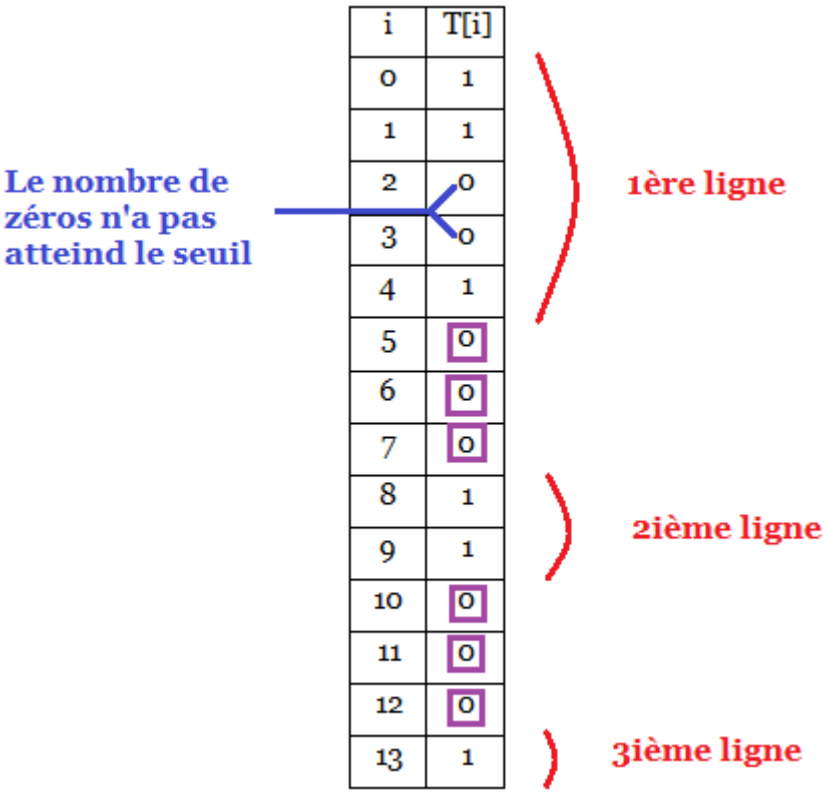


Figure 9 : Détermination des lignes

Quatre données sont sauvegardées : (Figure 10)

- ✓ Le nombre des lignes du texte nombreLignes

- ✓ L'image qui constitue chaque ligne
- ✓ Un tableau débutLigne de taille égale au nombreLignes, avec débutLigne[i] = l'indice où commence la ligne numéro i.
- ✓ Un tableau finLigne de taille égale au nombreLignes, avec finLigne[i] = l'indice où termine la ligne numéro i.

i	T[i]
0	1
1	1
2	0
3	0
4	1
5	0
6	0
7	0
8	1
9	1
10	0
11	0
12	0
13	1

nombreLigne =3

débutLigne[] = [0,8,13]

finLigne[] = [4,9,13]

Figure 10 : Données sauvegardées après l'étape de la segmentation en lignes

4. Segmentation en mots

Pour la segmentation en mot, chaque ligne (enregistrée dans l'étape de la segmentation en lignes) subit le même traitement ; il s'agit d'appliquer la méthode de la projection verticale cette fois-ci, et de la même manière utiliser un tableau dont la taille est égale à la largeur de l'image (image de la ligne).
Si la colonne i de l'image contient au moins un point noir, alors t[i]=1 sinon t[i]=0 ;
Ainsi, si nous trouvons un nombre de zéros successifs qui dépasse un seuil (5 dans notre cas) alors il s'agit d'un saut de ligne. (Figure 11)

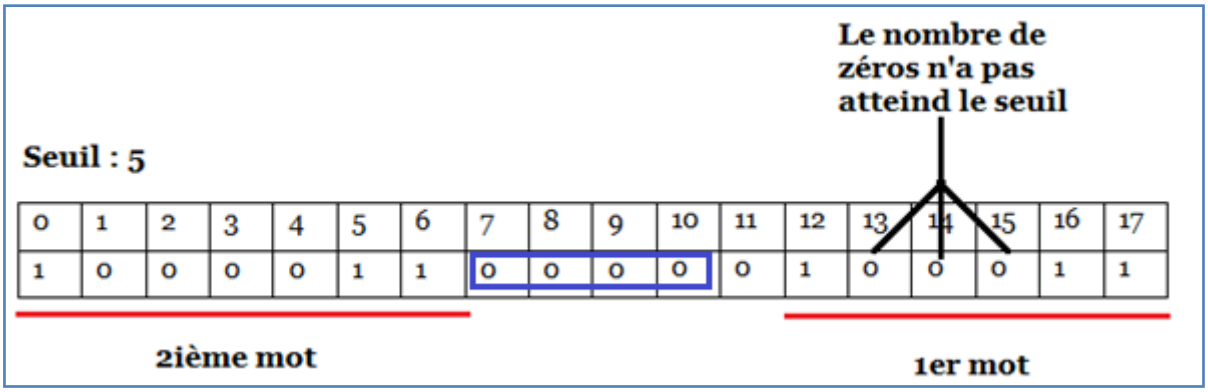


Figure 11 : Détermination des mots

Pour chaque ligne, nous enregistrons : (Figure 12)

- ✓ Le nombre des mots de la ligne
- ✓ Les images des mots obtenus
- ✓ Un tableau débutMot de taille égale au nombreMots, avec débutMot[i] = l'indice où commence le mot numéro i.
- ✓ Un tableau finMot de taille égale au nombreMot, avec finMot[i] = l'indice où termine le mot numéro i.

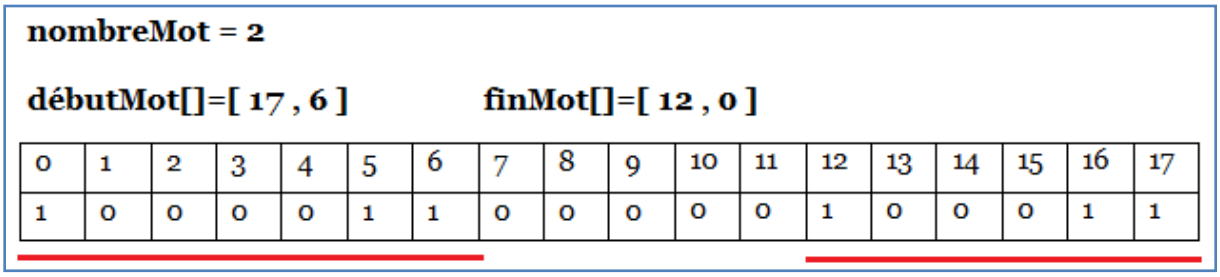


Figure 12 : Données sauvegardées après l'étape de la segmentation en mots

5. Segmentation en caractères

La langue arabe est une langue cursive (ou semi-cursive), c'est-à-dire que la majorité des lettres sont attachées, et par la suite la segmentation du mot en caractères est une opération très difficile.

Nous ne pouvons pas appliquer directement la méthode de la projection verticale parce qu'il n'y a pas généralement de 'vide' séparant les caractères du mot arabe (Figure 13)

Figure 13 : Problème de la segmentation des mots en caractères à cause de la cursivité de la langue arabe

Notre but est de créer ces 'vides' entre les caractères en éliminant la ligne de base (Figure 14)

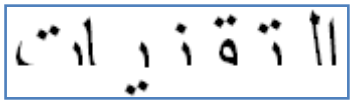


Figure 14 : Résultat souhaité

Le traitement s'effectue mot par mot, et suit plusieurs étapes :

a. Détermination de la ligne maximale

Nous voulons dire par la ligne maximale celle ayant un nombre maximal de points noirs dans le mot. Notons par ***max*** le nombre de points noirs dans la ligne maximale.

b. Elimination de la ligne de base

Pour éliminer la ligne de base, nous supprimons (mettons les points noirs de la ligne au blancs) toute ligne qui dépasse un seuil (60% dans notre cas) par rapport à ***max***. Càd : si le nombre de points noirs de la ligne est supérieur de 0.6* le nombre des points noirs de la ligne maximale.

Figure 15 : Elimination de la ligne de base

c. Mots qui ne sont pas concernés par l'élimination de la ligne de base

Il y a des cas où nous n'avons pas besoin d'éliminer la ligne de base :

i. Mots constitués d'un seul caractère

En effet les mots constitués d'un seul caractère n'ont pas besoin d'être segmentés, par la suite ils sont ajoutés à la liste des caractères obtenus directement. (Figure 16)

Question : Comment identifier les mots constitués d'un seul caractère ?

Réponse : Les mots de cette catégorie sont identifiés par les images dont la largeur ne dépasse pas un seuil (15 dans notre cas)

Figure 16 : Exemples des mots constitués d'un seul caractère

ii. Mots constitués uniquement des caractères isolés

Nous avons dit que nous avons opté pour l'idée d'éliminer la ligne de base des mots pour isoler les caractères cursifs de la langue arabe, or les mots qui sont constitués uniquement des caractères isolés n'ont pas besoin de subir le même traitement que les autres mots.

Question : Comment identifier les mots qui sont constitués uniquement des caractères isolés ?

Réponse : Notons p = max/largeur, avec max = le nombre des points noirs de la ligne maximale du mot et largeur = la largeur du mot.
Les mots constitués uniquement des caractères isolés sont identifiés par un p inférieur à un seuil (40% dans notre cas) (Figure 17)

Figure 17 : Exemples des mots constitués uniquement de caractères isolés

d. Détection des vides entre les caractères dans le mot sans ligne de base

Maintenant que nous avons le mot avec la ligne de base éliminée, nous pouvons appliquer la méthode de la projection verticale.

Nous sauvegardons les positions des débuts et fins des caractères en détectant les vides entre les caractères dans le mot sans ligne de base, et les images des caractères segmentés en utilisant le mot original et les positions sauvegardés. (Figure 18)

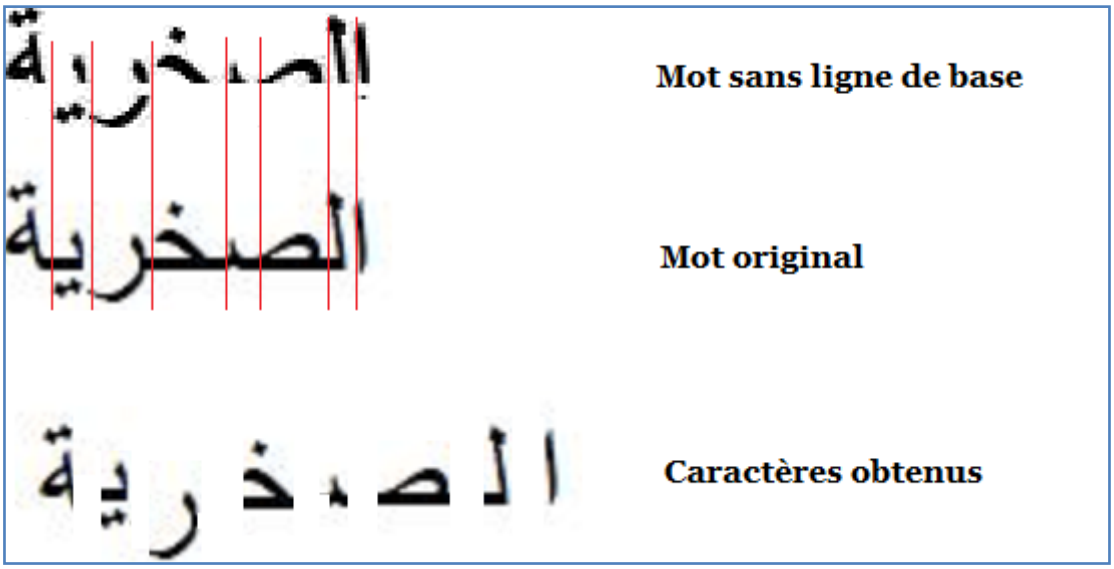


Figure 18 : Utilisation des mots original et sans base pour construire les caractères segmentés

Nous sauvegardons à cette étape pour chaque mot :

- Le nombre de caractères obtenus *nombreCaractères*.

- Un tableau *débutCaractères* de taille égale au *nombreCaractères*, avec *débutCaractères*[i] = l'indice où commence le caractère numéro i.
- Un tableau *finCaractères* de taille égale au *nombreCaractères*, avec *finCaractères*[i] = l'indice où termine le caractère numéro i.
- Les images des caractères obtenus.

e. Problème de chevauchements des caractères arabes

Le chevauchement des caractères est l'un des problèmes qui rendent la segmentation des mots arabes en caractères une tâche complexe. (Figure 19)

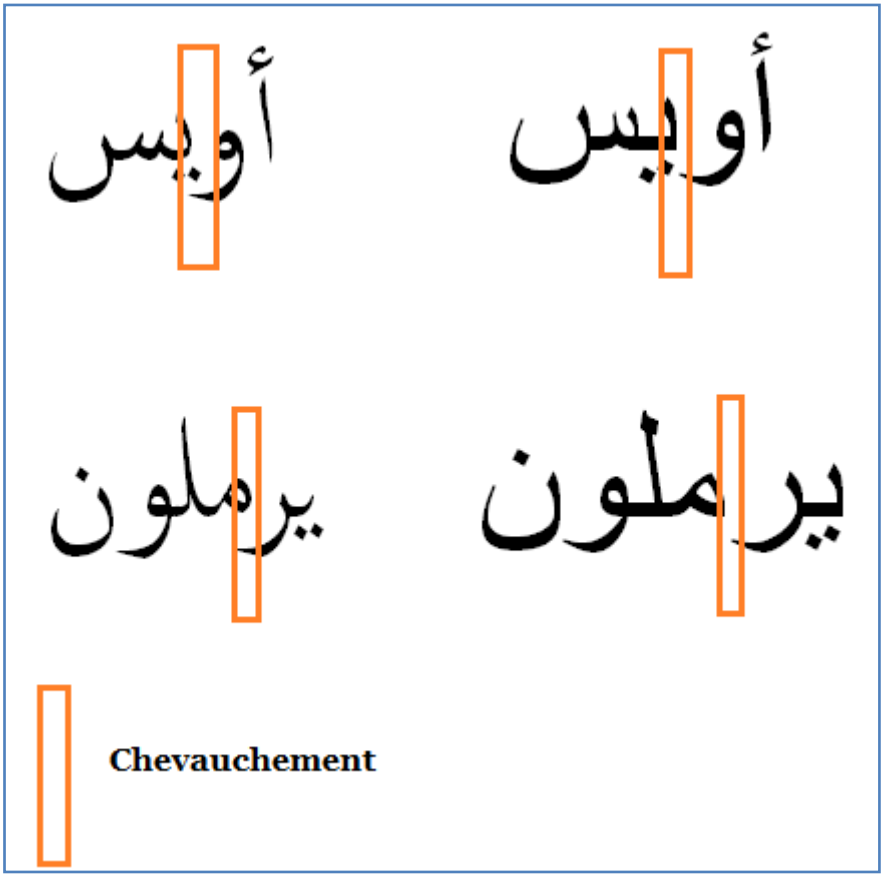


Figure 19 : Exemples de chevauchement des caractères arabes

Dans le cas d'un chevauchement de deux caractères arabes, nous obtiendrons dans le résultat des caractères segmentés les deux caractères chevauchés comme étant un seul caractère (Figure 20)

Figure 20 : Exemple montrant l'influence du chevauchement sur la segmentation en caractères

i. Comment détecter le chevauchement de deux caractères ?

Si la largeur de l'image du caractère obtenu dépasse un seuil (17 [2] dans notre cas) alors nous pouvons savoir qu'il s'agit d'un chevauchement d'au moins deux caractères.

ii. Comment résoudre le problème du chevauchement ?

Si la largeur de caractère obtenu dépasse le seuil, alors il subit le même traitement de segmentation des mots en caractères, avec une seule modification.
Dans la détection des vides pour les mots, nous faisons le test au long de toute la colonne or pour le caractère constitué d'un chevauchement, nous faisons le test juste sur l'intervalle [ligneMaximale – 5 , ligneMaximale+5]. La figure 21 permet d'éclaircir la différence entre les deux traitements.

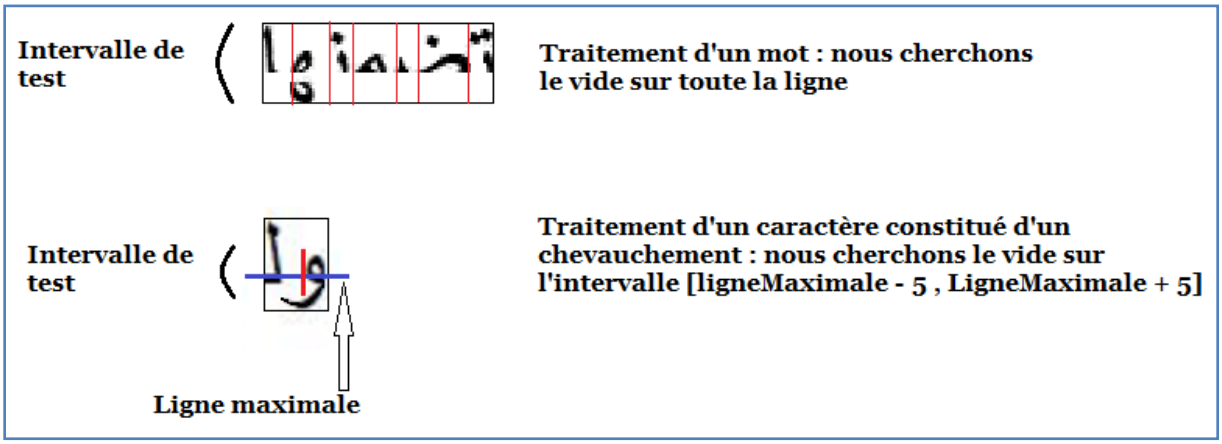


Figure 21 : Différence entre le traitement d'un mot et de caractère constitué d'un chevauchement

[2] La largeur maximale des caractères de la fonte traitée dans ce projet est 17.

f. Problèmes restants (limites) de l'approche proposée pour la segmentation en caractère

i. Problèmes liés à l'élimination de la ligne de base

Parfois (mais très rarement) il y a des mots où la ligne de base n'est pas complètement éliminée, à cause de la présence de quelque caractères isolés dans le mot. (Figure 22)

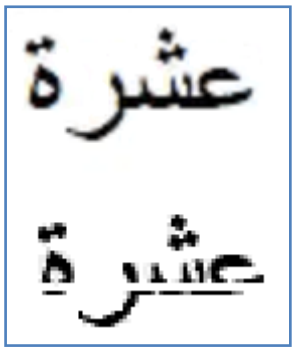


Figure 22 : Problème de l'élimination de la ligne de base

ii. Problèmes liés à la forme de quelques caractères arabes

- ✓ Les lettres « ص » et « ض » sont constituées d'une boucle suivie d'un pic, lors de la segmentation nous obtenons deux caractères comme le montre la figure 23.

Figure 23 : Exemple d'un mot segmenté contenant la lettre ص

- ✓ Le résultat de la segmentation de la lettre « ش » donne aussi deux caractères (Figure 24)

Figure 24 : Exemple d'un mot segmenté contenant la lettre ش

- ✓ Le résultat de la segmentation de la lettre « س » donne trois caractères (Figure 25)

Figure 25 : Exemple d'un mot segmenté contenant la lettre س

- ✓ Le résultat de la segmentation des lettres « ب », « ت », « ث », « ن » situés à la fin du mot donne trois caractères (Figure 26)

Figure 26 : Exemples des mots segmentés contenant les lettres ب ت ث ن

Ces problèmes seront résolus dans la phase de la reconnaissance.

III. Construction de la base de données

Nous avons créé des dossiers numérotés de 1 jusqu'au nombre de caractères que nous utilisons (lettre d'alphabet, points de ponctuation et caractères spéciaux). Chaque dossier contient plusieurs occurrences d'un seul caractère, en effet le caractère peut prendre différentes formes selon sa position dans le mot ou le résultat de la segmentation. La figure 27 montre les différentes formes prises par la lettre ع

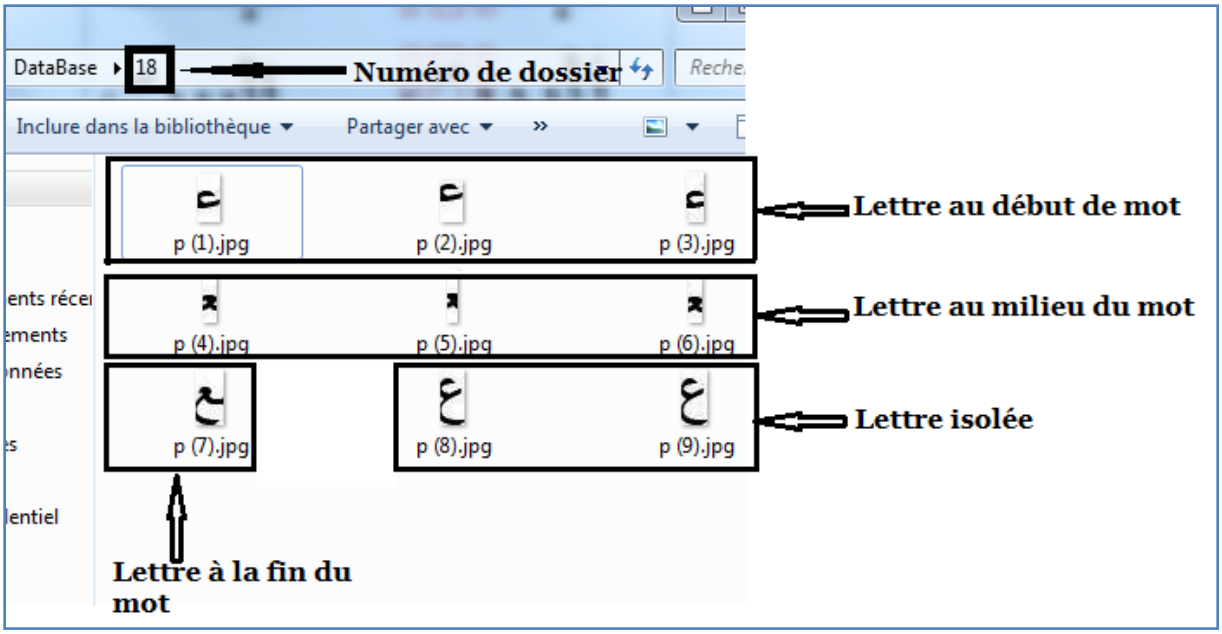


Figure 27 : Exemple d'un dossier de la base de données

1. Méthode classique pour la construction des caractères

L'approche classique ou directe pour construire les caractères consiste à faire un découpage manuel (une segmentation manuelle) en utilisant la souris pour séparer les caractères dans l'image et enregistrer chacun dans une image spécifiée.
Lors de la segmentation, quelques caractères obtenus ne sont représentés que par des parties des caractères originaux (Figure 28), d'où le problème de la segmentation manuelle car ces caractères ne trouveront pas de modèles dans la base de données.

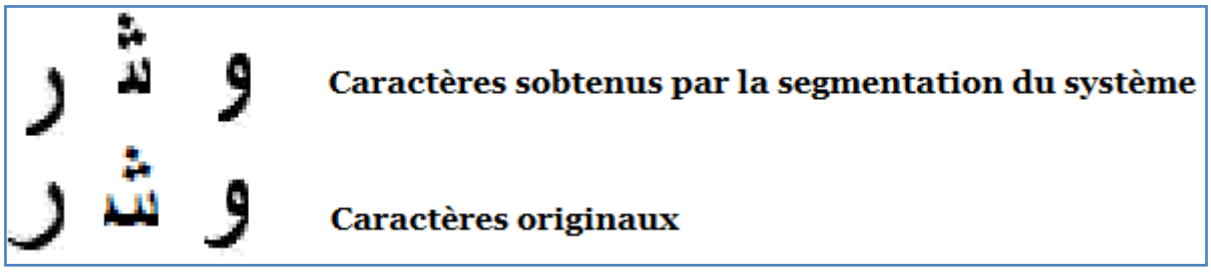


Figure 28 : Exemples des caractères obtenus par la segmentation

2. Méthode proposée pour la construction des caractères

Pour remédier au problème de la méthode citée ci-dessus, nous avons effectué des segmentations faites par le système (automatiquement) sur un échantillon des

images textes, puis nous avons déplacé l'image de chaque caractère au dossier correspondant. Avec cette idée nous assurons que les caractères requêtes correspondent exactement aux caractères stockés dans la base de données.

IV. Etape de la reconnaissance

La reconnaissance se fait caractère par caractère (approche analytique)

1. Prétraitement sur les images des caractères requêtes

Les caractères requêtes subissent un prétraitement avant de les comparer avec les modèles de la base de données

a. Squelettisation

Nous avons utilisé une implémentation (publiée dans le site developpez.net[3]) du squelette de la méthode décrite dans l'article [82]

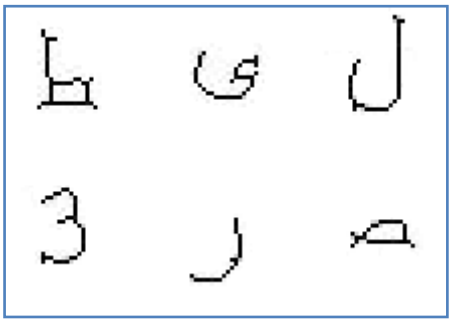


Figure 29 : Exemples des squelettes des caractères

b. Rogner[4] l'image

Après la squelettisation de l'image, nous rognons l'image pour ne garder que la partie de l'image contenant l'information. (Figure 30)

[3] http://www.developpez.net/forums/d344035/autres-langages/algorithmes/contribuez/image-filtre-squelette-imagej/
[4] Rogner : Découper les bords. Ex Rogner une image. [Dictionnaire linternaute.com]

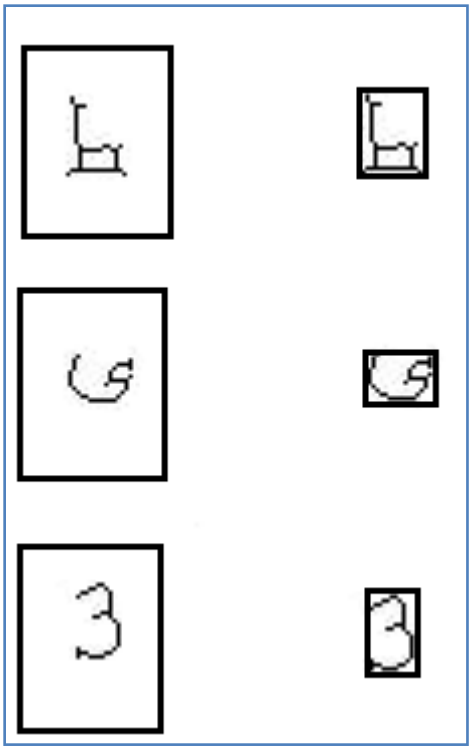


Figure 30 : Exemples des images de caractères rognées

2. Les points d'intérêts du squelette

Notre idée du choix des points d'intérêts est inspirée de l'algorithme de la corde [83] qui permet de rechercher une forme polygonale du contour coïncidant au mieux avec la forme d'origine. (Figure 31)

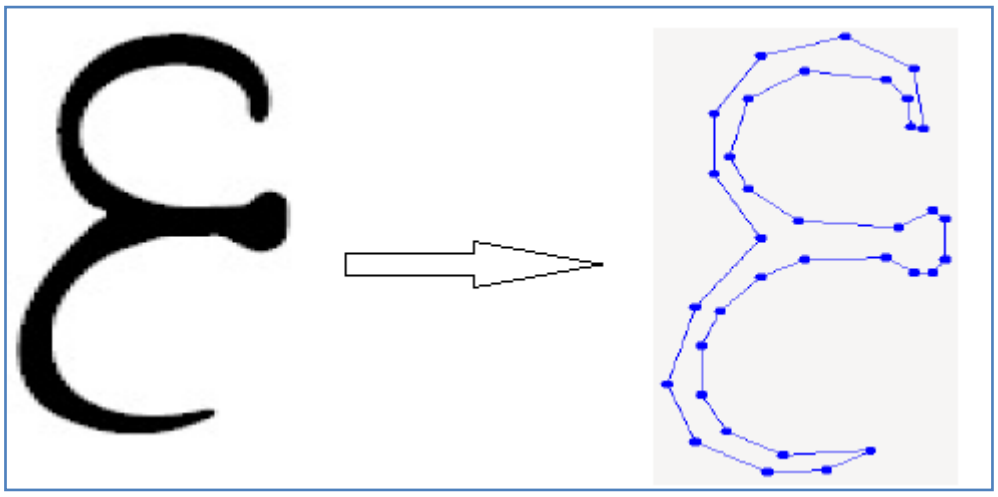


Figure 31 : Polygonisation du contour de la lettre ع [84]

Notre méthode consiste à calculer pour chaque image de squelette :

- Le point barycentre *Bary* (Xc , Yc)=1/N*(∑Xi , ∑Yi), avec :
    - Xc et Yc Sont les coordonnées du point barycentre.

- N est le nombre des points du squelette.
- Xi et Yi sont les coordonnées du point d'indice i.

- Le point *A* qui est le point le plus éloigné du barycentre et qui appartient au squelette.
- Le point *B* qui est le point le plus éloigné du point *A* et qui appartient au squelette.
- Le point *M* qui est le point le plus éloigné du segment [*A,B*] et qui appartient au squelette.
- Le point *M1* qui est le point le plus éloigné du segment [*A,M*] et qui appartient au squelette.
- Le point *M2* qui est le point le plus éloigné du segment [*B,M*] et qui appartient au squelette.

3. Comparer un caractère requête avec les modèles de la base de données

Les images de la base de données subissent la même procédure que les images des caractères à reconnaitre.

Notons (A1, B1, M1, M11, M21) et (A2, B2, M2, M12, M22) les points d'intérêts des caractères requêtes et modèles.

Pour chaque modèle, nous calculons un indice de dissimilarité *d* défini par : d = distance (A1, A2) + distance (B1, B2) + distance (M1, M2) + distance (M11, M12) + distance (M21, M22).

Le modèle dont l'indice de dissimilarité est plus faible est considéré comme le modèle le plus proche au caractère requête.

4. Reconnaissance du texte

La reconnaissance se fait caractère par caractère, pour reconnaitre un texte il suffit de concaténer les caractères retournés par le système et ajouter des espaces à la fin de chaque mot.

5. Correction des défauts de la segmentation

Nous avons cité quelques problèmes de la segmentation liés à la forme de quelques caractères (Chapitre III-5-f-ii), alors nous proposons ces solutions :

- ✓ Si le caractère retourné est la lettre ص alors le caractère suivant est négligé. (Figure 32.a)
- ✓ Si le caractère retourné est la lettre ش alors le caractère suivant est négligé. (Figure 32.b)
- ✓ Si le caractère retourné est la lettre س alors les deux caractères suivants sont négligés. (Figure 32.c)
- ✓ Si le caractère retourné est l'une des lettres ن ،ب ،ت ،ث à la fin du mot, alors le caractère suivant et précédents sont négligés. (Figure 32.d)

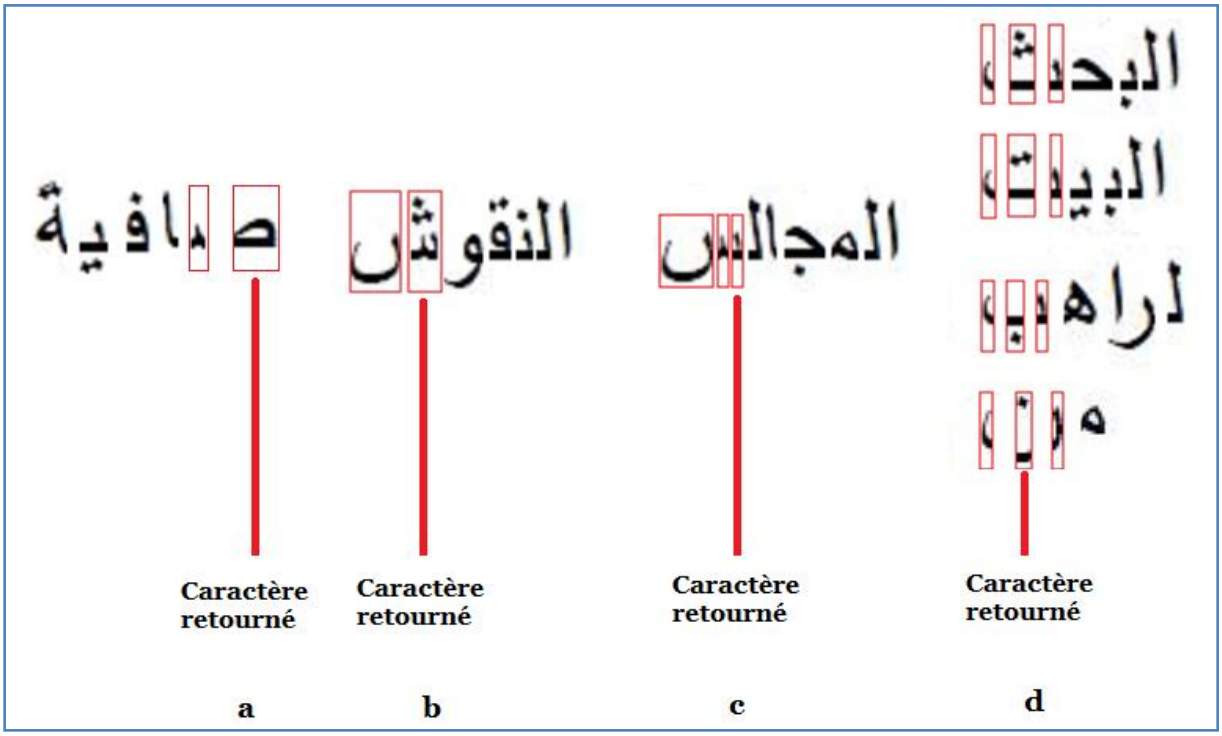


Figure 32 : Gestion des défauts de la segmentation

6. Amélioration de la reconnaissance (Post-OCR)

Parce qu'un caractère peut prendre différentes formes après la segmentation, il est possible qu'un caractère soit non reconnu par le système. Alors dans la phase du test des résultats, chaque fois que nous trouvons un caractère mal reconnu, nous ajoutons son image dans la base de données pour avoir presque toutes les occurrences des caractères.

7. Optimisation de la base de données

L'inconvénient de la méthode proposée pour améliorer la reconnaissance est que nous pouvons avoir plusieurs occurrences différentes du même caractère qui ont les mêmes points d'intérêts, ce qui augmente la taille de la base gratuitement. Pour remédier à ce problème, nous effectuons un balayage lors du chargement de la base de données, en effet si nous supprimons une image s'il existe une autre ayant les mêmes points d'intérêts.

CHAPITRE IV : APPLICATION ET RESULTATS

Dans la partie pratique de ce projet nous avons réalisé un système complet de l'OCR arabe imprimé mono-fonte.
Dans ce chapitre nous présentons l'application réalisée, les résultats obtenus de chaque étape, les taux de réussite de chaque opération et nous faisons une comparaison des résultats avec d'autres systèmes OCR.

1. Outil de développement : Eclipse

Eclipse est un environnement de développement intégré libre extensible, universel et polyvalent, permettant de créer des projets de développement mettant en œuvre n'importe quel langage de programmation.
Eclipse IDE est principalement écrit en Java (à l'aide de la bibliothèque graphique SWT, d'IBM), et ce langage, grâce à des bibliothèques spécifiques, est également utilisé pour écrire des extensions.
La spécificité d'Eclipse IDE vient du fait de son architecture totalement développée autour de la notion de plugin (en conformité avec la norme OSGi) : toutes les fonctionnalités de cet atelier logiciel sont développées en tant que plug-in.
Plusieurs logiciels commerciaux sont basés sur ce logiciel libre, comme par exemple IBM Lotus Notes 8, IBM Symphony ou WebSphere Studio Application Developer. [85]

2. Segmentation en lignes

La figure 33 présente l'image d'entrée du système capturée du livre «Gravures rupestres poétiques» en format « pdf ».


• تحقيق نصوص النقوش الصخرية الشعرية

عدد النقوش الصخرية الشعرية التي تضمنها البحث عشرة نقوش تتضمن أربعة عشر بيتا؛ هي:

(1) النقش الأول:

أفنى الجديد تقلب الشمس وطلوعها من حيث لا تمسي

وطلوعها بيضاء صافية وغروبها صفراء كالورس

• المصدر : ورد النقش في ثلاثة مصادر:

1) مجلة المنهل، المجلد (39)، محرم 1398هـ-1977م، استكشافات آثارية إسلامية عريقة على صخور قرب عرفة، عبد القدوس الأنصاري، الصفحات (392-394).

Figure 33 : Image à traiter

a. Résultats de la segmentation en lignes


• تحقيق نصوص النقوش الصخرية الشعرية

عدد النقوش الصخرية الشعرية التي تضمنها البحث عشرة نقوش تتضمن أربعة عشر

بيتا؛ هي:

(1) النقش الأول:

أفنى الجديد تقلب الشمس وطلوعها من حيث لا تمسي

وطلوعها بيضاء صافية وغروبها صفراء كالورس

• المصدر : ورد النقش في ثلاثة مصادر:

1) مجلة المنهل، المجلد (39)، محرم 1398هـ-1977م، استكشافات آثارية

إسلامية عريقة على صخور قرب عرفة، عبد القدوس الأنصاري، الصفحات

(392-394).

Figure 34 : Image segmentée en lignes

b. Taux de réussite de la segmentation en lignes

Le taux de réussite de cette opération est de 100%.

3. Segmentation en mots

a. Résultats de la segmentation en mots

Figure 35 : Image segmentée en mots

b. Taux de réussite de la segmentation en mots

Le taux de réussite de cette opération est aussi 100%.

4. Segmentation en caractères

a. Résultats de la segmentation en caractères

Figure 36 : Image segmentée en caractères

b. Taux de réussite de la segmentation en caractères

Les mots mal segmentés dans l'image ci-dessus sont :

عشرة: à cause du problème de la ligne de base (Figure 22)

عشر: même problème que le mot précédent

الأنصاري : le dernier caractère de ce mot est divisé en deux, parce qu'il a dépassé le seuil donné pour la largeur des caractères, et par la suite il a été considéré et traité comme un chevauchement de deux caractères.

Au total nous avons 322 caractères dans l'image ci-dessus, 6 caractères n'étaient pas bien segmentés, ce qui donne un taux de réussite de : 316/322 = 98%.

5. Reconnaissance des caractères

a. Résultats de la reconnaissance

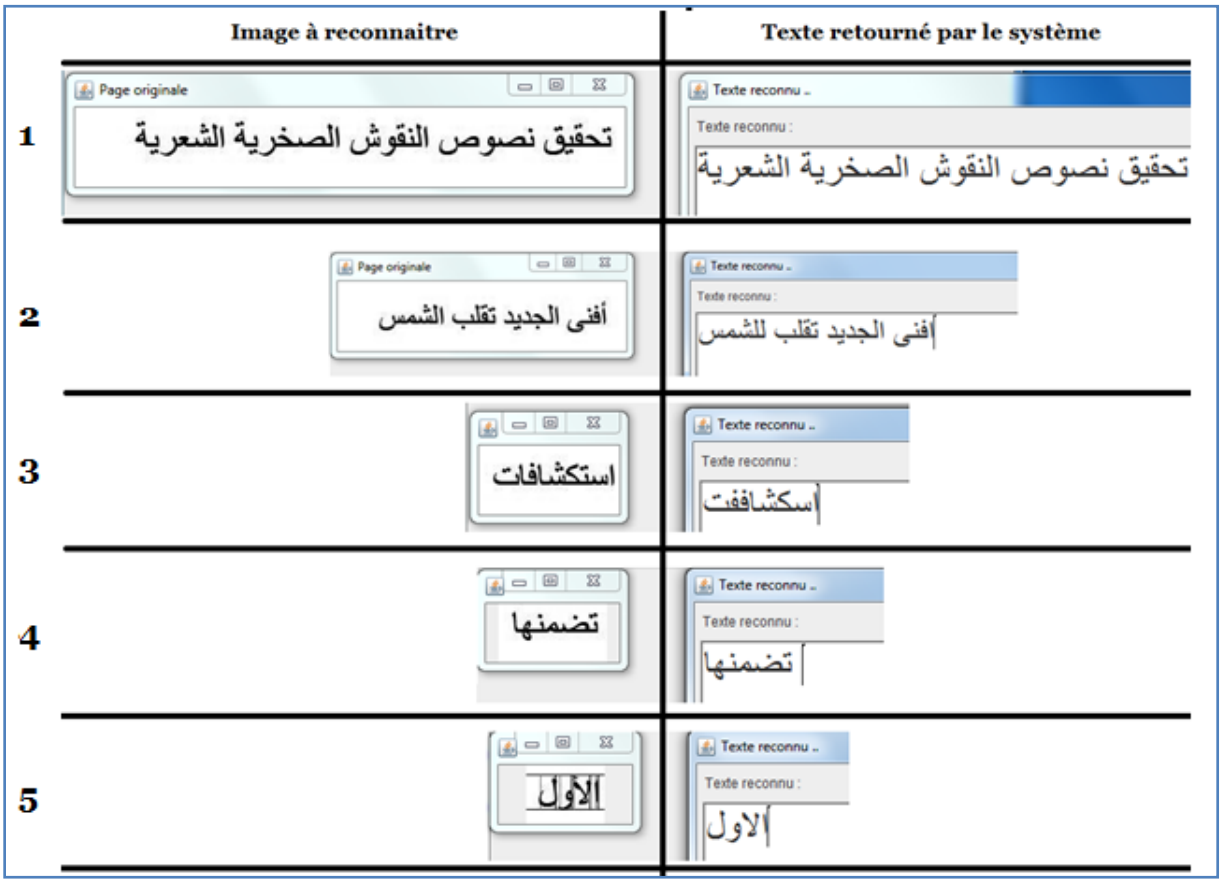


Figure 37 : Résultats de la reconnaissance de texte

b. Taux de réussite de la reconnaissance :

Pour les images 1, 4 et 5 de la figure 37, le taux de réussite de la reconnaissance de texte est de 100%.

Pour les images 2 et 3 de la même figure le taux de réussite est respectivement : 94 % et 77%.

6. Comparaison avec d'autres systèmes

a. Comparaison de la segmentation en caractères avec le système [86]

Pour faire une comparaison juste, nous utilisons la même phrase utilisée par l'auteur de l'article [86] :

المملكة المغربية دولة تقع في شمال افريقيا وعاصمتها الرباط

Figure 38 : Comparaison entre notre système et le système [86]

Le taux de réussite de la segmentation du système [86] est de : 95%

Le taux de réussite de la segmentation de notre système est de : 97%. En effet, nous avons une seule lettre «ي» du mot « **في**» qui est mal segmentée, or pour l'autre système, les lettres «ص » du mot « **عاصمتها**» et « ط » du mot « **الرباط**» se sont découpés en deux parties après la segmentation.

b. Comparaison de la reconnaissance avec le système [86]

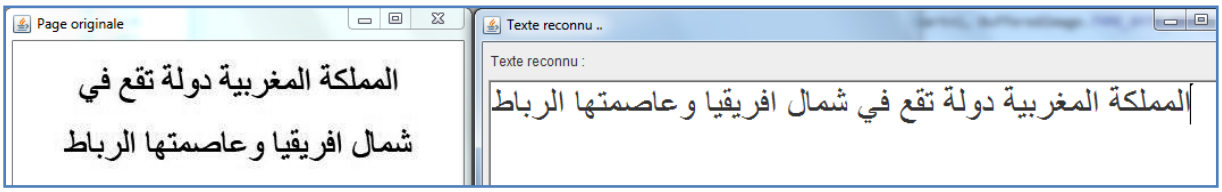


Figure 39 : Résultat de la reconnaissance retourné par notre système.

Notre système a pu reconnaitre la phrase sans problème, même l'erreur de la segmentation pour la lettre « ي » est rectifiée dans l'étape de la reconnaissance.

Pour le système [86], il a retourné deux fautes de reconnaissance : la lettre « ش » est considérée comme « ع », et la lettre « ص » est considérée comme « ه ».

c. Comparaison de la reconnaissance avec le système Sakhr

Le texte retourné par le système Sakhr est :

المملكة المإربية دولة تغع في شملى افريفيا وعاصمتها الرباط

Le système Sakhr a retourné 5 erreurs : la lettre « غ » n'est pas reconnu , la lettre « ق » est considérée comme « غ », la lettre « ا» est considérée comme « ل », la lettre « ل» est considérée comme « ى » et la lettre « ق » est considérée comme «ف»

7. Conclusion

Dans ce travail, nous avons essayé au maximum d'implémenter nos propres idées dans les diverses étapes du système de reconnaissance.

Selon les comparaisons que nous avons faites avec d'autres systèmes, nous pouvons dire que nous avons obtenu des résultats encourageants.

Nous avons consacré la majorité du temps pour la phase de la segmentation, car la robustesse de tout le système dépend d'elle et d'autre part parce qu'elle est une tâche délicate et qui reste toujours un domaine de recherche.

Concernant la reconnaissance, nous avons trouvé des problèmes dus à la qualité et la petite taille du document étudié. Les résultats au début sont acceptables et peuvent être optimisés dans la phase de post-traitement.

Le point faible du système réside dans la grande taille de la base de données ce qui influence négativement sur le temps de réponse du système.

CONCLUSION ET PERSPECTIVES

Malgré les efforts et les travaux intensifs réalisés dans le domaine de la reconnaissance optique de l'écriture, aucun système OCR n'est jugé fiable à 100%. Mais au fur et à mesure les auteurs essayent d'améliorer les scores pour de meilleurs résultats.

Les problèmes majeurs influençant la recherche en AOCR sont le manque de normalisation des calligraphies des caractères arabes, l'absence d'outils tels que dictionnaires, bases de données et statistiques se rapportant à l'écriture arabe. Concernant notre travail, nous avons obtenu des résultats satisfaisants, nous essayerons encore de les améliorer sous l'encadrement des professeurs du LSIA de la FSTF et les chercheurs d'ILC de Pise en Italie, espérant pouvoir contribuer à la recherche dans le domaine de l'AOCR.

Perspectives :

Ce travail n'est que le début d'un projet en collaboration de l'équipe de LSIA travaillant sur l'AOCR, les chercheurs d'ILC de Pise et le chercheur Maxim ROMANOV du département Classics & Perseus Project – Université de Tufts en USA.

Les perspectives les plus importantes sont :

- Améliorer la phase de la reconnaissance pour optimiser les résultats obtenus et le temps de réponse du système, ainsi qu'automatiser.
- Automatiser l'opération du post traitement.

REFERENCES

[1] Segmentation de textes en caractères pour la reconnaissance optique de la langue arabe (HAITAAMAR Schahrazed)

[2] La reconnaissance de l'écriture (Ben Ammara)

[3] Une approche hybride pour la reconnaissance d'écriture arabe manuscrite (Azizi Rebiai)

[4] T. Steinherz, E. Rivlin, N. Intrator : «Off-line cursive word recognition : a survey ». International journal on document analysis and recognition, 2(2), pp. 90-110, 1999.

[5] A. Belaïd et Y. Belaïd, Reconnaissance des formes méthodes et applications, InterEdition, 1992.

[6] B. Al-Badr, S.A. Mahmoud : « Survey and bibliography of Arabic optical text recognition ». Signal processing , vol. 41, pp. 49-77, 1995.

[7] A.Belaid : «Analyse de documents: de l'image à la représentation par les normes de codage». Cours de l'INRIA 1997

[8] T. Hu, R. Ingold : « A mixed approach toward an efficient logical structure recognition from document image ». Electronic publishing, vol.6(4), pp. 457-468, December 1993

[9] K. Etemad, D. Doermann, R. Chellappa : « Page segmentation using decision integration and wavelet packets ». International conference on pattern recognition, 1994.

[10] R.M. Haralick : « Document image understanding : geometrical and logical layout ». IEEE. Proc. International conference on computer vision and pattern recognition, vol. 8, pp. 385-390, 1994.

[11] Y.Y. Tang, M. Cheriet, J. Liu, J.N. Said, C.Y. Suen : « Document analysis and recognition by computers ». Handbook of pattern recognition and computer vision Chap 8, Editeurs: C.H. Chen, I.P. Pau et P.S.P. Wang.

[12] S. Mao, T. Kanungo : « Empirical performance evaluation of page segmentation algorithms ». Proc. SPIE on document recognition and retrieval, vol. 3967, pp. 303-314, 2000.

[13] R.G. Casey, E. Lecolinet : «A survey of methods and strategies in character segmentation ». IEEE Transactions on pattern analysis and

machine intelligence, vol. 18, No. 7, pp. 690-7, july 1996.

[14] C.C. Tappet, C.Y. Suen , T. Wakahara : « The state of the art in on-line handwritten recognition ». IEEE. Transaction on pattern analysis and machine intelligence, vol. 12, No 8, pp. 787-808, 1990.

[15] T.M. Ha, G. Kaufmann, H. Bunke : « Text localization and handwriting recognition». Technical report, university of Berne, 1996.

[16] B. Al-Badr, R.M. Haralick : « Symbol recognition without prior segmentation ». Conference SPIE-EI 1994.

[17] M. Bulmenstein, B.Verma : « A neural based segmentation and recognition technique for handwritten words». IEEE. Proc. Of the international conference on neural networks. Vol. 3, pp. 1738- 1742, 1998.

[18] M. Bulmenstein, B. Verma : « An artificial neural network based segmentation algorithm for off-line handwritten recognition ». Proc. Of the international conference on computational Intelligence and multimedia applications (ICCIMA'98), pp. 27-33 1998.

[19] M. Bulmenstein, B.Verma : « A neural based solution for the segmentation and recognition of difficult handwritten words from a benchmark database». Proc. Of the 5th international conference on document analysis and recognition (ICDAR'99). pp. 281-284, Bangalore, India, 1999.

[20] K.M. Sayre : « Machine recognition of handwritten words : a project report ». Pattern recognition, vol. 5, pp. 213-228, 1973

[21] E. Lecolinet, J.P. Crettez : « A grapheme-based segmentation technique for cursive script recognition». IEEE. Proc. International conference on document analysis and recognition (ICDAR'91), Saint-Malo, France 1991.

[22] C.J.C. Burges, J.I. Be, C.R. Nohl : « Recognition of handwritten cursive postal words using neural networks ». Proc. USPS 5th Advanced technology conference. 1992.

[23] J. Trenkle, A. Gillies, E. Erlandson, S.Schlosser : « Arabic character recognition ». Proc. symposium on document image understanding technology (SDIUT'95), 1995.

[24] J. Trenkle, A. Gillies, S.Schlosser : « An off-line Arabic recognition system for machine printed documents ». Proc. Of the symposium on

document image understanding technology (SDIUT'97), pp. 155-161 1997.

[25] J. Trenkle, A. Gillies, E. Erlandson, S.Schlosser, S. Cavin : « Advances in Arabic text recognition ». Proc. of the symposium on document image understanding technology (SDIUT'01), Maryland, Columbia, April 23-25, 2001.

[26] A. Gillies, E. Erlandson, J. Trenkle, S. Schlosser : « Arabic text recognition system ». Proc. Of the symposium on document image understanding technology, Annapolis, Marryland, 1999.

[27] N.E. Ayat, M. Cheriet, C.Y. Suen : « Un système neuro-flou pour la reconnaissance de montants numériques de chèques arabes ». Proc. Of CIFED'00, pp. 171-180, 2000.

[28] S. Kermi : « Classifieur neuronal base connaissances, application à la reconnaissance des caractères arabes isolés manuscrits ». Thèse de magister, université Badji Mokhtar, Annaba, Algerie 1999.

[29] N. Benamara : « Utilisation des modèles de Markov cachés planaires en reconnaissance de l'écriture arabe imprimée ». Thèse de doctorat, spécialité Génie Electrique, Université des sciences, des Techniques et de médecine de Tunis II, 1999.

[30] J. Anigbogu : « Reconnaissance de textes imprimés mutifontes à l'aide de modèles stochastiques et métriques ». thèse de doctorat, Université de Nancy I, 1992.

[31] P. Burrow : « Arabic handwriting recognition ». Master of science thesis. School of Informatics, university of Edinburg, England, 2004.

[32] J.L. Amat, G. Yahiaoui : « Techniques avancées pour le traitement de l'information». Edition CEPADUES 1996.

[33] L. Souici, Z. Zmirli, M. Sellami : « Système connexionniste pour la reconnaissance de l'arabe manuscrit ». 1ères journées scientifiques et techniques (JST FRANCIL), pp. 383-388, Avignon, France, 1997.

[34] K. Seymore, A. McCallum, R. Rosenfeld : « Learning Hidden Markov model structure for information extraction ». AAAI. Workshop on machine learning for information extraction, pp. 37-42, 1999.

[35] Leila Cherqui. La Théorie de la Résonance Adaptative et les Moments de Zernike pour la Reconnaissance de Mots Arabes Manuscrits

[36] B. Al-Badr, R.M. Haralick : « Segmentation-free word recognition with application to Arabic ». IEEE. Proc. 3rd International conference on document analysis and recognition (ICDAR'95),
pp. 355-359, Montreal, Canada, 1995.

37] A.M. Elgammal, M.A. Ismail : « A graph-based segmentation and feature extraction framework for Arabic text recognition». IEEE. Proc. 6th international conference on document analysis and Recognition (ICDAR'01),2001.

[38] H.Y. Abdelazim, M.A. Hashish : « Interactive font learning for arabic OCR». Proc. 1ST Kuwait computer conference, pp 463-486 Kuwait, 1986.

[39] H.Y. Abdelazim, M.A. Hashish : « Arabic typeset : an OCR approach». Proc. 5TH EUSIPCO-90, European signal processing conference, pp 1019-1022, Barcelona, Spain 90.

[40] I.S.I. Abuhaiba, P. Ahmed: «Restoration of temporal information in off-line Arabic handwriting». Pattern recognition, vol. 26, No 7, pp 1009-1017, 1993.

[41] S. Al-Emami, M. Usher : « On-line recognition of handwritten Arabic characters ». IEEE Transactions on pattern analysis and machine intelligence, vol. 12, No 7, 1990.

[42] H. Aissaoui, A. Haouari : « Une méthode structurelle pour la reconnaissance de textes arabe manuscrits » 14èmes journées tunisiennes en électrotechnique et automatique (JTEA'94), pp. 203-207, Hammamet, Tunisie, 1994.

[43] H. Aissaoui, A. Aissaoui, A. Haouari : « Une approche neuronale pour la reconnaissance de textes arabe imprimés multifonte» 16èmes journées tunisiennes en électrotechnique et automatique (JTEA'96), pp. 402-409, Hammamet-Nabeul, Tunisie, 1996.

[44] H. Aissaoui, A. Haouari:«Application des transformations unitaires à la reconnaissance de textes arabe » 17ème journées tunisiennes en électrotechnique et automatique (JTEA'97, pp. 333-340, Nabeul, Tunisie, 1997.

[45] A.M. Alimi , O.A. Ghorbel : « Etude de l'influence du nombre de prototypes dans la reconnaissance en ligne de lettres arabes moulées ». Actes du 3ème Colloque national sur l'écrit et le document (CNED'94),

pp. 293-298, Rouen, 1994.

[46] H. Almuallim, S. Yamagushi : « A method of recognition of Arabic cursive handwriting». IEEE Transactions on pattern analysis and machine intelligence, vol. PAMI-9, No5, pp. 715-722, 1987.

[47] H.S. Al-Yousefi, S.S. Udpa : « Recognition of Arabic characters ». IEEE Transactions on pattern analysis and machine intelligence, vol. 14, No 8, pp. 853-857, 1992.

[48] A. Ameur, K.Romeo-Packer, Y. Lecourtier : « 'arabe manuscrit et sa reconnaissance informatique ». Actes du colloque : langue arabe et technologies informatiques avancées, pp. 215-232, Casablanca Maroc 1993.

[49] A.Ameur, K. Romeo-Packer, H. Miled, M. Cheriet : « Approche globale pour la reconnaissance des mots manuscrits arabes ». Actes du 3ème Colloque national sur l'écrit et le document (CNED'94), pp. 151-156, Rouen, 1994.

[50] A.Ameur, K. Romeo-Packer, H. Miled, M. Cheriet : « Coupling observation/letter for a markovian modelisation applied to the recognition of Arabic handwriting ». IEEE. Proc. 4th International conference on document analysis and recognition (ICDAR'97), pp. 580-583, Ulm, Germany, 1997.

[51] A. Amin : « Machine recognition of handwritten Arabic words by the IRAC II system ».Proc. of the 6th international joint on pattern recognition, Munich, 1982.

[52] A. Amin, J.F. Mari : « Machine recognition and correction of printed Arabic text ». IEEE Transaction on system, man, cybernetics, vol. 19, No 5, pp. 1300-1304, 1989.

[53] A. Amin, S. Al-Fedaghi : « Machine recognition of printed Arabic text utilizing natural language morphology ». IEEE Transactions on systems, man and cybernetic. Vol. 35, No 6, pp.769-788, 1991

[54] R. Azmi, E. Kabir : « A new segmentation technique for omnifont Farsi text ». Pattern Recognition letters. No 22, pp 97-104, 2001.

[55] N. Benamara, N.Ellouze : « A robust approach for Arabic printed character segmentation ». IEEE. Proc. 3rd International conference on document analysis and recognition (ICDAR'95) pp. 865-868, Montreal,

Canada, 1995.

[56] F. Bouslama : «Arabic character recognition by fuzzy techniques» Proc. 5th European congress on intelligent techniques and soft computing, Aachen, Gerrmany, 1997.

[57] F. Bouslama, H. Kishibe : « Fuzzy logic in the recognition of machine printed Arabic characters». IEEE. Proc. 6th international conference on neural information processing (ICONIP'99), vol. 3 pp. 1150-1154, 1999.

[58] S.S. El-Dabi, R. Ramsis, A. Kamel : « Arabic character recognition system : a statistical approach for recognizing cursive typewritten text». Pattern recognition, vol. 23, No ¾, pp. 337-346, 1990.

[59] F. El-Khaly, M.A. Sid-Ahmed : «Machine recognition of optically captured machine printed Arabic text». Pattern recognition, Vol.23 No 11, pp. 1207-1214, 1990.

[60] T. El-Sheikh, R. Guindi : «Computer recognition of Arabic cursive scripts ». Pattern recognition, vol. 21, No 4, pp. 293-302, 1988.

[61] T. El-Sheikh, S.G. El-Taweel : « Real time Arabic handwritten character recognition ». Pattern recognition , vol. 23, No 12, pp. 97-105, 1990.

[62] M.C. Fehri, M. Ben Ahmed : « A new approach to Arabic character recognition in multifont documents ». Proc. 4th international conference and exhibition on multi-lingual computing (Arabic and Roman script), pp.2.5.1-2.5.7, university of Cambridge, London, UK, April 1994.

[63] M.C. Fehri, M. Ben Ahmed : « off-line handwriting recognition». Computational engineering in systems applications (CESA'98), pp. 1-3, Nabeul-Hammamet, Tunisie, 1998.

[64] A. Gillies, E. Erlandson, J. Trenkle, S. Schlosser : « Arabic text recognition system ». Proc. Of the symposium on document image understanding technology, Annapolis, Marryland, 1999.

[65] H. Goraine, M. Usher : « Printed Arabic text recognition». Proc. 4th International conference and exhibition on multi-lingual computing (Arabic and Roman script) , pp. 2.6.1-2.6.8 , university of Cambridge, London, UK 1994.

[66] F. Hadj-Hassen : « Printed Arabic text recognition ». Arabian Journal of Engineering science, vol. 16, No 4, 1991.

[67] K.M.Hassibi : «Machine-printed Arabic OCR using neural networks» Proc. 4th International conference and exhibition on multi-lingual Computing (Arabic and Roman script), pp. 2.3.1-2.3.11, university of Cambridge, London, UK 1994.

[68] K.M. Jambi : « An approach for segmenting handwritten Arabic words ». Actes du colloque : langue arabe et technologies informatiques avancées, pp. 233-245, Casablanca, Maroc 1993.

[69] M.B. Kurdy, A. Joukhadar, A. Wabbi : « Multifont Arabic/latin optical character recognition system ». Actes du Colloque : langue arabe et technologies informatiques avancées, pp. 245-256, Casablanca, Maroc 1993.

[70] M.A. Mahjoub : « Reconnaissance en-ligne des caractères arabes isolés par les chaînes de Markov cachées ». 16èmes journées tunisiennes en électrotechnique et automatique (JETA'96), pp. 358-367, Hammamet-Nabeul, Tunisie, 1996.

[71] M.A. Mahjoub, N. Ellouze : « Reconnaissance en-ligne des PAWs par les modèles de Markov cachés non stationnaires ». 6èmes colloque Magrebin sur les modèles numériques de l'ingénieur, pp. 335-340, Tunis, Tunisie, 1998.

[72] H. Miled : « Stratégie de reconnaissance de l'écriture arabe manuscrite ». Actes JED'96, Premières journées sur l'écrit et le document, jeunes chercheurs, pp. 27-28, juillet 1996.

[73] H. Miled, M. Cherit, C. Olivier, Y. Lecourtier : « Modelisation Markovienne de l'écriture arabe manuscrite : une approche analytique ». 1er colloque international Froncophone sur l'écriture et le document (CIFED'98), Quebec, Canada, mai 1998.

[74] D. Motawa, A. Amin, R. Sabourin : « Segmentation of Arabic cursive script». IEEE Proc. 4th international conference on document analysis and recognition (ICDAR'97), pp. 625-628, Ulm, Germany, 1997

[75] C. Olivier, H. Miled, K. Romeo-Pakker, Y. Lecourtier:«Segmentation and coding of Arabic handwritten words » . IEEE Proc. 13th international conference on pattern recognition (ICPR'96), pp. 264- 268, Vienne, Autriche, 1996.

[76] J. Trenkle, A. Gillies, E. Erlandson, S.Schlosser, S. Cavin : « Advances

in Arabic text recognition ». Proc. of the symposium on document image understanding technology (SDIUT'01), Maryland, Columbia, April 23-25, 2001.

[77] A. Zahour, B. Taconet, A. Faure : « Une méthode de reconnaissance de l'écriture arabe cursive». Proc. 1st international conference on document analysis and recognition (ICDAR'91), pp. 454-462, Saint-malo, France, 1991.

[78] A. Zahour, A. Djematene, S. Kebairi, A. Bennasri, B. Taconet : « contribution à la reconnaissance de l'écriture manuscrite arabe». Proc. du 1er colloque international francophone sur l'écrit et le document (CIFED 98), pp. 218-227, Québec, Canada, 1998.

[79] B.M.F. Bushofa, M. Spann : « Segmentation of Arabic characters using their contour information ». IEEE. Proc. International conference on digital signal processing, vol 2, pp 683-686, 1997.

[80] M.M.M. Fahmy, S.Al Ali : « Automatic recognition of handwritten Arabic characters using their geometrical features ». Studies in informatics and control journal (SIC journal), vol. 10, No 2, 2001.

[81] O. Hachour : « Reconnaissance hybride des caractères arabes imprimés ». Traitement automatique de l'arabe (JEP-TALN 2004), Fès, 20 Avril 2004

[82] A novel single pass thinning algorithm

[83] V. Ramer, "An Iterative Procedure for the Polygonal Approximation of Plane Curves", 1972

[84] I. Chakir. Nouvelle approche pour la reconnaissance des caractères arabes imprimés

[85] Site université Paris Sud : http://hebergement.u-psud.fr/

[86] N. El Makhfi. Segmentation multi échelle des caractères pour la reconnaissance de l'écriture arabe imprimée.