République algrienne démocratique et populaire

الجمهورية الجزائرية الديمقراطية الشعبية

UNIVERSITE EL-HADJ LAKHDHAR BATNA

Mémoire présenté

Pour l'obtention du diplôme de

Magister en informatique

Option

INFORMATIQUE INDUSTRIELLE

Thème

SEGMENTATION DE TEXTES EN CARACTERES POUR LA RECONNAISSANCE OPTIQUE DE L'ECRITURE ARABE

Présenté par : HAITAAMAR Schahrazed

Devant un jury composé de:

Président : Dr ZIDANI Abdelmadjid Maître de conférence université de Batna

Rapporteur : Dr BATOUCHE Mohamed Professeur université de Constantine

Examinateurs: Dr BILAMI Azzeddine Maître de conférence université de Batna

Dr CHIKHI Salim Maître de conférence université de Constantine

Soutenue le 08 Juillet 2007

RESUME

Le présent travail porte sur une étude concernant le domaine de reconnaissance optique de caractères arabes imprimés. Une étude générale sur les systèmes de reconnaissance de l'écriture a été développée, puis elle a été affinée par un intérêt particulier à une phase considérée comme cruciale dans le procédé de reconnaissance: la phase de segmentation.

Nous avons présenté un état de l'art des méthodes de segmentation des caractères, ensuite nous avons présenté la langue arabe et le domaine de l'OCR, nous avons soulevé certains problèmes de normalisation dans l'écriture arabe qui peuvent poser des problèmes dans la réalisation de bon systèmes de reconnaissance.

Nous avons aussi étudié les méthodes actuellement utilisées dans la segmentation des caractères arabes, puis donné liste détaillée des travaux de plusieurs auteurs. Après une comparaison de méthodes de segmentation de caractères arabes imprimés, nous avons terminé ce travail par la contribution par un algorithme simple de segmentation.

La méthode adoptée dans cet algorithme est basée sur un principe déjà utilisé qui est le principe de projections verticales, ce qui a été proposé est une étape de post traitement corrigeant les erreurs, durant la phase de segmentation. Des scores proches du 100% ont été obtenus

MOTS-CLES: OCR, segmentation, caractères arabes, pseudo-mot, post-traitement.

Table de matières

Liste des figures

Liste des tableaux

Abréviations

AOCR : Arabic Optical Character Recognition.

ASCII : American Standard Code Information Interchange.

ASMO : Arabic Standard Metrology Organization.

CCW : Counter Clock Wise.

CXX : Composantes Connexes.

C-LAG : Compressed Line Adjacency Graph

DSP : Decisive Segmentation point.

HMM : Hidden Markovian Model.

KNN : K Nearest Neighbor.

LAG : Line Adjacency Graph.

OCR : Optical Character Recognition.

PAO : Publication Assistée par Ordinateur.

PAW : Peace of Arabic Word.

PS : Pen Size.

PSP : Primary Segmentation Points.

INTRODUCTION

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 cheques, …

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 avancé que pour d'autres langues, deviennent plus intensifs qu'avant. Dans ce travail nous présentant un aperçu sur la reconnaissance des caractères, les étapes suivies pour la réalisation d'un système OCR puis nous nous intéressons plus particulièrement à la phase de segmentation en caractères de façon générale. Nous introduisons ensuite les caractéristiques de l'écriture arabe de point de vue calligraphie tout en soulevant les problèmes liés aux normalisations et au codage de cette écriture, suivie par une étude de l'état de l'art de la segmentations des caractères arabes, puis nous affinons cette étude par la présentation d'algorithmes pour la segmentation des caractères arabes imprimés. Nous terminons le travail par une proposition d'un algorithme de segmentation de l'écriture arabe imprimée hors ligne.

PLAN DE LECTURE DU MEMOIRE

Ce mémoire est constitué de cinq 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'écrit.
- Le deuxième chapitre étudie l'OCR et la langue arabe. La première section rappelle certaines données de la calligraphie arabe, suivie de notions d'OCR sur l'écriture arabe.
- Le troisième chapitre est spécifique à l'état de l'art de la segmentation des textes dans le cas général. Nous avançons graduellement de la détection des objets dans une page, à la segmentation des blocs de texte en lignes puis en mot puis en caractères. Nous mettons l'accent sur les méthodes utilisées dans ce type de segmentation.
- Le quatrième chapitre est l'étude de l'état de l'art de la segmentation des textes arabes. Dans ce chapitre nous étudions les différentes approches utilisées dans la segmentation en caractères des mots, puis un survol sur les différents travaux dans ce domaine. Nous terminons par l'étude de quatre algorithmes type de segmentation.
- Le cinquième chapitre est une petite contribution par un algorithme simple et assez efficace pour la segmentation des textes arabes imprimés multifonte.

Nous terminons le travail par des échantillons de résultats de cet algorithme et des perspectives de ce travail.

CHAPITRE I

LA RECONNAISSANCE DE L'ECRITURE.

I-1- INTRODUCTION

Toute information écrite peut être reprise dans une chaîne de traitement informatisée à différentes fins : la rédaction et l'édition de rapports, la diffusion de documents dans un système de messagerie … conduisent à exploiter des informations disponibles seulement sur papier. La reconnaissance optique de caractères (OCR) est une opération informatique rapide permettant de réaliser la transformation d'un texte écrit sur papier en un texte sous forme d'un fichier informatique en représentation symbolique (par exemple pour les écritures latines, le codage opéré est le code ASCII *(American Standard code for information interchange)*, tandis que pour l'arabe on utilise généralement le code ASMO *(Arabic Standard Metrology Organization)*.

I-2- DIFFERENTS ASPECTS DE L'OCR

Il n'existe pas de système universel d'OCR qui permet de reconnaître n'importe quel caractère dans n'importe quelle fonte. Tout dépend du type de données traitées et bien évidemment de l'application visée [Benamara 99]. Il existe plusieurs modes de classification des systèmes OCR parmi lesquels on peut citer :

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

I-2-1- RECONNAISSANCE EN-LIGNE ET HORS-LIGNE

Ce sont deux modes différents d'OCR, ayant chacun ses outils propres d'acquisition et ses algorithmes correspondants de reconnaissance.

a) ***La reconnaissance en-ligne (on-line)*** :

Ce mode de reconnaissance s'opère en temps réel (pendant l'écriture). Les symboles sont reconnus au fur et à mesure qu'ils sont écrits à la main.

Ce mode est réservé généralement à l'écriture manuscrite . c'est une approche « signal » où la reconnaissance est effectuée sur des données à une dimension . l'écriture est représentée comme un ensemble de points dont les coordonnées sont fonction du temps [Lecolinet 93], [Al-Badr 95].

La reconnaissance en-ligne présente un avantage majeur c'est la possibilité de correction et de modification de l'écriture de manière interactive vu la réponse en continu du système [Lallican 00].

L'acquisition de l'écrit est généralement assurée par une tablette graphique munie d'un stylo électronique.

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

Démarre après l'acquisition. Elle convient aux documents imprimés et les manuscrits déjà rédigés. Ce mode peut être considéré comme le cas le plus général de la reconnaissance de l'écriture. Il se rapproche du mode de la reconnaissance visuelle. L'interprétation de l'information est indépendante de la source de génération [Tsang 00].

La reconnaissance hors-ligne peut être classée en plusieurs types :

- ***Reconnaissance de texte ou analyse de documents*** : Dans le premier cas il s'agit de reconnaître un texte de structure limitée à quelques lignes ou mots. La recherche consiste en un simple repérage des mots dans les lignes, puis à un découpage de chaque mot en caractères [Benamara 99]. Dans le second cas (analyse de document), il s'agit de données bien structurés dont la lecture nécessite la connaissance de la typographie et de

la mise en page du document. Ici la démarche n'est plus un simple prétraitement, mais une démarche experte d'analyse

de document il y'a localisation des régions, séparation des régions graphiques et photographique, étiquetage sémantique des zones textuelles à partir de modèles, détermination de l'ordre de lecture et de la structure du document [Trenkle 97].

- ***Reconnaissance de l'imprimé ou du manuscrit*** : Les approches diffèrent selon qu'il s'agisse de reconnaissance de caractères imprimés ou manuscrits. Les caractères imprimés sont dans le cas général alignés horizontalement et séparés verticalement, ce qui simplifie la phase de lecture [Benamara 99]. La forme des caractères est définie par un style calligraphique (fonte) qui constitue un modèle pour l'identification. Dans le cas du manuscrit, les caractères sont souvent ligaturés et leur graphisme est inégalement proportionné provenant de la variabilité intra et inter-scripteurs. Cela nécessite généralement l'emploi de techniques de délimitation spécifiques et souvent des connaissances contextuelles pour guider la lecture [Fahmy 01].

  Dans le cas de l'imprimé, la reconnaissance peut être monofonte, multifonte ou omnifonte .

  Un système est dit *monofonte* s'il ne peut reconnaître qu'une seule fonte à la fois c'est à dire qu'il ne connaît de graphisme que d'une fonte unique. C'est le cas le plus simple de reconnaissance de caractères imprimés [Anigbogu 92].

  Un système est dit *multifonte* s'il est capable de reconnaître divers types de fontes parmi un ensemble de fontes préalablement apprises [Benamara 99].

  Et un système *omnifonte* est capable de reconnaître toute fonte, généralement sans apprentissage préalable. Cependant ceci est quasiment impossible car il existe des milliers de fontes dont certaines illisible par l'homme (sauf bien sure pour celui qui l'a conçue) et avec in logiciel de création de fonte n'importe qui peut concevoir des fontes à sa guise

[Anigbogu 92]. Anigbogu [Anigbogu92] a présenté une autre définition pour ce terme c'est l'expression « *polyfonte* » et a qualifié un système polyfonte de système capable de reconnaître un très grand nombre de fontes.

I-2-2- RECONNAISSANCE GLOBALE OU ANALYTIQUE

L'approche globale considère le mot comme une seule entité et le décrit indépendamment des caractères qui le constituent. Cette approche présente l'avantage de garder le caractère dans son contexte avoisinant, ce qui permet une modélisation plus efficace des variations de l'écriture et des dégradations qu'elle peut subir. Cependant cette méthode est pénalisante par la taille mémoire, le temps de calcul et la complexité du traitement qui croient linéairement avec la taille du lexique considéré, d'où une limitation du vocabulaire [Al-Badr 95],[Amin97].

L'approche analytique : contrairement à l'approche globale, le mot est segmenté en caractères ou en fragments morphologiques significatifs inférieurs au caractère appelés graphèmes. La reconnaissance du mot consiste à reconnaître les entités segmentés puis tendre vers une reconnaissance du mot, ce qui constitue une tâche délicate pouvant générer différents types d'erreurs [Amin 97],[Lecolinet 93]. Un processus de reconnaissance selon cette approche est basé sur une alternance entre deux phases : la phase de segmentation et la phase d'identification des segments. Deux solutions sont alors possibles : la segmentation explicite (externe) ou la segmentation implicite (interne) [Casey 95]. Par ailleurs, les méthodes analytiques par opposition aux méthodes globales, présentent l'avantage de pouvoir se généraliser à la reconnaissance d'un vocabulaire sans limite à priori, car le nombre de caractères est naturellement fini. De plus l'extraction des primitives est plus aisée sur un caractère que sur une chaîne de caractères [Al-Badr 94].

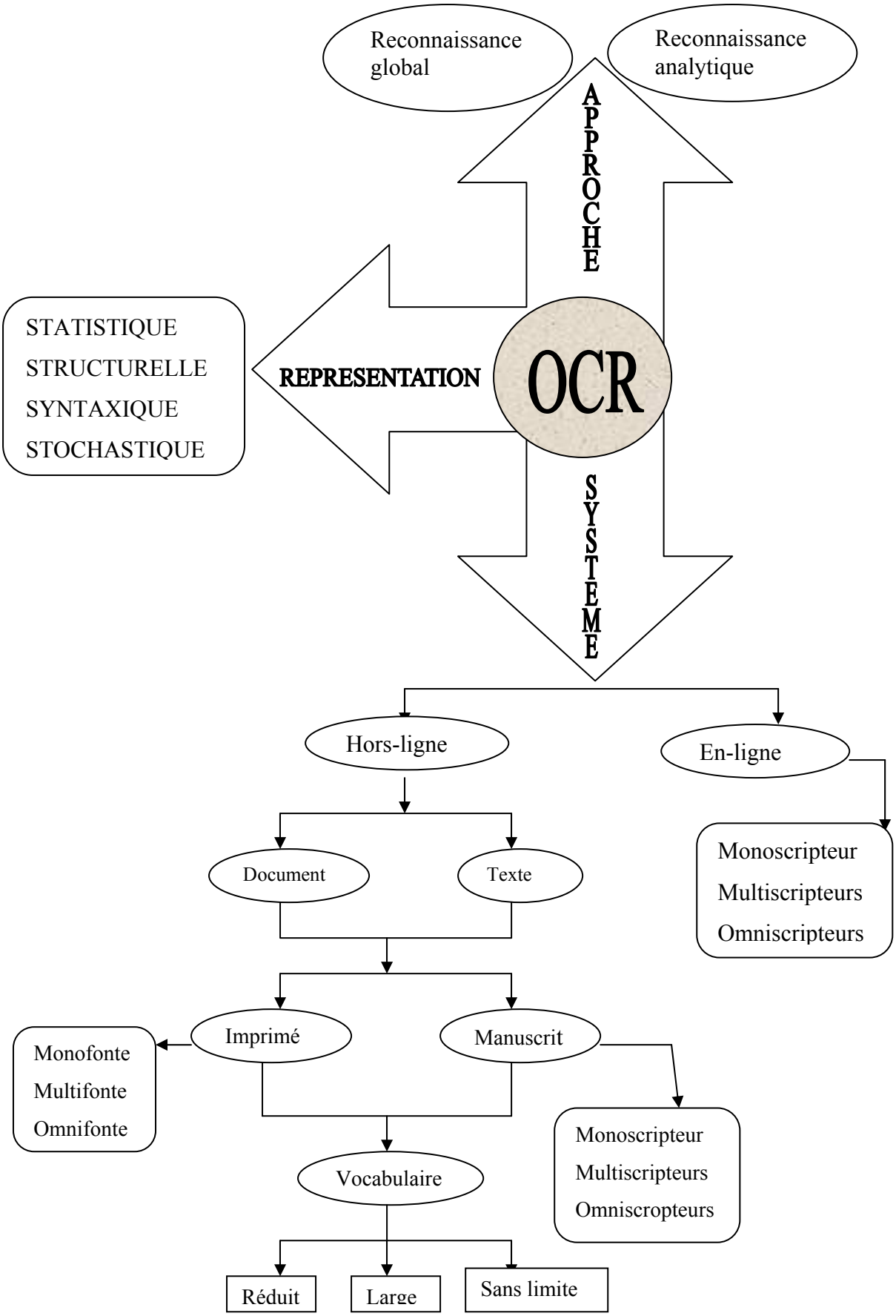


Figure I-1 *: Différents systèmes, représentations et approches de reconnaissance.*

Tiré de [Benamara 99]

I-3- PROBLEMES LIES A L'OCR

La tâche de l'OCR n'est pas aisée, divers problèmes compliquent le processus de reconnaissance, parmi lesquels on peut citer [Al-Badr 95], [Benamara 99]:

- La qualité du document : un document télécopié ou photocopié plusieurs fois est plus difficile à traiter que la copie originale. L'écriture peut devenir plus mince ou au contraire plus épaisse, dégradée avec des parties du texte qui manquent ou de tâches qui apparaissent, des ouvertures ou des bouchages de boucles …
- L'impression : un document composé est de meilleure qualité qu'un document dactylographié qui, à son tour, est plus clair qu'un texte issu d'une imprimante matricielle. Une imprimante à jet d'encre peut introduire des tâches d'encre et un étalement des caractères, une imprimante laser peut générer des lignes ou des fonds …
- La discrimination de la forme : selon le style de la fonte utilisée, son corps et sa graisse…, le caractère change de graphisme. Le nombre de formes est d'autant plus important que le nombre de styles d'écriture est élevé. De plus, plusieurs caractères présentent une forte ressemblance tels que :
  - pour l'arabe : ه et د ,صـ et ر,مـ et و
  - pour le Latin : U et V , O et 0, S et 5, Z et 2 .
- Le support de l'information, tel que le papier, joue également sur les performances de la reconnaissance par sa qualité : son grammage, sa granulation et sa couleur.
- L'acquisition : la numérisation en temps réel introduit souvent des distorsions dans l'image. Dans le cas hors-ligne la qualité du texte numérisé est un compromis entre les variations de la position (inclinaison, translation, rétrécissement…), la propreté de la vitre du dispositif de numérisation et sa résolution.
- Les variations des dimensions : un « pitch » de 10, 12 ou de 16 … (10, 12 ou 16 cpi (character per inch)). Un pitch de 10 implique des caractères plus grands aussi bien en largeur qu'en hauteur que ceux d'un pitch de 12.

En plus de ces problèmes un système OCR devrait être capable de distinguer entre un texte et une figure, de reconnaître les caractères ligaturés et d'être indépendant des variations de l'espace aussi bien inter-mots que de l'interligne.

Les problèmes posés par la reconnaissance optique de l'écriture manuscrite, sont plus complexes que ceux liés à l'écriture imprimée. Les erreurs de lecture dans le cas du manuscrit sont dues aux variations infinies de l'écriture de nature aléatoire qui dépendent de facteurs particuliers du scripteur et des conditions de l'écriture.

I-4-ORGANISATIONGENERALE D'UN SYSTEME DE RECONNAISSANCE

Un système de reconnaissance fait appel généralement aux étapes suivantes : Acquisition, prétraitements, segmentation, extraction des caractéristiques, classification, suivis éventuellement d'une phase de post-traitement (figure I-3)

I-4-1- PHASE D'ACQUISITION

La phase d'acquisition consiste à capter l'image d'un texte au moyen des capteurs physiques (scanner, caméra,…) et de la convertir en grandeurs numériques adaptés au système de traitement, avec un minimum de dégradation possible.

I-4-2- PHASE DE PRETRAITEMENTS

Le prétraitement consiste à préparer les données issues du capteur à la phase suivante. Il s'agit essentiellement de réduire le bruit superposé aux données et essayer de ne garder que l'information significative de la forme représentée. Le bruit peut être dû aux conditions d'acquisition (éclairage, mise incorrecte du document, …) ou encore à la qualité du document d'origine.

Parmi les opérations de prétraitement généralement utilisées on peut citer : l'extraction des composantes connexes, le redressement de l'écriture, le lissage, la normalisation et la squelettisation (figure I-2).

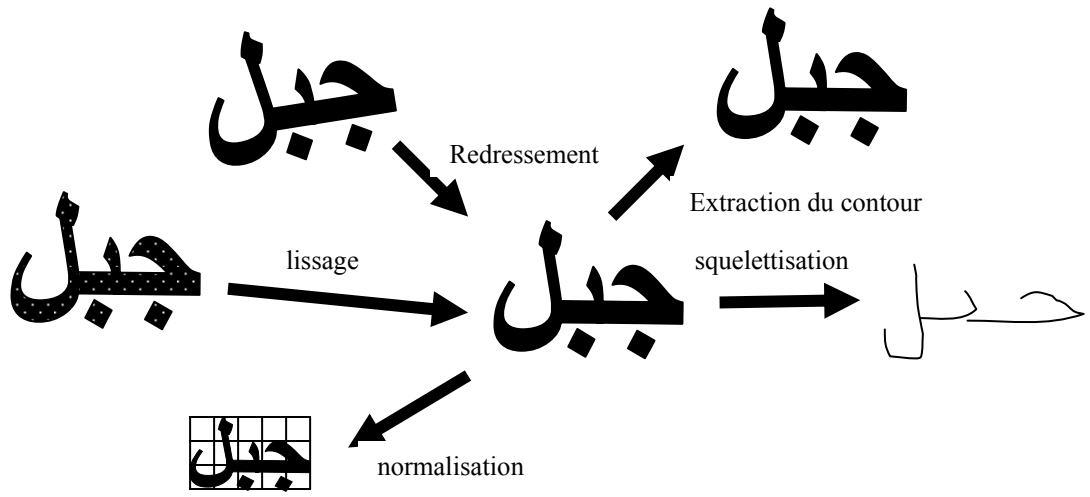


***Figure I-2** : effets de certaines opérations de prétraitement.*

a) *<u>Extraction de composantes connexes</u>* :

Une composante connexe (CXX) est un ensemble de points dans le plan. Elle peut correspondre à un point diacritique, un accent, au corps d'un caractère ou d'une chaîne de caractères… Une fois localisés les CXX sont regroupées pour former les mots. Cette technique est utilisée pour le repérage des points diacritiques dans les images de textes arabes [Benamara 99].

b) <u>Redressement de l'écriture</u> :

L'un des problèmes rencontrés en OCR est l'inclinaison des lignes du texte, qui introduit des difficultés pour la segmentation. L'inclinaison peut provenir de la saisie, si le document a été placé en biais, ou être intrinsèque au texte. Il convient alors de le redresser afin de retrouver la structure de lignes horizontales d'une image texte. Si α est l'angle d'inclinaison, pour redresser l'image, une rotation isométrique d'angle –α est opérée grâce à la transformation linéaire suivante [Steinherz 99] :

$$\begin{cases} x' = x\cos\alpha + y\sin\alpha \\ y' = y\cos\alpha + x\sin\alpha \end{cases}$$

c) *Lissage* :

L'image des caractères 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. Les techniques de lissage permettent de résoudre ces problèmes par des opérations locales qu'on appelle opérations de bouchage et de nettoyage [Burrow 04].

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. Plusieurs autres techniques similaires sont utilisées dont la méthode statistique, une méthode basée sur la morphologie mathématique … (pour plus de détail sur ces technique le lecteur peut se référer à [Benamara 99]).

d) *Normalisation* :

Après la normalisation de la taille, les images de tous les caractères se retrouvent définies dans une matrice de même taille, Pour faciliter les traitements ultérieurs. Cette opération introduit généralement de légères déformations sur les images. Cependant certains traits caractéristiques tels que la hampe dans les caractères (ط ظ ل ا par exemple) peuvent être éliminées à la suite de la normalisation, ce qui peut entraîner à des confusions entre certains caractères [Steinherz 99].

e) 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 [Steinherz 99].

La squelettisation des caractères arabes peut induire en erreur : deux points diacritiques sont souvent confondus avec un seul [Benamara 99].

Remarque :

Selon la qualité du document à traiter, le type de l'écriture et la méthode d'analyse adoptée, une ou plusieurs techniques de prétraitement sont utilisées. Mais pas forcément toutes.

I-4-3- PHASE DE SEGMENTATION :

Dans cette phase les différentes parties logiques d'une image sont extraites. 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 mot puis les caractères (ou parties du caractère) [Al-Badr 95]. Cette phase va être revue en détails dans les chapitre III et IV.

I-4-4 - PHASE D'ANALYSE OU D'EXTRACTION DES CARACTERISTIQUES

C'est l'une des étapes les plus délicates et les plus importantes en OCR. 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 : caractéristiques structurelles, caractéristiques statistiques, transformations globales, et superposition des modèles et corrélation [Kermi 99] [Al-Badr 95].

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 [Kermi 99]:

- 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, etc).

Plusieurs autres caractéristiques peuvent être tirés, suivant qu'ils soient extraits 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 utilisés pour la reconnaissances de textes arabes sont : le zonage (zonning), les caractéristiques de lieu géométrique (Loci) et les moment [Kermi 99].

- Le zonage consiste à superposer une grille n×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×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 [Al-Badr 95].
- 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 le 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 [Al-Badr 95]. 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 [Kermi 99].

I-4-5- PHASE DE 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 [Kermi 99].

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 [Benamara 99], [Al-Badr 95]. 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 [Al-Badr 95].

Les procédés d'apprentissage sont différents selon qu'il s'agisse de reconnaissance de caractères imprimés ou manuscrits ou de reconnaître des textes monofonte ou multifonte. 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 [Benamara 99], [Kermi 99].
- 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 [Benamara 99].

b) ***Reconnaissance et décision* :**

La décision est l'ultime étape de 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* [Benamara 99].

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

1) 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 » [Benamara 99].

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 [Al-Badr 95].

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

L'approche bayésienne

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 [Anigbogu 92].

La méthode du plus proche voisin

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 [Benamara 99 , burrow 04].

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.

Les réseaux de neurones

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 [Amat 96].

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 [souici 97].

Par ailleurs, le point fort des réseaux de neurones réside dans leur capacité de générer une région de décision de forme quelconque, requise par un algorithme de classification, au prix de l'intégration de couches de cellules supplémentaires dans le réseau [Lippman 87].

2) *Approche structurelle* :

Les méthodes structurelles reposent sur la structure physique des caractères. Elles cherchent à trouver des éléments simples ou primitives, 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 [Anigbogu 92], [Ha 96]. 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 [Benamara 99].

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 [Benamara 99].

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 [Benamara 99].

3) 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. [Benamara 99]. Ces méthodes sont robustes et fiables du fait de l'existence d'algorithmes d'apprentissage

efficaces [Seymore 99] . 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).

4) 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.

I-4-6- PHASE DE 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 objet 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.

I-5- CONCLUSION

Dans ce chapitre, nous avons présenté certains concepts généraux liés à la reconnaissance optique des caractères, en précisant les principales méthodes de reconnaissance. Nous avons aussi énuméré les principaux problèmes rencontrés par l'OCR. Ensuite nous avons abordé les différentes étapes intervenant dans la conception d'un système de reconnaissance de caractères et nous avons précisé qu'il existait différentes issues pour aborder ce domaine.

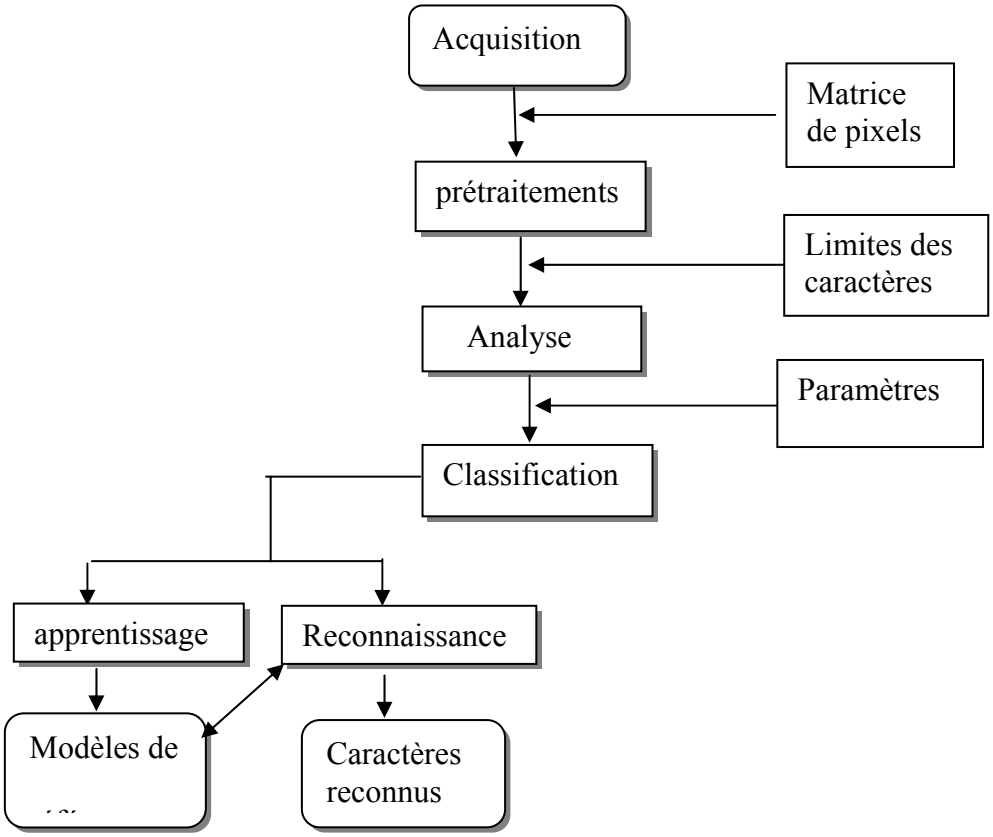


Figure I-3 : *Schéma général d'un système de reconnaissance de caractères.*

CHAPITRE II
L'OCR ET L'ARABE

II-1- INTRODUCTION

Dans ce chapitre, nous présentons les caractéristiques morphologiques de l'écriture arabe. Ensuite nous exposons les principaux travaux développés en OCR arabe, tout en soulevant les problèmes majeurs rencontrés dans ce domaine.

II-2- CALLIGRAPHIE ET TYPOGRAPHIE ARABES

II-2-1- CARACTERISTIQUES DE L'ECRITURE ARABE

L'arabe est écrit par plus de cent millions de gens, dans plus de vingt pays différents. L'écriture arabe a été développée à partir d'un type d'Araméen. La langue araméenne comporte moins de consonants que l'arabe, alors de nouvelles lettres ont été créée en ajoutant des points aux lettres déjà existantes. D'autres petites marques appelées diacritiques sont utilisées pour indiquer de courtes voyelles, mais elles ne sont généralement pas utilisées [Burrow 04].

L'arabe est une écriture consonantique qui utilise un alphabet de 28 lettres (Tableau II-1-a) auquel il faut ajouter la Hamza «ء», qui est le plus souvent considérée comme signe complémentaire [Al-Badr 95b]. 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 II-1-c).

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.

La considération du symbole « ~» qui s'écrit uniquement sur le support du caractère «ا», fait apparaître d'autres graphismes (Tableaux II-1-c et II-1-d). L'écriture arabe a ainsi plusieurs spécificités que nous citons ci-après.

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

(a)

caractère	initiale	médiane	finale	isolé
Ta marbouta			ـة	ة
Lamalif			ـلا	لا

(b)

caractère	initiale	médiane	finale	isolé
Alif+~			ـآ	آ
Alif+ء			ـأ	أ
			ـإ	إ
Waw+ء			ـؤ	ؤ
Ya+ء	ئـ	ـئـ	ـئ	ئ

(c)

caractère	initiale	médiane	finale	isolé
Lamalif +~			ـلآ	لآ
Lamalif +ء			ـلأ	لأ
			ـلإ	لإ

(d)

Tableau II-1-

(a) l'alphabet arabe dans ses différentes formes.

(b) Les caractères additionnels

(c) (c) et (d) Hamza et Madda et les positions qu'elles occupent en association avec Alif, Waw et Ya.

- 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 II-1).

ligne de base 

Figure II-1 – exemple d'écriture 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 II-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 II-2). 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 [Benamara 99].

Initiale	Médiane	Finale	Isoléé
عـلم	مـعلم	سمـع	ورع
هـمس	مـهـد	لعبـه	منتزه

Tableau II-2 – Les quatre formes des caractères « ain » et « he » en fonction de leur position dans la chaîne de caractères.

- 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 [Al-Badr 95].

- Certains caractères arabes incluent une boucle qui peut avoir différentes formes (Figure II-2).

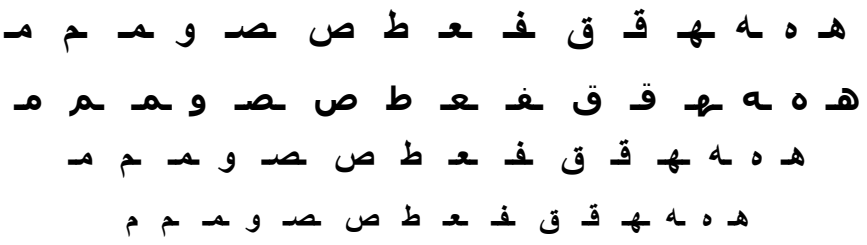


Figure II-2- Exemple de formes de boucles dans des styles différents

- certains caractères ne peuvent être rattachés à leur gauche et de ce fait ne peuvent se trouver qu'en position isolée ou finale, ce qui donne quand ils existent, des mots composés d'une ou de plusieurs parties qu'il est convenu d'appeler généralement PAW (*peace of arabic word*) ou encore pseudo-mot [Al-Badr 95]. Un PAW correspond donc à une chaîne d'un ou de plusieurs caractères (Tableau II-3). L'écriture arabe est ainsi semi-cursive plutôt que totalement cursive.

5 PAWs/mot	4 PAWs/mot	3 PAWs/mot	2 PAWs/mot	1PAW/mot
الرمادية,الوردية, الأزهار	الجزائر, الزهور, البذور	العالية, التاجر, البقاع	تونس ,برد ,عابد	مكثر, مطر, بلد

Tableau II-3- Exemple de mots composés de la droite vers la gauche de 1,2,3,4 et 5 PAWs respectivement.

Comme le caractère, le PAW peut se trouver dans des mots différents à des positions différentes, mais contrairement au caractère, le PAW présente une structure morphologique stable, il garde la même calligraphie dans les différentes positions qu'il occupe (Tableau II-4)

Initiale	Médiane	Finale	Isolé
قرار	ر**قر**اق	أ**قر**	**قر**

Tableau II-4- Le PAW «قر» dans différents mots et différentes positions

- Pour des raisons de justification de texte et/ou d'esthétique, les ligatures horizontales peuvent être allongées en insérant entre les caractères d'une même chaîne une ou plusieurs élongations « matta » (ou tatwil), correspondant au symbole «–». L'élongation se situe toujours à gauche du caractère courant. Si le trait d'allongement est associé à un caractère en position de début ou finale, le caractère prend sa forme de milieu et voit sa chasse augmenter du nombre de « matta » insérées (Tableau II-5) [Benamara 98]. Au niveau du PAW, l'insertion de traits d'allongement affecte uniquement sa largeur, la morphologie reste la même comme indiqué dans la figure II-3 [Trenkel 01]. Les éditeurs de texte tels que Word de Microsoft, insèrent dans les lignes de texte, le nombre approprié de « Matta », pour la justification gauche-droite d'un texte arabe.

Avec 6 mattas	Avec 3 mattas	Avec 1 mattas	Sans mattas
بـــــ	بـــ	بـ	ب
حـــــ	حـــ	حـ	ح
ســـــ	ســـ	ســ	س
فـــــ	فـــ	فـ	ف
مـــــ	مـــ	مـ	م
يـــــ	يـــ	يـ	ي

Tableau II-5- Exemples de caractères avec et sans matta.

ماطر ـ مـــاطر ـ مـــــــــاطـــــــــر

مـــاطر

Figure II-3- Variation du mot «ماطر» écrit avec un nombre différent de traits d'allongement dans différentes positions.

- 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 [Benamara 99]. 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. Elle 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 II-6.

ـلـ , ـمـ	ـمـ , ـجـ	ـفـ , ـجـ	ـقـ , ـجـ
ـلـ , ـجـ	ـمـ , ـحـ	ـفـ , ـحـ	ـقـ , ـحـ
ـلـ , ـحـ	ـمـ , ـخـ	ـفـ ,ـخـ	ـقـ , ـخـ
ـلـ , ـخـ			

Tableau II-6- Caractères susceptibles d'être ligaturés verticalement selon [Benamara 99].

ل م ج ة : lettres disjointes .

لمجة : ligatures obligatoires .

لمـــجة : ligature esthétique entre les 2 premières lettres.

لمجـة : ligature esthétique entre les 3 premières lettres.

Figure II-4- Exemples de ligatures horizontales et verticales.

Ce chevauchement modifie les dimensions des PAW et souvent la morphologie de certains caractères (figure II-4 et II-5). De plus, la ligne de base dans ce cas, n'est plus horizontale [Benamara 98]. Ce processus dépend fortement des fontes : il existe des fontes qui ne présentent aucune ligature telles que les fontes « mudir » et simplified arabic »,

d'autres qui possèdes un ensemble de ligatures complètement différentes les unes des autres [Benamara 99].

Figure II-5- Exemple de formes de PAWs sans et avec caractères ligaturés verticalement (respectivement à droite et à gauche de « ≡ »).

- Contrairement au latin, la notion de majuscule et de minuscule n'existe pas en arabe. Cependant, tous les attributs de mise en forme tels que gras, italique, souligné sont valables dans les lettres arabes (figure II-6).

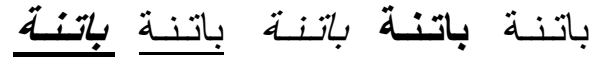


Figure II-6- le nom de ville « باتنة », de droite à gauche, en forme : normale, gras, italique, souligné et gras+italique+souligné.

- Les caractères arabes peuvent être voyellées. Les voyelles appelées aussi diacritiques dans certains documents tels que [Al-Badr 95] et [El_gammal 01] et courtes voyelles dans d'autres tels que [Burrow 04], 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 : عَـلـمٌ » ou « savoir : عِلْمٌ » ou encore « enseigner : عَـلَّمَ », selon sa voyellation.

Il existe 8 signes de voyellation qui peuvent se placer au dessus de la ligne d'écriture, tels que fathah (ـَـ) dhammah (ـُـ), soukoun (ـْـ) et chaddah (ـّـ) qui doit être accompagnée de l'une des vayellations fatha, Dammah ou kasrah, en dessous tels que Kasrah (ـِـ). De plus trois « tanwin » peuvent être formés à partir d'un double fatha (ـًـ), d'un double dhammah (ـٌـ) ou d'un double kasrah (ـٍـ).

Si en français 5 signes orthographiques (les accents grave, aigu et circonflexe, le tréma et la cédille) modifient certaines lettres, en arabe toutes les formes de consonnes sont

susceptible de porter chacune des huit signes de voyellation et souvent deux d'entre eux superposés (par exemple chaddah+voyelle et chaddah+tanwin). Outre cela et comme le montre ce paragraphe les caractères arabes voyellés nécessitent des matrices de dimensions importantes notamment en hauteur [Benamara 99].

II-2-2- ALPHABET ARABE : DONNEES GRAPHIQUES

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 [Benamara 99].

II-2-3- CONSEQUENCES TECHNIQUES DES CARACTERISTIQUES MORPHOLOGIQUES DE L'ARABE

La plupart des critiques apportées à l'écriture arabe sont faites, en général, par comparaison à l'écriture latine imprimée. On reproche à l'écriture arabe d'être sténographique, de ne noter qu'une partie des signes nécessaires à la lecture et donc d'obliger le lecteur à supplier les signes manquants, c'est à dire de savoir à l'avance ce qu'il doit lire, pourtant ces signes (voyelles) existent. Cependant, certains savants considèrent que la superposition des voyelles rend la lecture et l'écriture difficiles et lentes. En fait, ce qu'on reproche sur ce point à l'écriture arabe, c'est de ne pas avoir de voyelles incorporées dans le corps même des mots comme c'est le cas des langues latines où on est obligés de tout écrire (consonnes et voyelles), ce qui permet de tout lire. De plus nous avons vu qu'il existe une centaine de formes différentes de caractères arabes correspondants uniquement aux consonnes, sans aucune ligature ni voyellation, ni chiffre, ni signe de ponctuation. Ainsi les normes de claviers existants (90 à 120) sont largement dépassées. Toutefois, le changement

de forme d'un caractère en fonction de sa position dans le PAW, est géré dans les éditeurs de texte par un *analyseur de contexte* qui confère au caractère la bonne calligraphie.
C'est aussi par comparaison implicite ou explicite avec l'écriture latine que l'on reproche à l'arabe de ne pas avoir de lettres totalement isolées les unes des autres qui seraient utilisables par les techniques de composition de textes.

II-2-4- NOTIONS DE TYPOGRAPHIE ARABE

II-2-4-1- DEFINITION DE LA NOTION DE FONTE

Une *police* (fonte) est un ensemble de caractères d'une même famille, d'une même graisse et pour un corps donné. Ces caractéristiques typographiques sont normalisées dans l'imprimerie, tant au niveau du symbole (dimensions et dessin qui représente la forme et l'épaisseur du caractère), qu'à celui de la chaîne (mot : suite de symboles appartenant à la même fonte ou des fontes compatibles) dans chaque ligne du texte [Benamara 99].

- *La chasse* : comprend en plus de la largeur, l'espace entre caractères. La chasse dépend du dessin, du style et de la grosseur du caractère.
- *Le corps* : désigne la hauteur du caractère comprenant le blanc de séparation horizontale avec la ligne au dessus. Le corps varie en fonction de l'usage prévu pour le caractère : texte courant, titrage ou affiche. La dimension du corps s'exprime en points. Le point est l'unité de mesure typographique, équivalent 0.376 mm.

II-2-4-2- STYLES DE CALLIGRAPHIE ARABE

L'écriture arabe varie selon les milieux et les régions, d'une extrême simplicité formelle à la complexité exhaustive de l'arabesque. Ces deux formes les plus anciennes sont : l'une souple et cursive à l'origine du « Neskhi », l'autre plus raide et anguleuse, qu'on appela plus tard le « Koufi ». Ces deux écritures ont par la suite donné naissance à d'autres styles. La multiplicité de ces styles est due tout d'abord à la volonté des diverses populations converties à l'islam de conserver les textes coraniques et de les transcrire dans des styles adaptés à leur nature et à leur écriture d'origine. Elle est également due à l'adaptation de l'écriture aux dimensions et à la matière des supports, ainsi qu'au développement de la civilisation : on invente des styles spéciaux pour chaque usage (un style pour princesses, un pour l'administration, un autre pour la poésie etc). Elle est enfin due à certaines nécessités : pour des raisons stratégiques, on inventa par exemple le style pigeons voyageurs qui permettait de reconnaître l'expéditeur [Benamara 99].

Il existe une centaine de styles dont seulement quelques uns sont couramment utilisés dans le monde arabo-musulman, nous citons par exemple : le Neskhi, Thoulouthi, Roqa, Diwani, Koufi, Farsi… (Figure II-8). Le Neskhi demeure aujourd'hui la fonte la plus utilisée pour l'écriture imprimée.

Chaque style est régi par des lois particulières. D'un style à un autre les proportions d'une même lettre changent ; par exemple, dans le style Roqa la lettre alif : « ا » est plus petite que la lettre Ain : « ع », mais dans le style Koufi c'est l'inverse (figure II-7). De même les caractéristiques principales d'une lettre peuvent considérablement changer d'un style à un autre.

Figure II-7 – les lettres « ا» et « ع » dans les styles koufi et Roqa respectivement de gauche à droite.

Neskhi

C'est une écriture cursive, souple et arrondie, sans aucun angle brusque. C'est le style le plus usité dans les livres, les journaux, il est adapté à la machine à écrire et à l'imprimerie.

Thoulthi

Style cursif, il est le plus difficile en ce qui concerne son code et sa réalisation. Il est utilisé dans la décoration des monuments religieux et dans les compositions calligraphiques complexes, grâce à sa plasticité et à la possibilité d'étirement de ses lettres dans tous les sens.

Roqa

D'origine turque, ce style est l'écriture manuscrite orientale aux lettres courtes et ramassées. C'est une écriture qui comporte beaucoup de simplifications qui facilitent l'écriture manuelle mais qui ont un effet négatif sur la différenciation des signes. Actuellement, il est employé pour les gros titres de journaux de publicité.

Diwani

Style très lyrique, plein de souplesse, comportant souvent de grandes envolées gestuelles à la fin des mots. Il a été utilisé par les ottomans pour les lettres de la chancellerie. Maintenant il est d'emploi pour l'ornementation des certificats et en poésie.

Koufi

Considéré comme le plus ancien, le koufi est en général un style assez anguleux et géométrique. Il en existe de nombreuses variétés. Il sert le plus souvent à transcrire des textes religieux sculptés dans la pierre. Il est caractérisé par une base linéaire et des hampes montantes qui ont permis, en les tressant ou en les terminant par des éléments floraux, d'en faire un style très utilisé dans la décoration.

Il existe aussi des styles modernes qui viennent des styles classiques et qui ont été adaptés aux nécessités des moyens de reproduction. La figure II-8 Montre quelques exemples de fontes.

II-3- AVANCEES EN OCR ARABE

La reconnaissance l'écriture arabe (AOCR : Arabic OCR) remonte aux années 70, depuis, plusieurs solutions ont été proposées. Elle 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 [Benamara 99]. 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 II-8 (tiré de [Al-Badr 95] et [Benamara 99] et enrichi par des travaux récents), en fin de ce chapitre, 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.

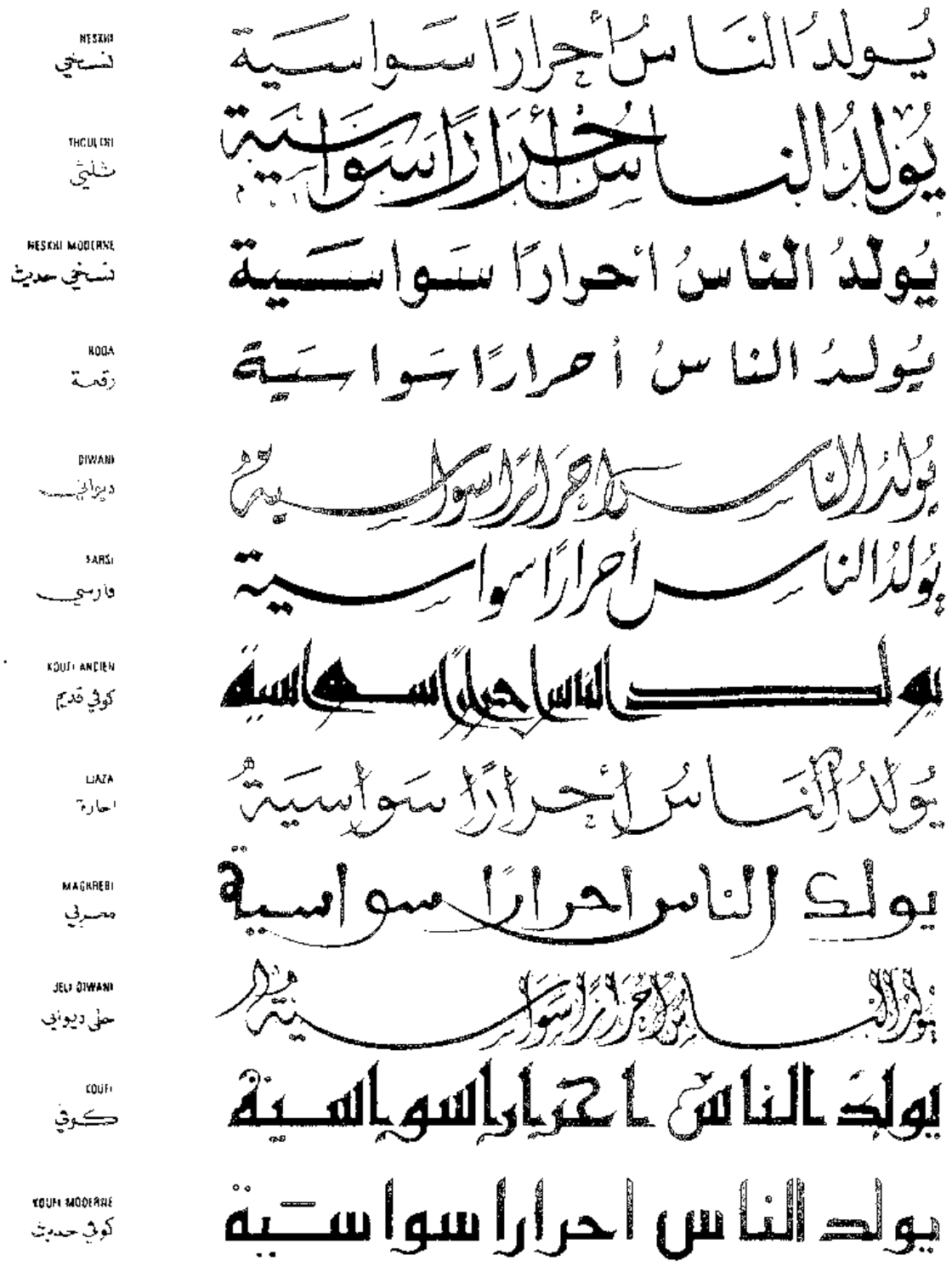


Figure II-8 – première phrase de la charte des droits de l'homme :« يولد الناس أحرارا سواسية » , répétée en huit styles différents

II-3-1- PRETRAITEMENTS

A ce stade, le problème classique est lié aux boucles qui risquent d'être bouchées ou ouvertes et aux points diacritiques qui peuvent être éliminés à la suite de certaines opérations de prétraitements ou encore confondus avec du bruit.
En effet, les prétraitements peuvent altérer surtout la forme des points diacritiques de manière à les confondre avec du bruit s'ils sont trop amincis. Ils peuvent également être accolés au corps du caractère associé à cause d'une dégradation ou d'une normalisation de taille. Un autre problème typique rencontré à la suite d'une mauvaise squelettisation, particulièrement dans le cas du manuscrit, est la confusion de deux point avec un seul ; très souvent, dans les deux cas on obtient un segment de droite [Altuwaijri 95].

Pour ces différentes raisons, dans la plupart des travaux, les points sont éliminés au début du traitement. Les étapes suivantes du traitement sont alors effectués sur le corps du caractère (ou du PAW), ainsi le nombre de formes considérés est réduit sensiblement, la phase de classification devient moins complexe et plus rapide. Pour retrouver l'identité exacte du caractère une fois son corps identifié, un algorithme d'assemblage corps/points est utilisé [Bushofa 97].

II-3-2- LA SEGMENTATION

La reconnaissance du PAW nécessite d'abord son extraction de la page, ceci suppose une décomposition de la page au préalable, qui consiste à retrouver la structure physique du document en délimitant les différentes parties homogènes (texte, graphe, photographie …).

1) Segmentation en lignes de texte

Les méthodes de traitement de l'arabe utilisent souvent la projection horizontale pour extraire les lignes. Cependant la présence des points/diacritiques complique cette extraction et conduit parfois à la fusion des paragraphes [El-Dabi 90]. Ce problème a lieu quand l'interligne est calculé par une simple moyenne des différents interlignes (figure II-9).

Pour remédier à ce problème, certains auteurs tels que [Parhami 91] identifient d'abord les différentes lignes d'écriture, ensuite regroupent les blocs de texte d'après leur proximité par rapport aux lignes d'écriture déjà localisées.

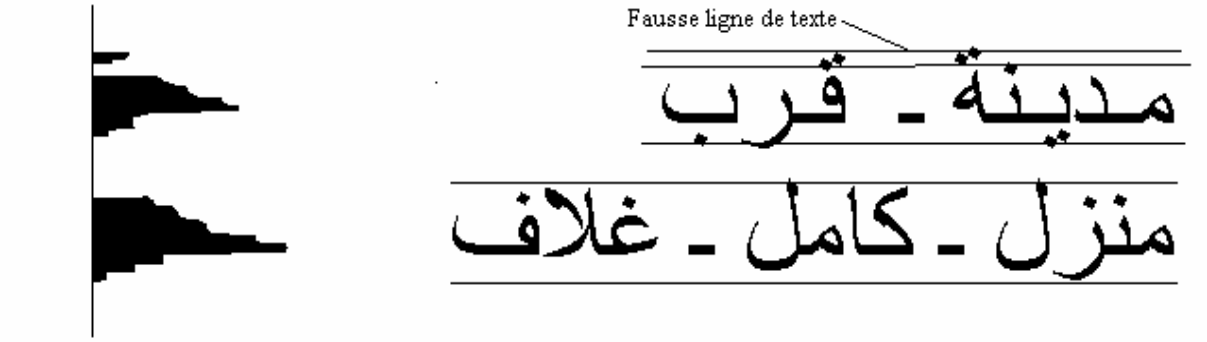


Figure III-9–
Exemple d'histogrammes horizontaux et d'une fausse ligne de texte qui en résulte .

Comme pour le latin, une fusion des lignes est aussi possible à cause des hampes et des jambages. Elle peut être accentuée dans le cas de l'arabe par la présence des points diacritiques. En cas de fusion, une méthode empirique de correction consiste à localiser d'abord la ligne qui contient le maximum de pixels noirs [Benamara 95]. Les parties au dessus et en dessous de cette ligne sont ensuite analysées en se basant sue les densités de pixels noirs des différentes lignes. Si la fusion a eu lieu par exemple dans la partie supérieure, la ligne ayant le minimum de densité de pixels dans cette partie, correspond à la frontière entre les lignes fusionnées.

2) Segmentation en PAWs

Elle est réalisée en déterminant l'histogramme des projections vertical des différentes lignes de texte [Amin 89]. Cependant, cette méthode n'est pas efficace dans le cas où les PAWs se chevauchent verticalement (figure II-10). Dans ce cas, d'autres techniques sont utilisées telles que la détermination du contour [Azmi 01], du squelette [Fahmy 01], ou encore des composantes connexes [Parhami 81]. Le choix de la technique est souvent guidé par la méthode d'analyse [Benamara 99].

كـرم ـ طرفة

Figure II-10 – exemple de chevauchement de PAWs
respectivement de droite à gauche entre « م, ر » et « ف, ر ».

3) Segmentation en mots

En AOCR, la segmentation est souvent réservée à l'extraction des PAWs ; le mot est plutôt considéré dans la phase de post-traitement (si elle est prévue) pour valider les résultats trouvés ou corriger les erreurs de reconnaissance. Par ailleurs afin d'éviter le problème d'une segmentation erronée en mots, certains auteurs introduisent dans leur systèmes, un seul mot à la fois [Amin 83] et [Al-Badr 95].

4) Segmentation en caractères

La segmentation en caractères (ou en graphèmes) constitue le problème le plus ardu lié à la reconnaissance de l'écriture arabe, ce sujet fait objet de notre recherche, et va être considéré en détails dans le chapitre suivant.

Les difficultés rencontrées à ce niveau sont du même type que celles affrontées lors de la reconnaissance du latin manuscrit (cursif), mais souvent plus complexes à cause de la diversité des formes du caractère arabe, de la courte liaison qui existe entre les caractères successifs, de l'allongement des ligatures horizontales et de la présence des ligatures verticales [Benamara 99].

II-3-3- EXTRACTION DES PRIMITIVES, CLASSIFICATION

La synthèses des travaux étudiés, montre que les différents types de primitives (structurelles, géométriques, statistiques, transformations globales, corrélations…) et les différentes méthodes de classification (statistiques, structurelles, syntaxique…) qui existent dans la littérature, ont été pratiquement toutes utilisées dans la description de l'écriture arabe.

Toutefois, nous constatons que le calcul des moments et l'utilisation des descripteurs de Fourier, pour l'extraction des primitives, sont appliqués dans un nombre relativement important de travaux [Alqaisy 85], [El-Sheikh 88], [Al-Yousefi 88], [El-Ramly 89], [El-Dabi 90], [Mahmoud 94] et [Miled 98]. Ces méthodes sont connues pour leur invariance à la translation, à la rotation et à l'homothétie, de plus, elles tolèrent les faibles variations de formes.

De même la classification en arbre de décision sont également populaires. Quatre arbres de décision sont élaborés, afin de déterminer l'identité du caractère selon sa position dans le PAW.

Les classifieurs connexionnistes constituent un nouveau paradigme en reconnaissance de formes, les travaux utilisant cette approche en AOCR, sont relativement récents [Amin 94], [Souici 97]. Les modèles utilisés par la majorité des travaux, appartiennent à la famille des réseaux à couches. Le principe des réseaux à couches est de transmettre l'information recueillie sur une couche d'entrée vers une couche de sortie qui exprime la réponse du réseau.
Par ailleurs, peu de travaux ont utilisés des méthodes de classification hybrides [Almuallim 87], [Abdelazim 89], [Khella 92], [Amin 94], [Bouslama 98]. Les études récentes en OCR recommandent cette approche, toutefois le choix ainsi que nombre de classifieurs, qui devraient être complémentaires, dépend de l'application considérée.

II-3-4 POST-TRAITEMENT

Des vérifications contextuelles classiques telles que la recherche dans un dictionnaire, les probabilités d'occurrence de bigramme et de trigramme…, sont appliquées dans les différents travaux qui prévoient un post-traitement.
La méthode du dictionnaire est traditionnellement simplifiée pour accélérer la recherche et réduire la complexité du calcul : le dictionnaire est construit à partir de mots réduit à leurs racines, les suffixes et les préfixes sont éliminés. Cependant des modèles sont élaborés afin de spécifier la relation racine-suffixe-préfixe [Benamara 99].
Par ailleurs, le post-traitement, malgré l'amélioration des scores qu'il peut apporter, n'est pas très utilisé en AOCR, ce qui peut s'expliquer par le manque de dictionnaires de validation et de statistiques élaborées par rapport au vocabulaire de référence. Or les statistiques sont relatives à l'application considérée et au vocabulaire de test.

II-4- CONCLUSION

Nous avons présenté dans ce chapitre, les principales propriétés morphologiques et typographiques de l'écriture arabe. Le manque de normalisation des typographies a montré la complexité de l'adaptation de l'écriture arabe aux exigences technologiques modernes. Donc la reconnaissance optique de l'arabe reste une tâche encore non résolue. La simplification des formes calligraphiques des caractères arabes faciliterait le problème de la composition de texte et en grande partie la tâche de l'AOCR.

Nous avons aussi passé en revue dans la dernières partie de ce chapitre certains travaux qui ont été réalisés dans le domaine de la reconnaissance de l'écriture arabe, les problèmes majeurs dans ce domaine se ramènent à la cursivité de l 'écriture et à la sensibilité de certaines caractéristiques topologiques de l'arabe à la dégradation, en l'occurrence les points diacritiques et les boucles.

Référence	Système	Approche	Segmentation	Primitives	Classification	Performance
[Abdelazim 89]	Hors-ligne, imprimé MF	Analytique	Externe	Structurelles / Statistiques	Structurelle / Statistique/arbre de décision	RC 99%
[Abdelazim 90]	Hors-ligne, imprimé	Analytique	Externe	Dimensions du graphème	Préclassification / mise en correspondance /reconstruction	RC 96%
[Abuhaiba 93]	Hors-ligne, manuscrit	Analytique	Externe	Structurelles	Transformation Off-line /on-line	-
[Al-badr 95b]	Hors-ligne, imprimé	Globale	-	Structurelles	Mise en correspondance spatiale de modèles de primitives	RM 73.13-99.39%
[Al-imami 90]	En-ligne, PAWs	Analytique	Externe	Structurelles	Arbre de décision	RM 86- 100%
[Aissaoui 94]	Hors-ligne, manuscrit	Analytique	Externe	Structurelles	Structurelle	-
[Aissaoui 96]	Hors-ligne, MF	Analytique	Externe	Statistiques	Réseaux de neurones	RC 64 – 100%
[Aissaoui 97]	Hors-ligne, MF	Analytique	Externe	Statistiques	Distance quadratique	RC 87.87 – 95.24%
[Alimi 94,95]	En-ligne, caractères isolés	-	-	Chaînes de codes	Programmation dynamique	RC 95%
[Almuallim 87]	Hors-ligne, manuscrit mot	Analytique	Externe	Statistiques	Syntaxique/ Distance	RC 91%

Tableau II-7 – 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

Référence	Système	Approche	Segmentation	Primitives	Classification	Performance
[Al-yousefi 92]	Hors-ligne, manuscrit.	Analytique	Externe	moments	Classifieur bayesien.	RC 99.5 %
[Ameur 93]	Hors-ligne, manuscrit MS	Analytique	Externe	structurelles	Arbre de décision	SC 98.9 % RC 83 %
[Ameur 94]	Hors-ligne, manuscrit	Globale	-	Structurelles	Dictionnaire	-
[Ameur 97]	Hors-ligne, manuscrit	Analytique	Externe	Structurelles/ statistiques	KNN	RC 82.5 %
[Amin 89]	Hors-ligne, MF	Analytique	Externe	Chaîne de codes	Arbre de décision	RC 90 %
[Amin 96]	Hors-ligne, caractères MS	-	-	Structurelles	Réseaux de neurones	RC 90-92 %
[Amin 97]	Hors-ligne, mots	Globale	-	Structurelles	Réseaux de neurones	RC 98%
[Azmi 01]	Hors-ligne, Perse	Analytique	Externe	-	-	SC 93-98.9 %
[Ben amara 95]	Hors-ligne	Analytique	Externe	geometriques	-	SC 99-100 %
[Bouslama 97]	Hors-ligne, caractères isolés	-	-	Structurelles/ variables linguistiques	Logique floue	RC 100 %
[Bouslama 99]	Hors-ligne, caractères isolés	Analytique	Externe	Fuzzy linguistiques	Logique floue	RC 100%
[El-Dabi 90]	Hors-ligne, imprimé	Analytique	Interne-SWS	Moments	Table de correspondance	RC 94 %

Tableau II-7 – Tableau récapitulatif précisant les caractéristiques et les performances de certains systèmes AOCR (suite).

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

Référence	Système	Approche	Segmentation	Primitives	Classification	Performance
[Elgammal 01]	Hors-ligne, imprimé	Analytique	Externe	-	Grammaire reguliere	RC 93.4 %
[El-Khaly 90]	Hors-ligne, imprimé	Analytique	Externe	moments	Distance	RC 95-100 %
[El-Sheikh 88]	Hors-ligne, imprimé	Analytique	Externe	Descripteurs de Fourier	Classifieur topologique	RC 99 %
[El-Sheikh 90]	En-ligne, caractères isolés	-	-	Structurelles	Arbre « handcrafted »	RC 99.6 %
[Fehri 94]	Hors-ligne, MF	Analytique	Interne	Structurelles/ statistiques	Programmation dynamique	RC 98 %
[Fehri 98]	Hors-ligne, manuscrit	Analytique	Externe	Structurelles	Réseaux de neurones/ HMM	-
[Gillies 99]	Hors-ligne, imprimé	Analytique	Externe	Structurelles	Réseaux de neurones	RC 89-93.1 %
[Goraine 94]	Hors-ligne, imprimé	Analytique	Externe	Chaîne de codes	Sruct./ Mesure géom./ contexte	RC 95.87 %
[Haj-Hassan 91]	Hors-ligne, imprimé	-	Externe	Structurelles	Syntaxique	RC 99 %
[Hassibi 94]	Hors-ligne, imprimé	Analytique	Interne	Structurelles	Réseaux de neurones	RC 99 %
[Jambi 93]	Hors-ligne, manuscrit	Analytique	Externe	Structurelles	Dictionnaire	SC 95 %
[Kurdy 93]	Hors-ligne, imprimé MF	Analytique	Externe	Structurelles	Morphologie mathématique	RC 98 %

Référence	Système	Approche	Segmentation	Primitives	Classification	Performance
[Mahjoub 96]	En-ligne, Caractères isolés	-	-	Statistiques	HMMs	RC 98.1 %
[Mahjoub 98]	En-ligne, MS	Globale	-	statistiques	DHMMs & NSHMMs	RC 90-93.5 % 1S RC 86-90 % MS
[Miled 96]	Hors-ligne, Manuscrit	Analytique	Externe	Structurelles	-	SC 98.52 %
[Miled 98]	Hors-ligne, Manuscrit	Analytique	Externe	Topologiques/ statistiques	HMMs	RC 79.5-82.5 %
[Mitiche 97]	Hors-ligne, Imprimé MF	Analytique	Externe	Structurelles	Distance	RC 90 %
[Motawa 97]	Hors-ligne, manuscrit	Analytique	Externe	Structurelles	Morphologie mathématique	SC 81.88 %
[Olivier 96]	Hors-ligne, Manuscrit Ms	Analytique	Externe	Chaîne de codes	-	SC 97.41 %
[Souici 97]	Hors-ligne, manuscrit	Analytique	Externe	Statistiques/ Structurelles	Réseaux de neurones	RC76.17-85.75%
[Trenkel 97,01]	Hors-ligne, imprimé	Analytique	Externe	Chaîne de codes	Rés. Neurones/ arbres	RC R.N 89.06% RC AR 90.68%
[Zahour 91]	Hors-ligne, manuscrit	Analytique	Externe	Structurelles	Dictionnaire	RC 86 %
[Zahour 98]	Hors-ligne, manuscrit	Analytique	Externe	Structurelles	Mise en correspondance	RC 87 %

Tableau II-7 – Tableau récapitulatif précisant les caractéristiques et les performances de certains systèmes AOCR (suite).

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

Référence	Système	Approche	Segmentation	Primitives	Classification	Performance
[Alaa 01]	Hors-ligne, manuscrit	Analytique	Externe	Structurelles	Réseaux de neurones	SC 69.72 %
[Burrow 04]	Manuscrit	Globale	-	KNN, moments	-	RC 94%
[Bushofa 97]	Hors-ligne, Imprimé	Analytique	Externe	-	-	SC 97.01 %
[Fahmy 01]	Hors-ligne, Manuscrit	Analytique	Externe	Géométriques	Réseaux de neurones Logique floue	RC 69.7 %
[Hachour 04]	Imprimé, caract. isolés	Analytique	Externe	Morphologiques/ statistiques	-	-
[Kandil 04]	Hors-ligne, imprimé	Analytique	Externe	-	-	-
[Kavianifar 99]	Hors-ligne, imprimé	Globale	-	Structurelles	PHMM	-
[Masmoudi 02]	Hors-ligne, manuscrit	Analytique	Externe	-	Réseaux de Neurones	-
[Menhaj 02]	Hors-ligne, imprimé	Analytique	Externe	-	Réseaux de Neurones	SC 100%
[Nawaz 03]	Hors-ligne, imprimé	Analytique	Externe	Moments	Rés. Neurones -	RC 76 %
[Sarfaz 03]	Hors-ligne, impr	Analytique	Externe	Moments		RC 73 %
[Sari 02]	Hors-ligne, Manuscrit	Analytique	Externe	-		SC 86 %

Tableau II-7 – Tableau récapitulatif précisant les caractéristiques et les performances de certains systèmes AOCR (suite et fin).

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

CHAPITRE III

ETAT DE L'ART DE LA SEGMENTATION

III-1- INTRODUCTION

Dans ce chapitre nous allons exposer les différentes techniques de segmentation, et en particulier la segmentation du mot en caractères. Pour les différentes écritures (à savoir cursive ou non cursive).

III-2- SEGMENTATION DE LA PAGE

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 etc).

Cette classification permet ensuite d'orienter la reconnaissance vers des systèmes spécialisés dans l'analyse de chaque type de media [Belaid 97].

Une étude détaillée sur les techniques utilisées dans l'analyse de documents se trouve dans : ([HU 93], [Etemad 94], [Haralick 94], [Belaid 97], [Tang] et [Mao 00]).

III-3- SEGMENTATION D'UN BLOC DE TEXTE EN LIGNES

Cette étape consiste à séparer les différentes lignes du texte pour en extraire les mots puis les caractères composants les mots. La plupart des études proposées dans ce domaine s'appuient sur une décomposition de l'image en composantes connexes [Bennasri 99].

D'autres par contre utilisent des techniques s'appuyant en grande partie sur les histogrammes des projections horizontale [Al-badr 95]. Et certains auteurs optent pour des méthodes spécialisées telle que celle utilisée par Bennasri et Al dans [Bennasri 99] pour la segmentation en lignes de l'écriture arabe manuscrite.

III-4- SEGMENTATION DES LIGNES EN MOTS

La segmentation en mots est réalisée en déterminant l'histogramme des projections verticales des lignes pour détecter les espaces entre les mots et pouvoir les séparer. Cependant cette technique peut ne pas être efficace dans certains cas où les mots se chevauchent (cas par exemple de l'écriture arabe). Dans ce cas d'autres techniques sont utilisées telles que : le suivi du contour, détermination du squelettes ou la détermination des composantes connexes …

III-5- SEGMENTATION DES MOTS EN CARACTERES

La segmentation des caractères est une opération qui tente de décomposer une image de séquence de caractères (mot) en sous-images de symboles individuels. C'est l'un des processus de décision dans un système de reconnaissance optique de caractères. Son but est de décider si un motif isolé d'une image (caractère ou autre entité identifiable du mot) est correct ou non [Casey 96].

III-5-1- ORGANISATION DES METHODES

Certains auteurs tels que Tappet et Al dans [Tappet 90] parlent de segmentation *interne* et *externe*, dépendant de si la segmentation se fait séparément ou simultanément avec la reconnaissance. D'autres auteurs tels que Dann et Wong dans [Wong 92] 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 [Case 96] la classification des méthodes suivant l'utilisation ou non de la reconnaissance durant la phase de segmentation n'est pas un 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.

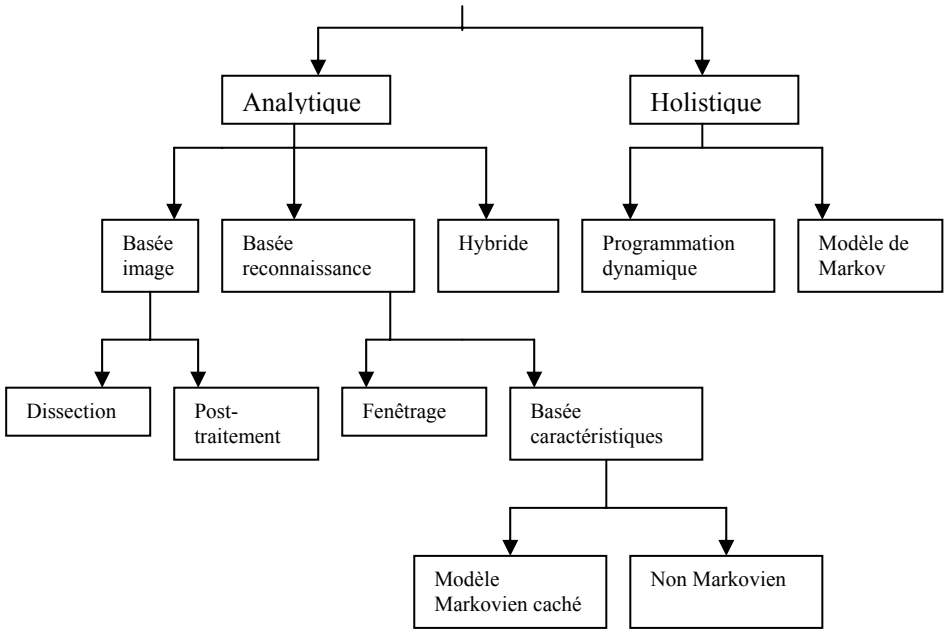


Figure III-1 : Hiérarchie des méthodes de segmentation selon R.G.Casey.

III-5-2- TECHNIQUES 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 [Casey 96]. 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.

1) Dissection directement en caractères :

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 [Yuan].

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 [Ha 96].

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 [Al-badr 94].

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.

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 [Casey 96].

- ***Analyse les « bounding boxes »*** **:**

  La distribution des boites informe beaucoup sur le segmentation des caractères non cursifs, en testant leur relation d'adjacence pour les fusionner ou leur taille et aspect pour les séparer [Casey 96]. 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 [Casey 96].
  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 [Ha 96].

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 [Bulmenstein 98a, 98b et 99].

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 [Sayre 73] 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 [Lecolinet 91] où la dissection est basée sur la détection des ligatures. L'algorithme de segmentation comportait deux étapes :

- la détection des zones de segmentation possibles
- 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 détectés en analysant le profile supérieur et inférieur et l'ouverture des concavités des mots.

  Plusieurs autres méthodes sont exposées en détail dans [Casey 96].

III-5-3- SEGMENTATION BASEE 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 [Casey 96]. Ces méthodes sont aussi appelées « méthodes sans segmentation » (« segmentation-free methods ») [Al-badr 94].

1) ***<u>Les méthodes cherchant l'image</u>*** **:**

La segmentation basée reconnaissance ici s'effectue en deux étape.

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 [Casey 96].

Une méthode utilisant une reconnaissance combinant la programmation dynamique et les réseaux de neurones était proposée dans [Burges 92]. 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 si 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.

2) *méthodes qui segmentent le représentation des caractéristiques de l'image* :

a) 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.

Pour plus de détails sur les H.M.M le lecteur peut se référé à [Anigbogu 92], [Benamara 96, 99, 2000] et [Miled 2001].

b) *Approches non Markovienne* :

Ces méthodes se sont inspirées des concepts utilisés dans la reconnaissance d'objets. Ici différentes caractéristiques et leurs positions d'occurrences sont enregistrés par image. Chaque caractéristique contribue à l'évidence de

l'existence d'un ou plusieurs caractères dans une position d'occurrence [Al-badr 94]. 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. [Casey 96]. 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.

Nous n'avons cité que deux méthodes, mais plusieurs sont exposées dans : [Kozima 93], [Casey 96], [Kosawat 2003] et [Coüasnon 96].

III-5-4- STRATEGIES MIXTES (SUR-SEGMENTATION)

Cette famille de méthodes utilise aussi le 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 (ie : 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 [Casey 96].cette technique a été utilisée pour la segmentation des mots arabes. Nous pouvons citer quelques références employant cette technique [Trenkle 95, 97, 2001] , [Gillies 99] et [Ayat 2000].

III-5-5- STRATEGIES 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.

La façon de comparer le mot inconnu avec les mots de référence fait la différence entre les différentes méthodes utilisant cette approche [Casey 96].

III-6- CONCLUSION

Dans ce chapitre nous avons essayé d'exposer les différentes méthodes utilisées dans la segmentation de mot. Ces méthodes ont connu beaucoup de progrès ces derniers temps. Des techniques variées influencées par l'évolution dans les domaines tels que la reconnaissance de la parole et la reconnaissance en ligne des caractères ont émergés.

La difficulté dans la réalisation d'une segmentation performante dépend généralement de la nature du document à lire et de sa qualité. Le taux de mauvaise segmentation croit progressivement à partir de l'écriture imprimée à l'écriture manuscrite jusqu'à l'écriture manuscrite cursive où la difficulté devient plus importante.

La performance d'un système de reconnaissance de l'écriture ne dépend pas seulement des résultats de la phase de segmentation, mais aussi du type de classifieur utilisé pour la reconnaissance de ces segments.

CHAPITRE IV
SEGMENTATION DES MOTS ARABES EN CARACTERES

IV-1- INTRODUCTION

Dans ce chapitre, nous allons présenter un état de l'art sur la segmentation des caractères arabes, ainsi qu'une présentation des différents travaux effectués dans ce domaine. Nous allons aussi exposer plusieurs algorithmes de segmentation couvrant la majorité des méthodes utilisées dans la segmentation des caractères arabes imprimés.

IV-2- ETAT DE L'ART DE LA SEGMENTATION DES MOTS ARABES EN CARACTERES

IV-2-1- INTRODUCTION

Après la phase de pré-traitement, la majorité des systèmes OCR isolent les caractères individuels avant de les reconnaître. Ceci est effectué durant la phase de segmentation.

Segmenter une page de texte peut être divisé en deux étapes : la décomposition de la page et la segmentation des mots. Lorsqu'on travaille avec des pages contenant différents types d'objets tels que les graphiques, formules mathématiques, blocs de texte … la décomposition de la page consiste à séparer les différents éléments de la page, produisant ainsi à partir des blocs de texte, des lignes et des pseudo-mots (PAWs).

La segmentation des mots quant à elle consiste à séparer les caractères d'un pseudo-mot. Les performances d'un système d'OCR dépendent généralement de comment sont isolés les caractères.

IV-2-2- DECOMPOSITION DE LA PAGE

La décomposition de la page est un sous_domaine de l'analyse de documents. L'analyse de documents étudie la structure des documents et identifie ses différentes parties logiques. Une étude détaillée sur l'analyse de documents se trouve dans [Belaid 97].

Dans certaines bibliographies traitant la langue arabe, la décomposition de la page est limitée à la séparation des différentes lignes d'un bloc de texte et l'extraction des pseudo-mots.

IV-2-3- SEGMENTATION DES MOTS

Les méthodes de segmentation de l'écriture latine cursive ont été étudiées de façon extensive. Bien que ces méthodes puissent être applicables à l'écriture arabe, mais elles sont généralement insuffisantes pour la segmentation de l'écriture arabe.

Lorsqu'on examine les méthodes de segmentation des mots arabes, on peut les classer selon cinq approches [Al-Badr 95].

1) Dans la première approche on suppose que le mot en entrée est déjà segmenté en caractères.
2) Dans la seconde approche le mot en entrée est segmenté en primitives plus petites que le caractère.
3) Dans la troisième approche le mot en entrée est segmenté en caractères.
4) Dans la quatrième approche le mot est reconnu de telle sorte que la segmentation soit un sous-module du module de reconnaissance.
5) Dans la cinquième approche le mot est reconnu comme un tout, sans segmentation.

IV-2-3-1- PREMIERE APPROCHE

Utilisée principalement dans le cas de l'écriture des caractères isolés. Mais comme les caractères isolés sont rarement utilisés dans l'écriture arabe, sauf à l'exception dans les formules mathématiques. Cette approche n'est utilisée que pour des cas particuliers. De tels systèmes nécessitent un sous système de segmentation qui identifie les caractères à l'intérieur du mot, avant de le reconnaître.

IV-2-3-2- SECONDE APPROCHE

Cette approche segmente un PAW cursif à tous les endroits qui paraissent être des points de connexion. Dans ce cas il est possible que le PAW soit découpé en primitives plus petites que le caractère tels que les traits, points d'intersection, les points d'inflexion et les boucles. Le schéma habituel de reconnaissance ici est de reconnaître les primitives puis les combiner en caractères, comme par exemple dans [Trenkle 01].
Cette approche est utilisée pour la segmentation de l'écriture en-ligne. Et l'écriture hors-ligne amincie.

Dans certains systèmes utilisant cette approche la segmentation consiste à examiner le squelette du PAW et de le découper aux points de connexions. Ces points correspondent aux points caractéristiques qui peuvent inclure : la fin d'une ligne, les points où les lignes se croisent, les points dans lesquels le trait change rapidement de direction…[Al-Emami 90], [Goraine 92].

D'autres systèmes segmentent le PAW en se basant sur l'histogramme des projections vertical. Où les points de segmentation représentent les points où l'histogramme descend en dessous d'un certain seuil prédéfini [Tolba 90].

L'avantage de ce type de méthodes est qu'il est plus facile d'identifier un ensemble potentiel de points de connexion qui peuvent inclure tous les points de connexion actuels que d'identifier directement les points de segmentation. Il est ensuite du rôle de la phase de classification de décider quels sont les points de segmentation, suivant la connaissance à priori qu'elle possède. Cette méthode est particulièrement appropriée pour la reconnaissance de l'écriture manuscrite où les bordures des caractères sont ambiguës.

IV-2-3-3- TROISIEME APPROCHE

Cette approche essaye de segmenter correctement un mot en caractères puis reconnaître les caractères. Dans cette approche l'étape de segmentation devient l'étape la plus critique dans le processus de reconnaissance. Beaucoup de techniques utilisées dans cette approche sont similaires à celles utilisées dans l'approche précédente, mais modifiées pour prévenir de la dissection du caractère en plus d'une partie.

El-Khaly et Sid-Ahmed dans [El-Khaly 90] segmentent un mot aminci en caractères en suivant la ligne de base du mot, en détectant quand les pixels commencent à monter au dessus ou descendre en dessous de cette ligne.

Le système IRAC II par exemple proposé dans [Amin 82], segmente à la fin d'une dent (سن). Mais comme un caractère peut comporter plusieurs dents (exemple 3 dents dans le cas de la lettre ســ) et comme les caractères comportant une seule dent doivent avoir des points diacritiques au dessus ou en dessous. Le système examine les points se trouvant autour de la dent pour décider de segmenter ou non.

D'autres méthodes cherchent les points de segmentation le long de la ligne de base, en utilisant l'histogramme des projections verticales. Ces points sont définis comme étant les endroits où l'histogramme descend en dessous d'un certain seuil [Benamara 95], [Fakir 93]. Mais comme ces descentes peuvent se trouver à l'intérieur même d'un caractère, les

chercheurs utilisent différentes méthodes pour prévenir de rupture d'un caractère en plus d'une partie. Par exemple certains chercheurs utilisent des règles heuristiques qui préviennent la segmentation de caractères de largeur inférieure à une certaine valeur [Amin 89], [Amin 91]. D'autres modifient l'histogramme de projections verticales en multipliant chaque entrée par la hauteur de la colonne relative à la ligne de base. Ceci a pour effet d'amplifier la distance de la ligne de base de manière à ne pas considérer les points loin de la ligne de base comme points de connexion [Tolba 90].
Plusieurs méthodes de segmentation du mot en caractères sont présentées en détail dans [Al-Badr 95].

IV-2-3-4 QUATRIEME APPROCHE

Cette approche reconnaît les caractères d'un mot connectés sur place (ie sans segmentation préalable). Quelques systèmes qui adoptent cette méthode commencent à l'extrême droite d'un pseudo-mot et examinent un ensemble de colonnes (de largeur égale au caractère le plus proche) et essayent de le reconnaître. Si la reconnaissance échoue. Ils ajoutent itérativement d'autres colonnes, jusqu'à reconnaissance du caractère. Une fois un caractère reconnu, il est enlevé du sous mot et le processus est répété [Abdelazim 90].
Le problème dans ce type d'approches est que si le système échoue à la reconnaissance d'un caractère dans n'importe quelle partie du mot et spécialement au début le reste du mot ne sera pas traité. Pour y remédier les chercheurs utilisent une reconnaissance arrière de gauche à droite et est déclenchée lorsque le système n'arrive pas à reconnaître un caractère au milieu [Al-Badr 95b].

IV-2-3-5 CINQUIEME APPROCHE

Les systèmes de cette approche reconnaissent un mot comme une entité. Il utilisent généralement des techniques de comparaisons globales pour comparer les mots en entrée à d'autres stockés dans une base de données. Cependant cette approche est limitée à la reconnaissance d'un ensemble de mots prédéfini (exemple : commandes d'ordinateurs dans les ordinateurs basés stylo), et n'est pas pratique pour la reconnaissance générale de texte à riche vocabulaire [Amin 82].

IV-2-4- ENUMERATION DE CERTAINS TRAVAUX DE SEGMENTATION DE MOTS ARABES EN CARACTERES

Plusieurs méthodes de segmentation ont été développées en utilisant les différentes techniques énumérées précédemment. Dans ce qui suit, nous allons présenter des travaux de différents auteurs sur la segmentation de textes arabes.

a) ***B.Al-Badr et R.Harlick dans [Al-Badr 95b] :*** Proposent une méthode pour la reconnaissance de caractères arabes imprimés sans segmentation préalable. La technique est basée sur la description des symboles en terme de primitives de forme et ce durant une phase d'apprentissage. Ces primitives représentent des modèles auxquels seront comparés les mots lors de la reconnaissance. Au temps de la reconnaissance, les primitives sont détectées sur l'image du mot, en utilisant des opérations de morphologie mathématiques. Le système compare les primitives détectées aux symboles modèles. Le mot candidat peut ressembler à un ensemble de symboles modèles, alors aux résultat obtenus sont affectés des probabilités représentant le taux de ressemblance du mot avec le modèle. Le système choisis le modèle ayant la plus haute probabilité comme reconnaissance du mot.

b) ***S.Al-Emami et M.Usher dans [Al-Emami 90]*** *:* proposent un système de reconnaissance en-ligne de caractères arabes. L'écriture est acquise en utilisant une tablette graphique et enregistrée dans une matrice sous forme de coordonnées de chaque position du stylo par rapport au premier point du mot. Le processus de segmentation est appliqué aux points. Au début trois points sont choisis à partir de la matrice de points.
La distance entre le premier point et le second et l'angle entre les trois points sont calculés, si l'ange et la distance sont supérieurs à des valeurs seuils alors le second point est considéré comme point de segmentation, sinon le troisième point est considéré comme second point et au troisième point est affectée la valeur du quatrième point et le processus est répété jusqu'à ce que tous les points de segmentation du mot soient retrouvés. Chaque segment est ensuite catégorisé selon sa direction (les auteurs ont pris les quatre directions Est, West, Nord et Sud). Les segments adjacents ayant la même direction sont alors concaténés et la pente de chaque segment est enregistrée. Les points diacritiques sont représentés par des drapeaux. Toutes les informations sur les segments trouvés sont enregistrées dans un arbre dit de décision, construit pendant la phase d'apprentissage. Cet arbre est utilisé entrée de la phase de reconnaissance.

c) ***A.Amin et H.B. Al-sadoun dans [Amin 92] :*** l'image du texte subit un prétraitement dans l'ordre d'en produire le squelette. A partir de l'image amincie, un arbre binaire est construit. Dans les nœuds de l'arbre sont décrits les informations sur les formes correspondantes aux parties composantes du pseudo-mot. ces informations sont, des pointeurs vers les nœuds père et fils et des champs contenant des primitives décrivant la structure de l'image binaire (les primitives sont définis à partir du code de Freeman) en plus d'informations complémentaires telles que le nombre de points au dessus ou en dessous de la ligne de base s'ils existent, les zigzag (hamza) … Après construction de l'arbre binaire. Il est en suite réduit, pour minimiser le nombre de nœuds. La segmentation du pseudo-mot consiste à diviser l'arbre le représentant en sous-arbres, de façon que chaque sous-arbre représente un caractère du pseudo-mot. ce sont ces sous graphes qui seront pris en compte par le classifieur pour la reconnaissance des caractères.

d) ***R.Azmi et E.Kabir dans [Azmi 01]*** **:** proposent un algorithme de segmentation des mots perses imprimés en caractères en se basant sur l'étiquetage conditionnel. après quelques prétraitements dont la détermination du contour supérieur du mot et la détection et l'ajustement de la ligne de base. Le contour est analysé point par point. A chaque point est affecté une étiquette, selon la position du point par rapport à la ligne de base (c à d le point est au dessus, en dessous ou au milieu de la ligne de base) et l'étiquette du point précédent. Un ensemble de règles est appliqué au contour libellé pour trouver les points de segmentations. Un post-traitement est ensuite effectué pour prévenir de la sur-segmentation de certains caractères.

e) ***N.Benamara et N.Ellouze dans [Benamara 95] :*** proposent une méthode de segmentation en caractères des mots arabes imprimés, basée sur la détection d'un ensemble de caractéristiques morphologiques de type hampe, jambage, boucle et point diacritique. La détection de chacune des primitives (dans l'ordre de leur citation) est supervisée par un vérificateur-correcteur d'erreurs basé sur un ensemble de règes de morphologie arabe. La vérification met en cause les limites de la segmentation à chaque fois que le caractère non homogène au sens de la topologie, sont regroupés ensemble.

f) B.M.F.Bushofa et M.Spann dans [Bushofa 97]: proposent un algorithme de segmentation de caractères arabes imprimés en utilisant les informations sur leur contour. La méthode commence par trouver le nombre total de lignes dans la page et détermine pour chaque ligne la ligne de base, les zones supérieure et inférieure de la ligne et calcule la distance entre les limites de ces zones et la ligne de base pour calculer un seuil à partir de ces valeurs. Pour chaque pseudo-mot le contour est tracé point par point et en sauvegardant les coordonnées de chaque point. Dans la procédure de segmentation, ils considère trois cas différents, la segmentation des caractères qui se touchent et qui ne devrait pas se toucher. Pour cela ils cherchent dans la partie inférieure du contour l'endroit où les caractères se touchent et les séparent. Le deuxième cas est la segmentation du caractère «ـي» se situant à la fin du mot, ce caractère est détecté à part et séparé de son voisin. Le troisième cas est la segmentation de caractères joints et ce en examinant la partie supérieure du contour pour trouver les points de segmentation. Ces points sont choisis en suivant le mouvement du contour, lorsque le contour monte puis redescend jusqu'à une valeur inférieure au seuil calculé précédemment, ce point est considéré comme point de segmentation. La procédure continue jusqu'à ce que toutes les coordonnées du pseudo-mot s'épuisent. La reconnaissance des caractères se fait au fur et mesure de la segmentation pour valider ou rejeter un segment et re-segmenter.

g) A.M. El-Gammal et M.A.Ismail dans [El-Gammal 01] : proposent une méthode de segmentation en caractères des mots arabes imprimés, basée sur la représentation du texte en utilisant un graphe appelé graphe d'adjacence de ligne (« line adjacency graph (LAG)»). La segmentation est accomplie en considérant la relation entre la ligne de base du texte et ce graphe. Le texte est segmenté en scripts (un script est considéré comme étant l'unité de reconnaissance, il peut correspondre à un caractère, une partie du caractère ou plus d'un caractère (dans le cas par exemple de caractères ligaturés verticalement)). Le LAG associé au pseudo-mot est parcouru en utilisant un algorithme de parcours pour détecter les sous-graphes associés à chaque script. Le classifieur analyse les sous graphes associés au script pour le reconnaître.

h) H.Goraine et al dans [Goraine 92] : proposent un système de reconnaissance hors-ligne de caractères arabes. Pour la segmentation des mots les auteurs utilisent avant des prétraitements dont l'amincissement pour obtenir une image binaire contenant des lignes d'un seul pixel d'épaisseur. Les mots sont segmentés en traits. L'algorithme de

segmentation est basé sur un principe de segmentation angulaire, il impose que la direction de la courbe entre les deux extrémités d'un trait ne doit pas excéder un certain angle seuil. Lors de la classification les traits sont analysés pour supprimer les redondances puis codés. Chaque chaîne de traits en entrée est comparée à un ensemble de références jusqu'à correspondance exacte avec le modèle, là la chaîne est sauvegardé dans une table et le processus est répété pour une autre chaîne jusqu'à reconnaissance de tout le mot.

i) A.Hamid et R.Haraty dans [Hamid 01] : proposent un algorithme neuro-heuristique pour la segmentation des mots arabes manuscrits. Le processus comporte six étapes. La première concerne l'acquisition du texte, la seconde sa binarisation en utilisant un algorithme heuristique générant une matrice prenant pour valeur 1 pour un pixel noir et 0 pour un blanc. La troisième étape consiste à extraire les pseudo-mots. La quatrième étape consiste à extraire les caractéristiques à partir de chaque colonne de la matrice de pixels. Les caractéristiques sont de type topographique. La cinquième étape est une génération de points de pré-segmentation. C'est une sur-segmentation basée sur les caractéristiques extraites pour chaque colonne de la matrice et d'une largeur approximative des caractères. La sixième étape est une vérification de la validité des points de pré-segmentation et pour cela un réseau de neurones de 52 entrées, 4 couches cachées et une sortie. Les entrées sont les attributs de caractéristiques d'un point de pré-segmentation et la sortie est la validité de ce point.

j) M.Kavianifaret A.Amin dans [Kavianifar99] : proposent une méthode de reconnaissance globale de l'écriture arabe imprimée. Après acquisition du l'image du texte, cette dernière subit plusieurs prétraitements, dont : une opération de seuillage dont le but est de choisir un seuil parmi d'autres existants et dans l'image du texte les pixels au dessus de ce seuil sont considérés comme fond (background). Les composants connectés sont aussi trouvés et regroupés dans des bounding boxes. Ces composants sont ensuite regroupés, l'algorithme de regroupement essaye de le concaténer avec un ensemble de groupes existants pour trouver des mots valides. Le texte est aussi ajusté. Après les prétraitements les caractéristiques sont extraites. Le procédé consiste à déterminé le contour de chaque pseudo-mot des mots existants, ce contour est analysé et classifié en trois types (contour du corps principal, contour de caractères complémentaires ou bruit) à la fin de cette étape un tableau d'informations sur le

contour du pseudo-mot de chaque mot est construit. A partir de ce tableau sont détectés les contours désignant des corps de pseudo-mots puis les contours représentant des caractères complémentaires ainsi que les coordonnées de leurs points d'occurrence. A la fin de cette extraction des caractéristiques un fichier de sortie est produit. Il contient toutes les informations concernant chaque mot du fichier image. Chaque ligne du fichier contient les informations suivantes : le nom du mot, le nombre pseudo-mots, nombre de pics dans l'histogramme de pseudo-mot, type et position des caractères complémentaires et le nombre de boucles.

k) M.B.Menhadj et al dans [Menhadj 02] : proposent un algorithme de segmentation et reconnaissance simultanée de textes arabe/perse. Après prétraitement, les lignes de texte sont détectés puis vient le tour de la segmentation reconnaissance. L'algorithme consiste à détecter chaque pseudo-mot et à tracer le contour de son corps. Ensuite il scrute autour de ce corps les détails (points et diacritiques). L'étape suivante de l'algorithme consiste à segmenter le corps du pseudo-mot en caractères en cherchant dans le contour supérieur du pseudo-mot les points minimaux locaux et les marques ainsi que les points de début et de fin du mot. Chaque segment est délimité par deux marques successives. Pour prévenir d'une éventuelle sur-segmentation des caractères, une procédure dite de validation des points de segmentation est utilisée. Cette procédure cherche les détails autour du segment et le soumet à reconnaissance, si le caractère n'est pas reconnu elle lui ajoute un segment voisin et répète la même procédure (chercher les détails et essayer de le reconnaître) jusqu'à ce que le caractère soit reconnu. Le module de reconnaissance est un classifieur neuronal.

l) S.T. Masmoudi et al dans [Masmoudi 02] : présentent l'étape de segmentation associée à un modèle basé PHMM (modèles Markoviens cachés planaires) pour la reconnaissance hors-ligne des noms de villes tunisiennes manuscrites. Le procédé consiste en trois étapes. La première est une segmentation horizontale qui consiste à diviser l'image du texte en cinq zones logiques : la zone de points diacritiques supérieurs, la zone de points diacritiques inférieurs, la zone de hampes, la zone de jambages et la zone médiane. Après cela il y'a détection des boucles qui caractérisent la zone médiane. La seconde étape est appelée segmentation naturelle. Elle consiste à séparer les différents pseudo-mots en détectant l'espace entre les pseudo-mots dans la zone médiane. La troisième étape est appelée segmentation verticale. C'est l'étape la

plus difficile dans le processus de segmentation. Elle consiste à examiner le contour supérieur du pseudo-mot pour localiser les points minimaux locaux, et les considérer comme points de segmentation. Le résultat de cette étape est un ensemble de graphèmes qui peuvent correspondre à un caractère, un caractère sur-segmenté (une partie du caractère) ou à un caractère sous-segmenté (deux caractères ou plus comme dans le cas de caractères ligaturés verticalement).

m) D.Motawa et al dans [Motawa 97] : décrivent un algorithme pour la segmentation des caractères arabes manuscrits. L'image scannée subit d'abord un redressement de l'écriture, puis les éléments connectés sont emboîtés dans des *« bounding boxes »*. l'algorithme utilisé pour déterminer les limites des *bounding boxes* est basé sur l'analyse des lignes de pixels de largeur prédéfini ligne par ligne pour déterminer les composants connectés. Chaque boite subit une segmentation. L'algorithme de segmentation consiste à déterminer les singularités et les régularités dans le pseudo-mot. Et ceci en supprimant toutes les lignes horizontales pour déterminer les singularités et en supprimant les singularités pour trouver les régularités. Les points de segmentation sont sensés appartenir aux régularités. Ils correspondent aux points ayant la densité minimale de pixels dans un bloc de pixels noirs. Si la densité de pixels est uniforme, alors le point de segmentation est le point extrême droit de régularité.

n) C.Olivier et al dans [Olivier 96] : ont développé un algorithme de segmentation en graphèmes des mots arabes manuscrits, en analysant le contour de chaque pseudo-mot pour générer des codes de variation du signal produit. Ces codes sont analysés de droite à gauche par un automate qui détermine les régions probables de segmentation notées PSP (Primary Segmentation Points). Un ensemble de critères est appliqué à ces régions pour en extraire les points décisifs de segmentation notés DSP (Decisive segmentation points).

o) M.Sarfaz et al dans [Sarfaz 03] : proposent un système de reconnaissance hors-ligne de caractères arabes imprimés. Après l'acquisition et le prétraitement de l'image du texte, vient la phase de segmentation, qui consiste à déterminer les limites des lignes, déterminer les zones du milieu de haut et de bas de la ligne et détecter la ligne de base en utilisant l'histogramme de projections horizontal. Chaque ligne est segmentée en pseudo-mots en utilisant l'histogramme de projections vertical. Chaque pseudo-mot est

ensuite divisé en caractères individuels. La procédure de segmentation consiste à trouver l'histogramme de projections vertical de la partie du milieu de la ligne, là où l'histogramme devient inférieur au 2/3 de l'épaisseur de la ligne de base, cet endroit est considéré comme un point de segmentation. La procédure est répétée sur tous les pseudo-mots de ligne. Après la segmentation il y'a extraction des caractéristiques dans ce case les auteurs ont utilisés un calcul des moments et d'autres descripteurs de forme. La classification se fait via un réseau de neurones artificiel à partir des caractéristiques calculés et les segments détectés.

p) ***F.Sari et al dans [Sari 03] :*** proposent un algorithme pour la segmentation des caractères arabes manuscrits appelé ACSA (Arabic Character Segmentation Algorithm). Après acquisition de l'image du texte, les mots sont normalisés puis lissés pour réduire le bruit et régulariser le contour des mots. Les pseudo-mots sont extraits de chaque mot par suivi du contour extérieur. L'étape suivante est l'extraction des caractéristiques et là ils utilisent différents types de caractéristiques dont on peut citer les boucles, hamza, hampes, jambages …, ces caractéristiques sont stockées dans une liste. Après détection de la ligne de base, en utilisant l'histogramme des projections horizontal. La ligne d'écriture est divisée en trois zones, une zone supérieure comportant les hampes, une zone médiane comportant les jambages et une zone médiane comportant le corps principal de l'écriture. Chaque pseudo-mot est ensuite considéré séparément pour être segmenté en caractères. Les points de segmentation sont considérés comme les points minimaux du contour extérieur inférieur du pseudo-mot. après détection de ces points, ils sont validés dans une autre étape en leur appliquant un ensemble de règles pour en générer un ensemble de points de segmentation valides (représentant une segmentation idéale en caractères). Ces segments sont remis au classifieur ainsi que leurs caractéristiques pour être reconnus.

q) ***J.trenkle et al dans [Trenkle 01] :*** proposent un algorithme pour la segmentation en caractères des mots arabes imprimés issus d'un FAX. Après quelques prétraitements dont la suppression des bruits et des bandes de fax et après s'être assuré que c'est de l'imprimé et non du manuscrit. L'algorithme de segmentation consiste à sur-segmenter le pseudo-mot de façon à produire des segments atomiques ne dépassant pas le caractère. Ces segments sont ordonnés de gauche à droite. Ils sont combinés en groupes de deux à cinq segments consécutifs. L'ensemble de segments contient les

segments atomiques et combinés. Parmi cet ensemble de segments se trouvent les segments représentant les caractères idéaux du pseudo-mot. Il est du rôle du classifieur d'identifier ces segments et de les reconnaître.

IV-3- ETUDE DE L'EXISTANT

Vu que le domaine de segmentation en caractères des mots est très vaste et qu'il existe des classes de méthodes pour chaque type d'écriture. Pour pouvoir balayer une multitude de facettes de la calligraphie de l'écriture arabe imprimée, nous avons choisis d'étudier les méthodes traitants l'écriture imprimée multifonte, dans le cas voyellé ou non et aussi avec et sans ligature verticale.

En examinant la bibliographie concernant la segmentation en caractères de l'écriture arabe imprimée, nous avons constaté quatre classes de méthodes. Une des classe utilise comme base de sa segmentation l'histogramme des projections vertical, une seconde classe utilise une sur-segmentation des mots normalisés puis regroupe les segments pour trouver le caractères, une troisième utilise comme base le contour des mots pour les segmenter en caractères et une quatrième utilisant une représentation à base d'arbres. Pour cela nous avons choisis comme exemple un algorithme de chaque classe traitant chacun une particularité de l'écriture arabe.

IV-4- CHOIX DE L'APPROCHE ET DES ALGORITHMES

Nous avons vu dans le paragraphe IV-2-3 que pour la segmentation des mots arabes il existait cinq approches différentes. La première approche assumait que les mots étaient déjà segmentés à l'entrée. La seconde segmentait les mots en primitives plus petites que le caractère et qu'elle était plus appropriée au cas du manuscrit. La troisième segmentait le mot en caractères pour les reconnaître, c'est l'approche adoptée pour les cas de notre étude. La quatrième approche reconnaissait les mots sans segmentation préalable, en utilisant des primitives morphologiques et des modèles pour la comparaison. Le problème avec cette approche est que la définition des primitives dépendait de la taille des caractères et de leur fonte ce qui limitaient le nombre de fontes et les tailles de caractères. Dans la cinquième approche des mots entiers étaient reconnus sans segmentation. Ce qui posait le problème de limite du vocabulaire.

Donc Dans notre cas d'étude nous avons choisi des algorithmes adoptant la troisième approche c.à.d segmentant les mots imprimés en caractères.

Pour le choix des algorithmes nous avons rencontré dans notre recherche bibliographique différents algorithmes. Beaucoup d'algorithmes utilisent le même principe et différent dans de petits détails. Alors nous avons opté pour ces quatre algorithmes parce qu'ils utilisent chacun une technique complètement différente de l'autre et nous semble couvrir l'ensemble des algorithmes utilisés pour la segmentation des caractères arabes imprimés. Le premier algorithme utilise un découpage en utilisant l'histogramme des projections vertical et traitant des textes sans voyelles ni ligatures verticales. Le second utilise une sur segmentation puis réuni les segments pour trouver les caractères idéaux et traite des textes sans voyelles ni ligatures verticales. Le troisième segmente les mots en caractères et utilise pour cela une représentation arborescente, il traite des textes arabes voyellés et ligaturés verticalement. Le dernier segmente les mots en utilisant leur contour et traite des textes perses sans voyelles ni ligatures verticales

IV-5- ETUDE DETAILLEE DE QUELQUES ALGORITHMES SEGMENTANT LES MOTS ARABES IMPRIMES EN CARACTERES

Dans ce paragraphe, nous allons détailler quelques algorithmes de segmentation en caractères des textes arabes imprimés. Nous avons choisi les algorithmes touchant différentes facettes des caractéristiques morphologiques de l'écriture arabe telles que les textes simples, textes contenant des ligatures verticales, textes voyellés et l'écriture perse. Chaque algorithme utilise une technique différente pour la segmentation, et parmi eux un algorithme qui utilise une reconnaissance simultanée à la segmentation.

IV-5 -1- ALGORITHME PROPOSE DANS [Benamara 95] :

Cet algorithme est conçu de façon à accomplir une segmentation sans reconnaissance explicite des caractères individuels. L'algorithme suit les étapes suivantes :

- Identification des lignes de texte.
- Segmentation des lignes en PAWs.
- Séparation des PAWs en composants connectés.
- Localisation approximative des limites des différents caractères dans le PAW.
- Calcul du maximum de segments noirs dans une ligne de pixels.
- Extraction des primitives.
- Utilisation d'un contrôleur d'erreurs pour la détection des erreurs de segmentation.

1) *Segmentation des lignes du texte* :

Le texte arabe est segmenté en lignes, en utilisant l'histogramme de projections horizontal où les pics correspondent aux lignes de texte et les minimums correspondent aux espaces entre les lignes. Une procédure de rectification est utilisée pour éliminer les fausses lignes de texte, correspondants aux blocs d'épaisseur inférieure à un certain seuil et les lignes sont concaténés. Un autre problème peut survenir lors de la séparation des lignes où un caractère dans la ligne inférieure soit assez haut pour entrer dans le champ de la ligne directement au dessus. Pour résoudre ce problème, la ligne de plus haute densité est localisée. Les parties inférieure et supérieure sont contrôlées, s'il y'a fusion des lignes, on implémente la procédure suivante : en supposant que la ligne fusionnée est dans la partie supérieure, on identifie dans cette partie la ligne ayant la plus haute et la plus basse densité en pixels noirs. Cette dernière est considérée comme limite entre les deux lignes fusionnées. La même procédure est utilisée su la fusion est détectée dans la partie inférieure. Si plus de deux lignes sont fusionnées, l'opération est répétée récursivement jusqu'à ce qu'il n'y ait plus de détection de fusion.

2) *Segmentation des lignes de texte en mots ou pseudo-mots* :

Dans cette phase chaque ligne de texte est segmentée en pseudo-mots (PAWs). Pour cela l'histogramme des projections vertical de la ligne est déterminé afin de trouver les différents blocs qui sont séparés par au moins une colonne vide. Les parties contiguës sont séparées. Ces parties peuvent être des mots entiers, dans le cas où les mots s'écrivent à partir de lettres attachées, ou des parties de mots, dans le cas où les mots contiennent des lettres qui de nature s'écrivent isolés. Les pseudo-mots subissent ensuite un redressement et lissage.

3) Séparation des pseudo-mots en composants connectés :

Cette étape consiste à séparer le pseudo-mot en composants connectés, qui sont des groupes de segments noirs ayant une connexion commune. L'image du pseudo-mot est

traitée ligne par ligne de pixels. Quand des segments noirs sont détectés, chaque groupe de pixels connectés ensemble est étiqueté. Le module donne ensuite les coordonnés

maximales et minimales des rectangles qui englobent chaque composant, le nombre de pixels noirs associés à chaque composant et la ligne de base du corps du pseudo-mot. Ceci permet d'identifier les points s'ils existent et de stocker leur nombre ainsi que leurs emplacements par rapport à la ligne de base.

4) Segmentation de pseudo-mot en caractères

C'est l'étape la plus délicate dans le procédé de segmentation. Le procédé de segmentation se fait comme suit. Dans une première étape sont choisis approximativement les limites des caractères dans le pseudo-mot en utilisant l'histogramme de projections vertical paramétré par la largeur verticale de l'écriture (calculée par une procédure interactive d'apprentissage de fontes I.F.L.P). Ce qui permet de ne garder dans l'histogramme que les valeurs supérieures à l'unité seuil (ne garder que les pics). Ensuite le nombre maximal de segments noirs que contient chaque supposé caractère est calculé. Dans une ligne de pixels prise au milieu et à travers une distance proportionnelle à la largeur de l'écriture, chaque caractère ne doit avoir qu'un seul segment noirs. Sauf pour certains caractères se trouvant à la fin ou isolés et pour les caractères tels que «س » et tous les caractères ayant une boucle dans leur forme, qui peuvent avoir plus d'un segment noir à la fois. Pour ces caractères, une valeur seuil dépendant de la largeur horizontale de l'écriture doit exister entre deux segments successifs. Pour les caractères isolés ou à la fin, le résultat de la segmentation est analysé ; si la largeur du dernier caractère est plus courte que la largeur approximative d'un caractère (calculée par la procédure I.F.L.P), le processus de segmentation est arrêté.

5) Extraction des caractéristiques

L'extraction des caractéristiques est très importante pour la reconnaissance des formes, car c'est la clé de la reconnaissance d'un objet inconnu. Après la pré-segmentation précédente, on procède par l'extraction des caractéristiques dans les différentes zones trouvées. Les caractéristiques choisies dans ce cas sont : hampe,

jambage, boucle et points. Chaque primitive est codée selon les différentes formes qu'elle peut avoir. A chaque fois qu'une primitive est extraite, un contrôleur d'erreurs vérifie s'il y'a compatibilité avec les caractéristiques de la Librairie de caractères arabes déjà existante sans essayer de le reconnaître.

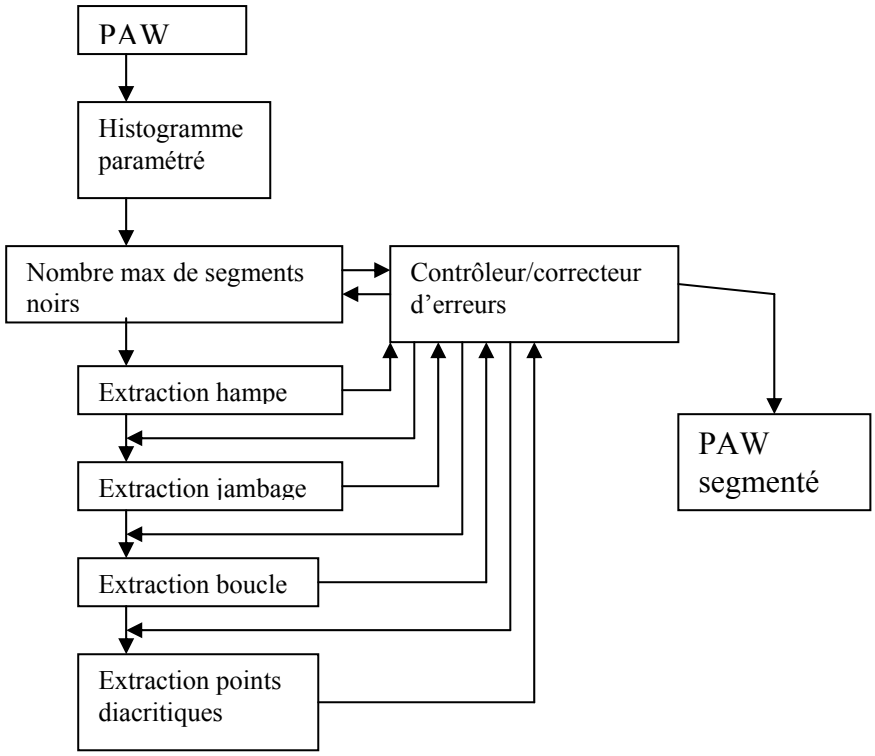


Figure IV-1- Diagramme de l'algorithme de segmentation des PAWs En caractères

Ce programme peut détecter différents types d'erreurs s'il y'a fausse segmentation dont :

- Un caractère ne peut pas avoir un jambage au début ou au milieu du pseudo-mot.
- Un caractère ne peut pas avoir une double hampe au début ou au milieu du pseudo-mot.
- Un caractère ne peut pas avoir de points à la fois au dessus et en dessous à n'importe quelle position du pseudo-mot.
- La largeur du caractère n'est jamais inférieure à la largeur horizontale de l'écriture.

Le programme suit les étapes montrées dans la figure IV-1. A chaque fois qu'une primitive est extraite, il y'a contrôle d'erreurs. Si une erreur se produit, les limites sont corrigées et les caractéristiques extraites avant cette étape sont revues.

6) ***Résultats expérimentaux :***

La digitalisation du texte se fait en utilisant un scanner HP scanjet 3P, connecté à un PC 486, avec une résolution de 300 dpi.

Le produit a été testé sur des textes contenant 500 à 1000 mots imprimés avec plusieurs fontes, telles que Neskhi, Bagdad et Mehdi en différentes tailles. Le résultat de segmentation était entre 99% et 100%.

IV-5-2- ALGORITHME PROPOSE DANS [Gillies 99]

A l'arrivée, l'image du texte est traitée par un module de décomposition de page. Ce module commence par la séparation des éléments graphiques et le nettoyage de la page. Ensuite il décompose les parties restantes en blocs de texte. Chaque bloc est segmenté en lignes individuelles de texte arabe. L'image de la ligne est normalisée à une hauteur de 40 pixels puis passée au module de segmentation des mots.

1) L'algorithme de segmentation :

Ce module exécute une sur-segmentation des mots, ayant pour but de produire des segments atomiques qui ne sont pas plus grands qu'un simple caractère. Lorsque cela est effectué, la segmentation idéale (un caractère idéal) peut être produite à partir de la combinaison des segments atomiques, dans des groupes appropriés. Le regroupement doit être fait de manière à maintenir la relation spatiale entre les segments atomiques. Ceci est réalisé par un algorithme spécial appelé « Vitebri beam search algorithm » (cet algorithme est détaillé dans [Gillies99 et Trenkle01]). Cet algorithme est un sous-module du module de reconnaissance.

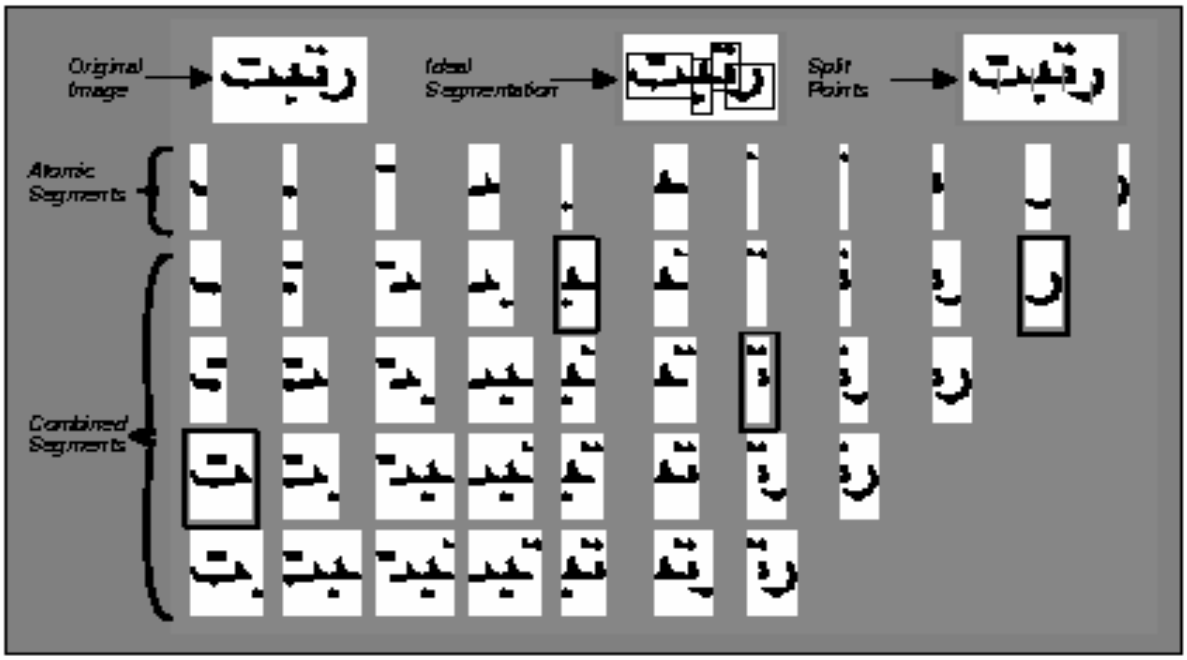


Figure IV-2- segmentation du texte.

Le procédé de segmentation commence d'abord par la détection des composants connectés de l'image du texte. Chaque composant est analysé pour voir où il y'a besoin de découper. Les points de segmentation sont calculés de façon heuristique, en utilisant deux méthodes. La première suggère comme points de segmentation les points où une fonction objective f(x) atteint un minimum local.

$$f(x) = \max (B - Top(x), 0) + \max (Bottom (x) - B , 0)$$

Top(x) : la coordonnée Y du pixel le plus haut de la colonne X du composant.

Bottom(x) : la coordonnée Y du pixel le plus bas de la colonne X du composant.

B : La coordonnée Y de la ligne de base de l'image normalisée.

La seconde méthode de découpage (scission) cherche un minimum local dans la coordonnée Y lors du tracé de la moitié haute du contour du composant. Les deux méthodes produisent souvent des points de scission identiques ou près de l'identique. Dans ce cas les points redondants sont rejetés et les segments atomiques sont ordonnés de gauche à droite dans des « bounding boxes ». Les segments sont combinés en groupes de deux à cinq segments atomiques consécutifs. L'ensemble entier de segments contient les segments atomiques et les segments combinés. Il contient K<5 N segments, où N est le nombre de

segments atomiques. Parmi les K segments se trouvent les caractères idéaux du pseudo-mot. Les K segments sont ensuite passés au module de reconnaissance pour être reconnus.

2) Résultats expérimentaux

Le système a été testé sur une totalité de 40 pages d'images. Dont 20 digitalisées à une résolution de 200x200 dpi et les mêmes 20 pages digitalisés à 200x100 dpi, en utilisant à chaque fois deux scanners différents de 200dpi.

Les résultats suivants ont été obtenus.

Ensemble de données	Résolution	pages	Lignes	Mots	caractères	Taux en %
Ens1	200x200	10	442	3495	17165	92.9
Ens2	200x200	10	447	3689	18129	93.2
Total	200x200	20	889	7184	35294	93.1
Ens1	200x100	10	442	3495	17165	90.0
Ens2	200x100	10	447	3689	18129	88.6
Total	200x100	20	889	7184	35294	89.3

Tableau IV-1- résultats expérimentaux de l'algorithme de [Gillies 99].

IV-5-3- ALGORITHME PROPOSE DANS [Elgammal 01]

Le processus de segmentation extrait les scripts de base d'une ligne de texte. Où le script est considéré comme l'unité de reconnaissance. Il peut correspondre à un caractère ou plus d'un caractère (dans le cas de caractères ligaturés verticalement). La segmentation d'une ligne de texte en scripts est basée sur la relation topologique entre la ligne de texte représentée sous forme d'un graphe et la ligne de base du texte.
L'algorithme comporte deux étapes : la détection de la ligne de base et l'extraction des pseudo-mots et scripts.

1) ***Détection de la ligne de base*** **:**

La détection de la ligne de base se fait en utilisant l'histogramme de projections horizontal. Elle correspond au sommet global de l'histogramme. La ligne de base est paramétrée par deux valeurs : « basetop » et « basebottom » qui représentent respectivement le rang haut et le rang bas de la ligne de base. Ces valeurs sont prises de telle manière qu'un certain pourcentage de pixels noirs de la ligne de texte se situe entre ces deux valeurs. La hauteur de la ligne de base est minimisée.

2) ***Extraction de pseudo-mots et scripts*** **:**

Chaque ligne est représentée par un graphe appelé « line adjacency graph » (LAG). Un LAG est un graphe comportant des nœuds représentants les pistes horizontales de pixels. Les nœuds de deux pistes appartenant à deux lignes adjacentes et se chevauchant sont liés par un arc. Le processus de construction d'une représentation LAG pour une ligne de texte produit un ensemble de sous-graphes isolés, représentant chacun le LAG d'un pseudo-mot de la ligne considérée. C'est pour cela que l'isolation des mots s'effectue en même temps que la construction des LAGs. Il peut correspondre à une combinaison de points diacritiques, un caractère isolé ou un pseudo-mot constitué de plusieurs caractères.

Le LAG associé à chaque pseudo-mot est ensuite transformé en un autre graphe de même type que le premier et comportant un nombre minimum de nœuds. Le nouveau graphe est appelé LAG compressé (C-LAG). Les nœuds du nouveau graphe sont

étiquetés « path » et « junction ». Manifestement les nœuds « path » ne sont jamais adjacents entre eux, mais les nœuds « junction » peuvent l'être.

La relation entre les nœuds du C-LAG et la ligne de base est très importante dans la phase d'extraction des caractéristiques et dans l'extraction des scripts. Chaque nœud du graphe est étiqueté de l'un des labels suivants, qui reflètent sa relation avec la ligne de base.

a) au dessus de la ligne de base.
b) en dessous de la ligne de base.
c) sur la ligne de base.
d) Au dessus de la ligne de base et y est connecté.
e) En dessous de la ligne de base et y est connecté.
f) Traversant la ligne e base.

A tout nœud « path » peut être assigné n'importe lequel de ces labels. A un nœud « junction » peut être assigné seulement les trois premiers labels. Les deux règles suivantes sont appliquées aux nœud du graphe étiqueté, pour éviter le problème d'incertitude dans la détection de la ligne de base, due à la variation des fontes, l'inclinaison de l'image ou tout autre facteur pouvant affecter la détection de la ligne de base.

Règle 1 :

Un nœud junction qui est adjacent à un autre nœud junction sur la ligne de base doit être libellé comme étant sur la ligne de base. Bien qu'il peut ne pas se trouver physiquement entre « basetop » et « basebottom ». cette règle doit être appliquée récursivement.

Règle 2 :

Si un nœud path est adjacent à un nœud junction qui est libellé sur la ligne de base. Ce nœud sera libellé au dessus ou en dessous de la ligne de base selon leur emplacement et connecté à la ligne de base, même s'il n'est pas actuellement sur la ligne de base. Cette règle est appliquée après la règle 1 mais non récursivement.

Pour trouver les sous-graphes correspondants aux scripts, le C-LAG est en premier lieu parcouru commençant par n'importe quel nœud path qui est libellé au dessus de la ligne de base, jusqu'à ce qu'on arrive à nœud junction libellé sur la ligne de base. Dans

ce cas les nœuds traversés constituent le sous-graphe d'un script et appelé aussi script. Le critère d'arrêt de l'algorithme de parcours du graphe est la rencontre d'un nœud junction libellé « sur la ligne de base ». Ce dernier sera considéré comme point de repère entre les scripts d'un même pseudo-mot. Le processus de parcours du C-LAG est répété jusqu'à détection de tous les scripts.

A ce stade seul le corps script est traité. Les points et les diacritiques ne sont considérés qu 'en phase de classification.

3) ***résultats expérimentaux*** :

Dans une des expérimentations, deux groupes de pages ont été utilisés. Le premier groupe contient 31 pages qui ne contiennent pas de diacritiques (voyellations). Tandis que le second contient 15 pages de texte voyellé. Toutes les pages ont été prises de cinq magazines arabes comportant plus de 10 fontes différentes et de taille variant entre 10 et 16 cpi. L'acquisition de ces pages est faite en utilisant un scanner de résolution de 300 dpi. Le premier groupe a donné un taux de reconnaissance de 95.2%. tandis que le deuxième un taux de 94.1%.

Deux autres classifieurs ont été utilisés pour la reconnaissance des point et des voyellations. Le premier reconnaît quatre classes de diacritiques (1 point, 2 points, 3 points et « hanza ») le second reconnaît 14 classes de diacritiques (les 4 premières classes + 10 voyellations).

Groupe de pages	classifieur	Taux de reconnaissance
Gr 1	4 classes	97.3 %
Gr 1	14 classes	94.7 %
Gr2	14 classes	91.7 %

Tableau IV-2- résultats expérimentaux de l'algorithme de [Elgammal 01].

Une autre expérimentation a été effectuée pour l'évaluation de l'algorithme sur des pages imprimées en utilisant des imprimantes laser d'une résolution de 300 dpi (de qualité inférieure à celle du cas précédent). Le taux de reconnaissance obtenu pour le script était de 94.6 % pour les points était de 96.6 %. Le taux de reconnaissance final était de 93.4 %.

IV-5-4- ALGORITHME PROPOSE DANS [Azmi 01]

Avant d'être segmentée l'image du texte subit quelques prétraitements dont la détection de la ligne de base et la détermination du contour.

1) *prétraitements* :

1-1- calcul de la taille du stylo :

Pour trouver la taille du stylo, un texte est parcouru colonne par colonne. La taille de pixels noirs la plus fréquente dans ces colonnes est prise comme taille du stylo (PS).

1-2- détection de la ligne de base globale :

La ligne de base est défini comme étant une ligne horizontale traversant toute une ligne de texte et dont la largeur est égale à PS. Et contenant le nombre maximal de pixels noirs.

1-3- ajustement de la ligne de base locale :

La technique utilisée pour l'ajustement de la ligne de base locale associée à chaque pseudo-mot est comme suit :

Le contour du pseudo-mot est tracé, il est représenté par sous forme des 8 directions de Freeman, avec une distance de $\frac{PS}{2}$ autour du bord supérieur de la ligne de base globale. La rangée de l'image contenant le maximum séquences de code 4 (notée n_4) est considérée comme limite supérieure de la ligne de base locale (notée ubl pour « upper bound of local base line »). La limite inférieure (lbl) est trouvée de façon similaire, en cherchant la rangée ayant le maximum de séquences de code 0 (notée n_0) en parcourant le bord inférieur de la ligne de base globale. Si la largeur de la ligne de base locale résultante est supérieure à 1.25xPS, alors si $n_4 > n_0$ alors ubl est retenu et lbl est décalé vers le haut jusqu'à ce que la largeur de la ligne de base locale devienne égale à PS. Sinon, si $n_4 \leq n_0$ alors ubl est décalé vers le bas de la même manière que précédemment.

2) Algorithme de segmentation

2-1- Etiquetage du contour :

La technique de segmentation proposée est basée sur l'étiquetage conditionnel du contour supérieur de chaque pseudo-mot. Le contour est tracé de droite à gauche en CCW (dans le sens contraire des aiguilles d'une montre), chaque point du contour est libellé dépendant de sa distance de la ligne de base et le label du point le précédent. Ces étiquettes sont u, m et d pour « Up », « middle » et « down » respectivement. Le processus d'étiquetage est montré dans la figure suivante :

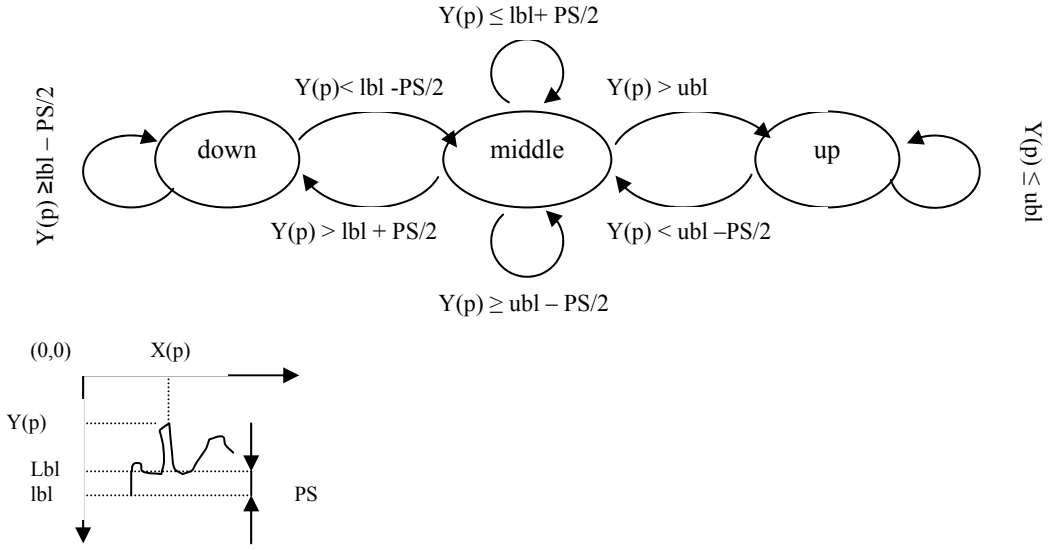


Figure IV-3- étiquetage conditionnel des points du contour

L'étiquette du premier point d'un contour est toujours u. les points voisins ayant la même étiquette forment un chemin. Si un chemin est plus court que PS/(2+1) il est lié au chemin précédent. De cette manière tout le contour est représenté par une chaîne de chemins étiquetés de droite à gauche.

2-2- segmentation en caractères :

Un point potentiel de segmentation est défini comme étant le dernier point d'un chemin de «m», qui satisfait les conditions suivantes :

- Le chemin m est plus long que PS/(2+1).
- Le chemin précédent est un chemin u plus long que PS.
- Le chemin suivant est plus long qu'une valeur seuil égale à 1.5xPS pour un chemin u et 4xPS pour un chemin d.

Cet algorithme de segmentation a tendance à sur-segmenter certains caractères et cela pour deux raisons : la première est que l'algorithme est sensé ne rater aucun point de segmentation si possible et la seconde est qu'il y'a certaines formes du corps d'un caractère qui ressemble à d'autres caractères. Cependant il est possible de laisser la correction de ces erreurs à la phase de classification. Mais il est plus pratique de les corriger pendant une phase de post-traitement. Les auteurs ont préféré prévoir une phase de post-traitement pour rectifier les erreurs de segmentation.

3)- *post-traitement* :

Dans cette phase les erreurs se produisant fréquemment sont détectées et corrigées de la manière suivante :

g1	ب پ ت ث ک گ
g2	ه ا
g3	د ذ
g4	م
g5	سد شد
g6	س ش
g7	ﺑ ﭘ ﺗ ﺛ ﻧ ﻳ
g8	صد ضد

Tableau IV-3- groupes de caractères (g1 – g8).

a) Les caractères dans g1, lorsqu'ils se trouvent à la fin d'un pseudo-mot peuvent avoir un chemin u qui cause un faux segment. Seulement les caractères de g2 ont un

chemin u similaire qui produit un segment correct. Le premier caractère est détectable par se hauteur et le deuxième par sa boucle, ce qui permet de détecter un faux segment et de le connecter à son voisin.

b) La même chose se produit pour les caractères de g3, elle est détectée par sa petite largeur.

c) Le caractère de g4 est divisé en deux segments. Le segment gauche est détectable par son étiquette et sa hauteur.

d) Les caractères de g5 sont divisés en 2 ou 3 segments. Si trois caractères de g7 sont joints ensemble, leur corps ressemble au corps des caractères de g5. pour pouvoir différencier les deux cas, il est nécessaire de localiser les points (simple, double ou triple). Après détection des points les erreurs sont corrigées par l'une des alternatives suivantes :

- Si au début d'un pseudo-mot, après un chemin u il y'a deux dents avec aucun point ou un triple point au dessus, les dents sont concaténés ensemble pour former un caractère de g5. La même chose pour trois dents au milieu du pseudo-mot.
- Si à la fin d'un pseudo-mot, il y'a deux dents sans points ou avec un triple point au dessus, suivi par un chemin u et un chemin d plus long que 3xPS ils sont concaténés pour former les caractères de g6.
- S'il y'a une dent sans points au dessus ou en dessous, elle est connectée à son voisin de droite cas des caractères de g8.

4- ***Résultats expérimentaux*** :

L'algorithme de segmentation a été testé sur un ensemble de textes imprimés dans vingt fontes différentes. L'ensemble de tests contenait 11347 caractères, dont 8056 connectés. Les résultats suivants ont été constatés :

Le taux de bonne segmentation avant le post-traitement pour les caractères connectés était de 91% et pour tous les caractères 93%.

Le taux de bonne segmentation après le post-traitement, pour les caractères connectés 98.5% et pour tous les caractères 98.9%.

IV-6- CHOIX D'UNE METHODE POUR L'IMPLEMENTATION

Les méthodes exposées précédemment ont montré leur efficacité du point de vue score de bonne reconnaissance. Mais pour ce qui est complexité, taille mémoire et temps d'exécution, chacun possède son inconvénient. En effet le principe utilisé dans les deux premiers algorithmes est très simple. Mais les erreurs de segmentation ne sont traitées qu'en phase de reconnaissance. De plus pour ce qui est du premier algorithmes les caractéristiques obtenues ne peuvent être exploités que par un classifieur de type modèles Markoviens cachés. Les deux derniers sont assez efficaces mais leur problème est dans la quantité de variables utilisées. Pour le dernier Le dernier plusieurs arbres sont utilisés pour représenter un seul caractère, or Pour représenter des milliers de caractères les structures doivent être énormes. Pour le troisième algorithme c'est le nombre de variables utilisées pour représenter les différentes variations de directions dans un même caractère. Ceci conduira impérativement à un temps assez important lors de l'exécution du programme sur des pages de texte pleines.

Pour notre implémentation, nous avons opté pour la simplicité tout en tenant compte de prise en charge des erreurs de segmentation avant la phase de reconnaissance. Nous allons utiliser un algorithme basé sur le principe d'histogramme de projections vertical, et nous prévoyons de corriger le maximum d'erreurs de segmentation par un module de post-traitement. L'algorithme est détaillé dans le chapitre suivant.

IV-7- CONCLUSION

Dans ce chapitre, nous avons abordé le domaine de segmentation des textes arabes. Dans la première partie, nous avons défini ce qu'est la segmentation et quelles sont les différentes étapes suivies pour la segmentation d'une image comportant du texte arabe dans le cas général. Nous avons vu précisément les méthodes utilisées pour la segmentation des mots en caractères (ou primitives inférieures au caractère). Nous avons terminé notre étude par exposer en détails quelques algorithmes utilisés pour la segmentation des mots arabes imprimés tout en précisant les taux de reconnaissance de chaque algorithme. Nous avons terminé par exposer les avantages et inconvénients de chaque méthode.

CHAPITRE V
CONTRIBUTION A LA SEGMENTATION DES MOTS ARABES IMPRIMES EN CARACTERES.

.

V-1- INTRODUCTION

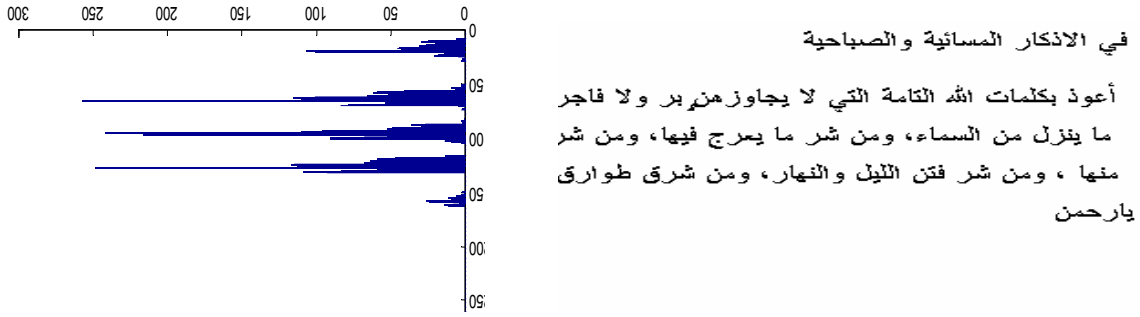
Dans ce chapitre nous présentons une méthode pour la segmentation en caractères des mots arabes imprimés. Cette méthode commence d'abord par quelques prétraitements visant l'amélioration de la qualité de l'image à traiter, ensuite le bloc de texte est analysé pour en extraire les lignes. Chaque ligne est ensuite traitée séparément pour en extraire les pseudo-mots. Ces derniers sont stockés dans une liste et sont segmentés un à un en caractères.

L'algorithme de segmentation segmente les pseudo-mots en caractères individuels ou fragments de caractère. Cet algorithme est suivi par une phase de post-traitement ayant pour but de corriger les erreurs de segmentation, pour enfin obtesnir des pseudo-mots segmentés en caractères individuels parfaits.

V-2- ACQUISITION ET PRETRAITEMENT

Le texte est scanné avec une résolution de 200 dpi et est stocké sous forme d'image binaire. Avant d'être analysée l'image subit quelques prétraitements, dont le redressement, le lissage et la normalisation. Les lignes du texte sont ensuite détectées en utilisant l'histogramme de projections horizontal (figure V-1).

Lorsque l'image est scanné plusieurs défauts peuvent apparaître. L'image est d'abord filtrée pour enlever les artefacts causés par le scanner, puis l'écriture subit un lissage pour le nettoyage de l'image en enlevant les parties supplémentaires et en complétant les parties manquantes. L'écriture est ensuite normalisée à une taille correspondant à hauteur de 17 pixels. Les lignes sont ensuite redressées pour obtenir des lignes parfaitement horizontales.



FigureV-1 : Exemple de texte et histogramme horizontal associé.

V-2-1- PRETRAITEMENTS

Les prétraitements sont effectués dans le but d'améliorer la qualité de l'image à Traiter. Les appareils d'acquisition tels que le scanner ou la caméra peuvent déformer l'image du texte. Certains défauts peuvent apparaître, comme par exemples l'inclinaison de l'écriture à cause du mal positionnement de la feuille dans le scanner, ou encore des taches noirs causées par des saletés ou de la poussière …

Pour prévenir de ces défauts, deux prétraitements sont envisagés : un lissage suivi d'un redressement. Un troisième est envisagé pour traiter de l'écriture de la même taille, c'est la normalisation.

1. Le lissage : consiste en deux opérations élémentaires. Le nettoyage qui consiste à détecter toutes les tâches ne faisant pas partie du texte à traiter et les éliminer de l'image du texte. La seconde opération est le bouchage qui consiste à boucher les trous internes à la forme du caractère et d'égaliser les contours de l'écriture.

   Ces deux opérations sont très délicates dans le sens où certaines tâches représentants des points diacritiques peuvent être confondus avec des bruits, des trous intra caractères peuvent être confondus avec des déformations du caractère et par conséquent bouchés par erreur.

2. Le redressement : consiste à redresser une écriture inclinée à cause d'une déformation de l'écriture lors de l'acquisition, ou de traitement d'écriture inclinée de nature. La procédure habituelle de redressement consiste à détecter l'angle d'inclinaison de l'écriture, et de corriger en une rotation isométrique d'un angle égal à – la valeur de l'angle d'inclinaison trouvé.

3. La normalisation : consiste à représenter tous les caractères à traiter dans une matrice de pixels de même taille, dans notre cas c'est une matrice de 17 lignes (définissant la hauteur maximale des caractères), le nombre de colonnes est variable selon le type de caractère.

V-2-2- SEGMENTATION DU TEXTE EN LIGNES

Les lignes du texte sont extraites en utilisant l'histogramme des projections horizontal (figure V-1). Un espace entre deux parties denses en pixels noirs, correspond à un interligne. Pour résoudre le problème de fausses lignes de texte nous avons fixé la valeur de l'espace entre deux lignes à au moins deux pixels. Un espace de moins de deux pixels est considéré comme fausse ligne de texte et est directement relié à la ligne traité. Nous n'avons pas traité le cas de lignes très rapprochées où l'interligne peut correspondre à un espace de un pixel.

V-2-3- CALCUL DE L'EPAISSEUR DU TRAIT

L'épaisseur du trait d'écriture notée (etr) est calculée en examinant une ligne du texte, colonne par colonne et en calculant à chaque fois le nombre de pixels noirs dans la colonne. La largeur de l'écriture correspond au nombre de pixels le plus fréquemment trouvé. Il correspondra par la suite à la largeur de la ligne de base.

V-2-4- DETECTION DE LA LIGNE DE BASE

La ligne de base est détectée en utilisant l'histogramme de projections horizontal. La ligne de base représente l'amplitude maximale de l'histogramme et aura pour largeur «etr» (figure V-1).

V-3- L'ALGORITHME DE SEGMENTATION

L'algorithme de segmentation (figure V-2) est un algorithme simple qui consiste à :

V-3-1- PHASE DE SEGMENTATION DES LIGNES EN PSEUDO_MOTS

Il s'agit d'analyser la ligne courante du texte pour la découper en mots ou pseudo-mots[1] (figure V-3). Pour réaliser ce découpage, la ligne de texte est examiné de haut en bas ligne par ligne de pixels pour déterminer si dans une paire de lignes, les pixels noirs sont connectés . Parmi les objets trouvés se trouveront les corps des pseudo-mots et les points diacritiques (figure V-4). Ils sont différenciés par leur taille et leurs emplacements par rapport à la ligne de base (au dessus ou en dessous de la ligne de base). Il seront associés au corps du pseudo-mot avant le post-traitement.

V-3-2- PHASE DE SEGMENTATION DES PSEUDO-MOTS EN CARACTERES

Les pseudo-mots trouvés sont examinés un à la fois. Chaque pseudo-mot est scanné de haut en bas pour détecter les point diacritiques et leur emplacement dans sa matrice de pixels, pour qu'ensuite le corps du pseudo-mot soit considéré sans diacritiques. La ligne de base du pseudo-mot courant est ensuite supprimée (figure V-5), puis un histogramme vertical est établi sur la nouvelle image. Les espaces blancs trouvés dans l'histogramme contiennent les points préliminaire de segmentation (le point de segmentation correspond généralement à la fin de l'espace blanc).

Une fois cette opération effectuée, les segments obtenus peuvent correspondre à des caractères complets, ou à des fragments de caractères. Les erreurs générées par la segmentation peuvent être laissées jusqu'à la phase de classification, ou être corrigés pendant la phase de segmentation en envisageant une étape de post-traitement. Dans notre cas nous envisageons une étape de post-traitement pour la correction des erreurs de segmentation.

[1] on entend par pseudo-mot un ensemble de caractères connectés composant un mot complet ou une partie du mot

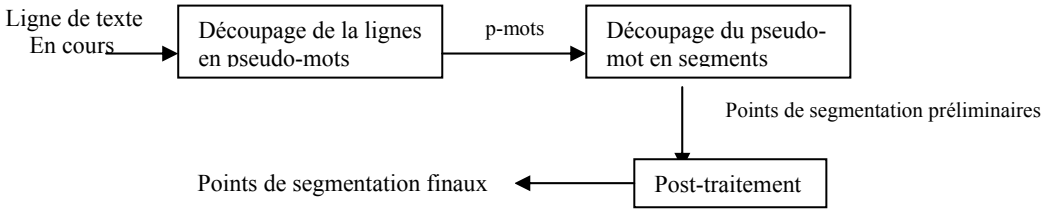


Figure V-2 : Algorithme de segmentation.

محطة كرم مدرسة براءة الوردية

Figure V-3 : mots arabes composés respectivement de droite à gauche de 1,2,3,4 et 5 pseudo-mots

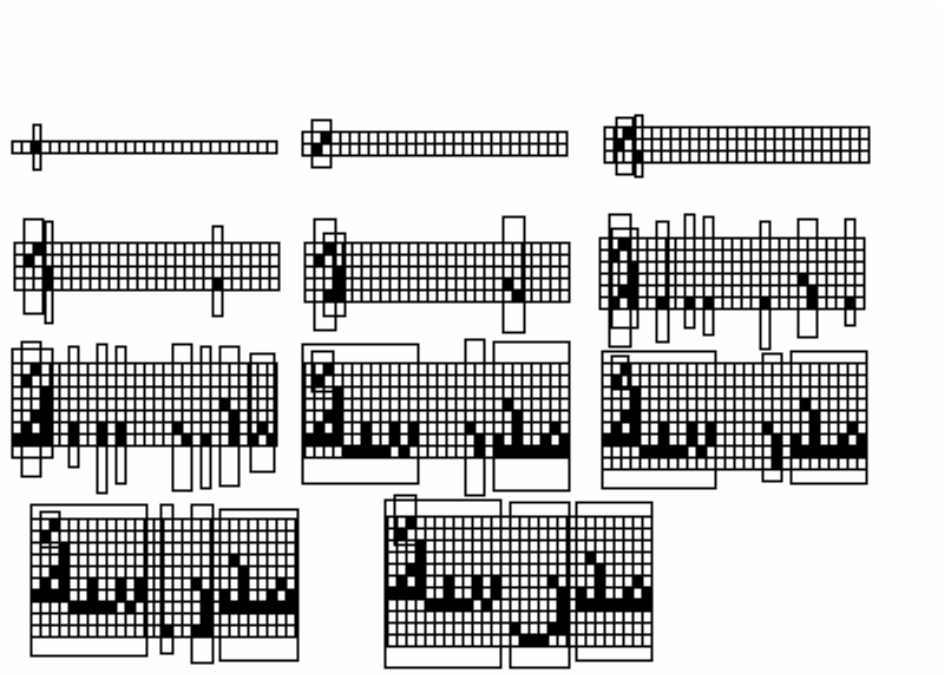


Figure V- 4 : exemple de parcours d'un mot ligne par ligne de pixels pour retrouver les pseudo-mots.

Figure V- 5 : le mot **مستطيل** avant et après suppression de la ligne de base

V-3-3- PHASE DE POST-TRAITEMENT

Les erreurs qui se produisent le plus fréquemment lors de la segmentation sont détectés, ils sont généralement de type sur-segmentation, c'est à dire qu'un caractère est découpés en deux segments ou plus.

La phase de post-traitement tente de corriger ce type d'erreurs , chacun suivant son cas. Le tableau ci dessous récapitule le type d'erreurs pouvant survenir lors de la segmentation. Les pseudo-mots sont parcourus de gauche à droite.

Avant de commencer la correction les point diacritiques sont ré-associés aux caractères, chacun dans son emplacement.

- pour les caractères de la classe C1 et C2 , ils sont détectés par leur taille et leur positions par rapport au pseudo-mot. En effet ils se situent toujours en fin du pseudo-mot (ou au début de la matrice du pseudo-mot). Ce type d'erreurs est corrigé en reliant toujours le dernier segment ayant sa taille au segment précédent.

Classe	C1	C2	C3	C4	C5
caractères de la classe	ك ب ت ث	د ذ	سـ شـ صـ ضـ	س ش	م
Exemple de segmentation d'un caractère	ـبـ	ـد	ـسـ ـصـ	ـس ـص	ـم

TableauV-1- récapitulatif des erreurs pouvant survenir lors de la segmentation

- Pour les caractères de la classe C3 les segments générés ici ont une forte ressemblance avec les caractères بـ , تـ, ثـ, نـ, يـ au début ou en milieu du pseudo-mot. La différence entre les deux cas est les points diacritiques qui apparaissent dans ces caractères mais pas dans les segments générés. Pour corriger ce type d'erreurs on teste sur la ponctuation au dessus ou en dessous du segment. Si le segment ne contient pas de ponctuation il est relié au segment suivant. Le problème qui se pose à ce moment

concerne le caractère ـسـ ,or le dernier segment du caractère ne comporte pas de point, il sera relié directement au caractère qui le précéde. Pour résoudre ce problème on teste sur le nombre de fois qu'on a effectué de liaisons dans une région égale approximativement à la largeur du caractère le plus long, et on n'en fera au maximum que deux liaisons, par ce qu'un caractère est au maximum fragmenté en trois segments.

- Pour les caractères de la classe C4. les segments obtenus ont une forte ressemblance avec le caractère ن. la différence réside dans la ponctuation. Donc pour corriger cette erreur on teste sur la ponctuation, si elle n'apparaît pas dans le segment il est relié au segment suivant.
- Pour le caractère de la classe C5 cette erreur n'apparaît que dans certaines fontes où la queue du caractère م s'écrit verticalement. Elle est détectée par sa largeur, sa position dans le pseudo-mot (or elle n'apparaît qu'à la fin) et par rapport à ligne de base. Elle est corrigée en reliant le segment au segment suivant.

V-4- STRUCTURE DU PROGRAMME

Le programme de segmentation a été réalisé par le langage MATLAB. Nous avons choisis ce langage pour les facilités qu'il offre dans la manipulation des matrices. En effet toute variable MATLAB est une matrice, et toutes les opérations sur les matrices sont prédéfinies dans le langage. La notion de procédure n'existe pas, mais on peut écrire des fonctions qui sont appelées par un programme principal.

Pour réaliser notre programme, nous avons définis 5 fonctions, chacune représentant une opération dans le procédé de segmentation ; ces fonctions sont :

- *fonction de découpage du bloc de texte en lignes (decoup)* : cette fonction a pour but de lire un fichier image bitmap (monochrome) et de le transformer en une matrice binaire, puis sa taille (ses dimensions) est enregistrée dans des variables. Ensuite la matrice est examinée (en établissant l'histogramme de projections horizontal) pour détecter les lignes de texte et les coordonnées de début et de fin de chaque ligne sont enregistrées dans un tableau appelé « lig », il contient 2 lignes représentants le numéro de la ligne dans la matrice du texte où débute la ligne et le numéro de la ligne où se termine la ligne courante. Le nombre de colonnes est variable selon le nombre de

lignes dans le bloc de texte à traiter. La première ligne est utilisée pour calculer l'épaisseur du trait d'écriture « etr ».

- ***Fonction de segmentation (segment)*** : cette fonction a pour paramètre le numéro de la ligne à traiter. Pour cette ligne un histogramme de projections vertical est établi pour détecter les espaces entre les mots, et enregistre les coordonnées de début et de fin de chaque mot dans un tableau à 2 lignes appelé « mot ». mais lorsque 2 mots (ou même les parties séparées d'un même mot) se chevauchent verticalement l'espace entre les mots n'apparaît pas dans l'histogramme. Alors une autre fonction est appelée, cette fonction est appelée :
- ***fonction de découpage du mot en composantes connexes (decxx)*** : cette fonction prend en charge le mot courant dans le tableau « mot », le copie dans une matrice à partir de la matrice de l'image et parcourt une à une les lignes de pixels du mot pour détecter les parties contigus du mot et les enregistre dans un autre tableau appelé « p_mot » (pour pseudo-mot). A ce stade tous les objet sont détectés dont les corps des pseudo-mots et les diacritiques. Les emplacements des diacritiques sont enregistrés dans un tableau appelé « diac », puis supprimés de la matrice de pseudo-mot.
- *Fonction de segmentation du pseudo-mot (segpmot)* : cette fonction utilise la matrice du pseudo-mot courant (sans diacritiques), détecte la ligne de base puis la supprime (sauvegarde une copie dans une variable). Un histogramme de projections vertical est établi et chaque espace dans l'histogramme est considéré comme une séparation entre deux caractères. Les points de segmentation sont définis et la ligne de base est replacée les emplacements des points de segmentation sont remplacés par des blancs leur largeurs est égal à un pixel sur l'épaisseur de l'écriture. Ensuite les diacritiques sont replacés pour un post-traitement.
- ***Fonction de post-traitement (posttr)*** : cette fonction a pour but de corriger les erreurs de segmentation. Chaque type d'erreur est traité séparément et suivant se classe. Le programme commence d'abord par tester si le caractère traité est un segment en fin du pseudo-mot. Si cela est vrai, alors si le segment est une « senna », il sera directement relié au segment suivant. Si le segment est un demi cercle vers le bas, alors s'il comporte un point au dessus, il est laissé tel quel sinon il est relié au segment suivant.

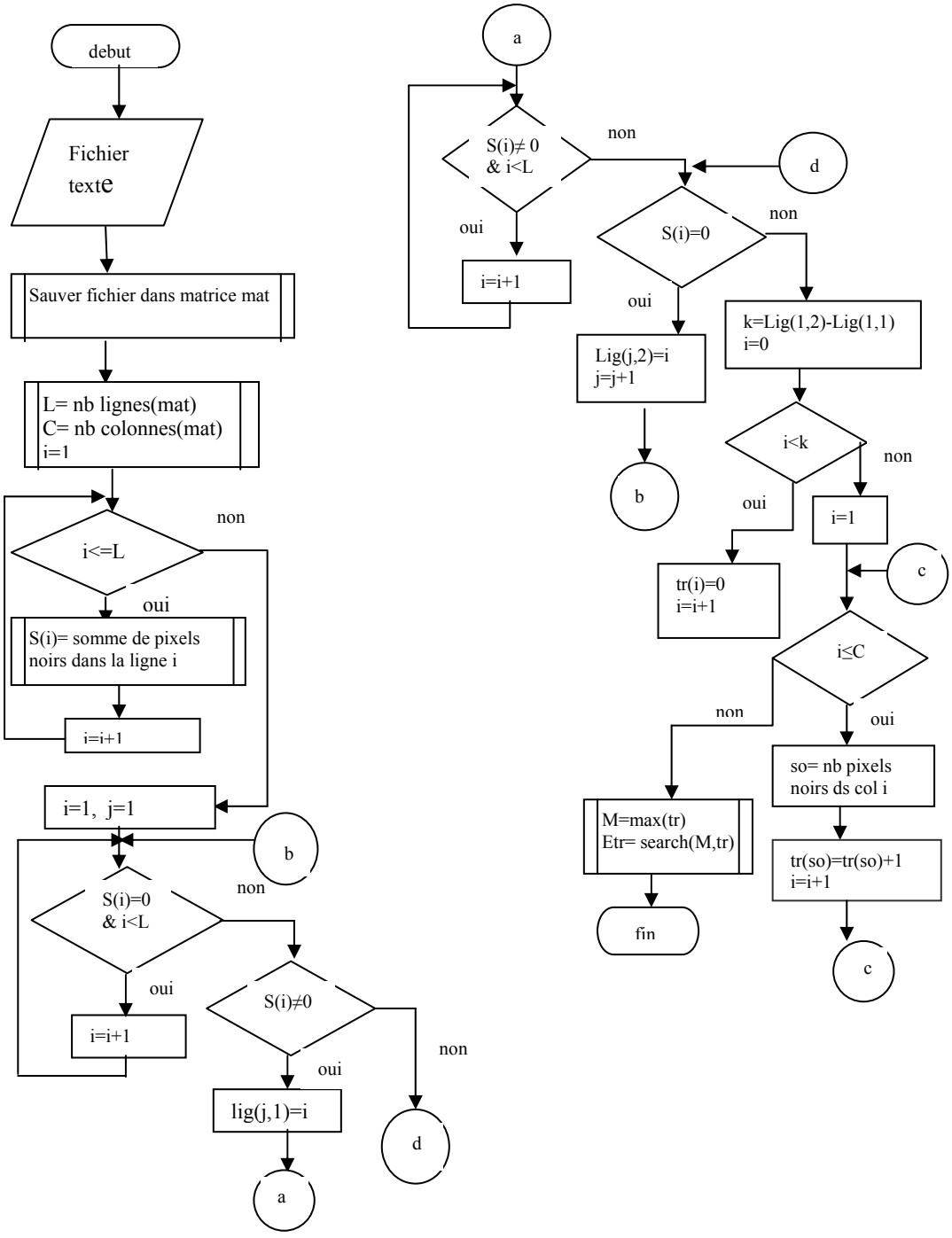
Si le segment ressemble à un « ا » mais en dessous de la ligne de base, il est aussi relié au segment suivant.

Si le segment n'est pas un segment de fin de pseudo-mot, alors on teste si le segment est une « senna », alors si elle comporte des points diacritiques au dessus ou en dessous. Si oui le segment est un caractère, il est laissé tel quel, sinon le segment est relié au segment suivant et un compteur est augmenté de 1 pour ne faire que deux liaisons au maximum. Lorsque le compteur atteint la valeur 2 le segment courant n'est pas relié au segment suivant même s'il ne comporte pas de diacritiques. Si le compteur est à 1 et le segment courant n'est pas une « senna » il est remis à 0.

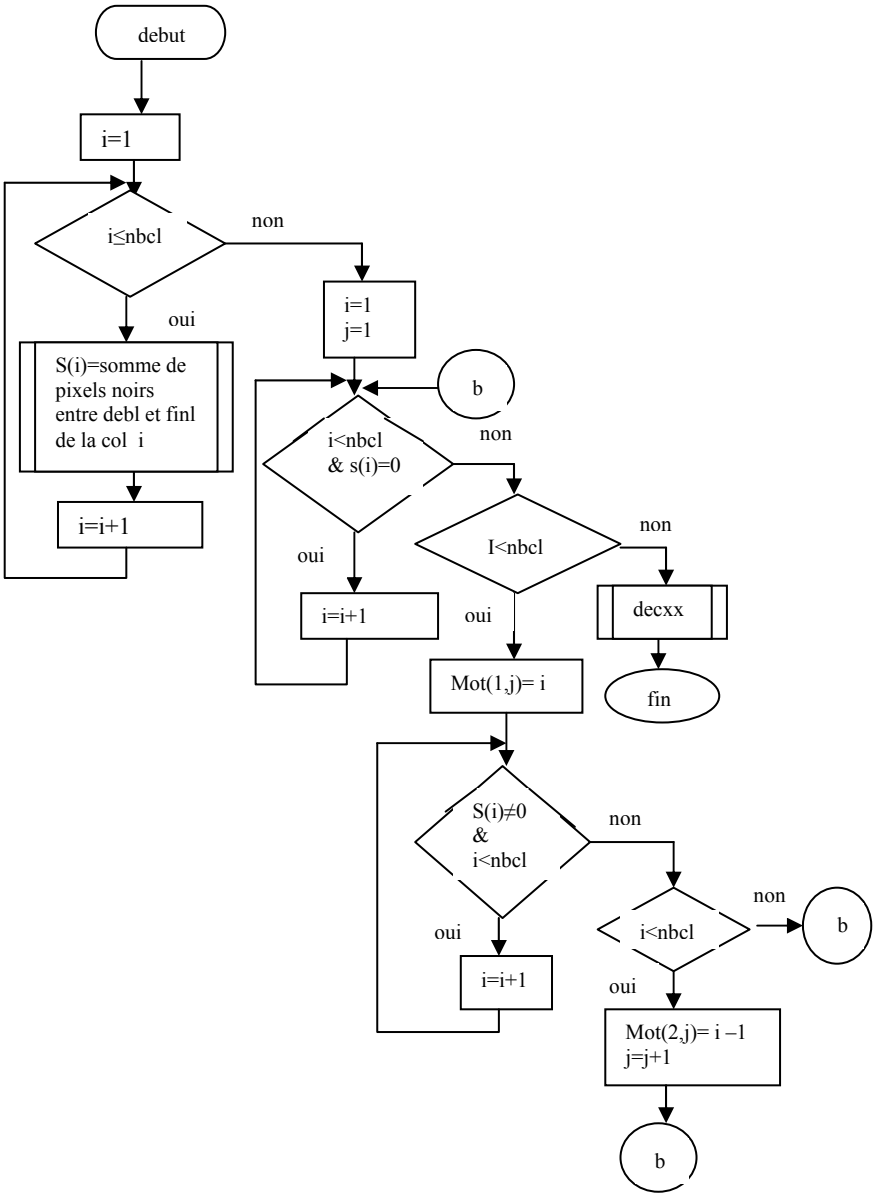
Un programme principal appelle chaque fonction à son tour. Il commence d'abord par lire le nom du fichier contenant le bloc de texte à traiter. Il le transmet à la fonction decoup et l'appelle, puis parcours le tableau de ligne « lig » et pour chaque ligne appelle la fonction segment en lui transmettant le numéro de la ligne à traiter. Une fois une ligne segmenté, il parcourt chaque entrée du tableau p-mot contenant les coordonnées des pseudo-mots et pour chaque pseudo-mot, il appelle la fonction segpmot ensuite la fonction posttr. Le résultat de ces fonctions est un pseudo-mot segmenté en caractères ce résultat est enregistré dans un fichier. Ensuite il passe au pseudo-mot suivant, jusqu'à épuisement de tous les pseudo-mots de la ligne courante, puis passe à la ligne suivante et recommence le procédé.

V-5- ORGANIGRAMME DE L'ALGORITHME DE SEGMENTATION

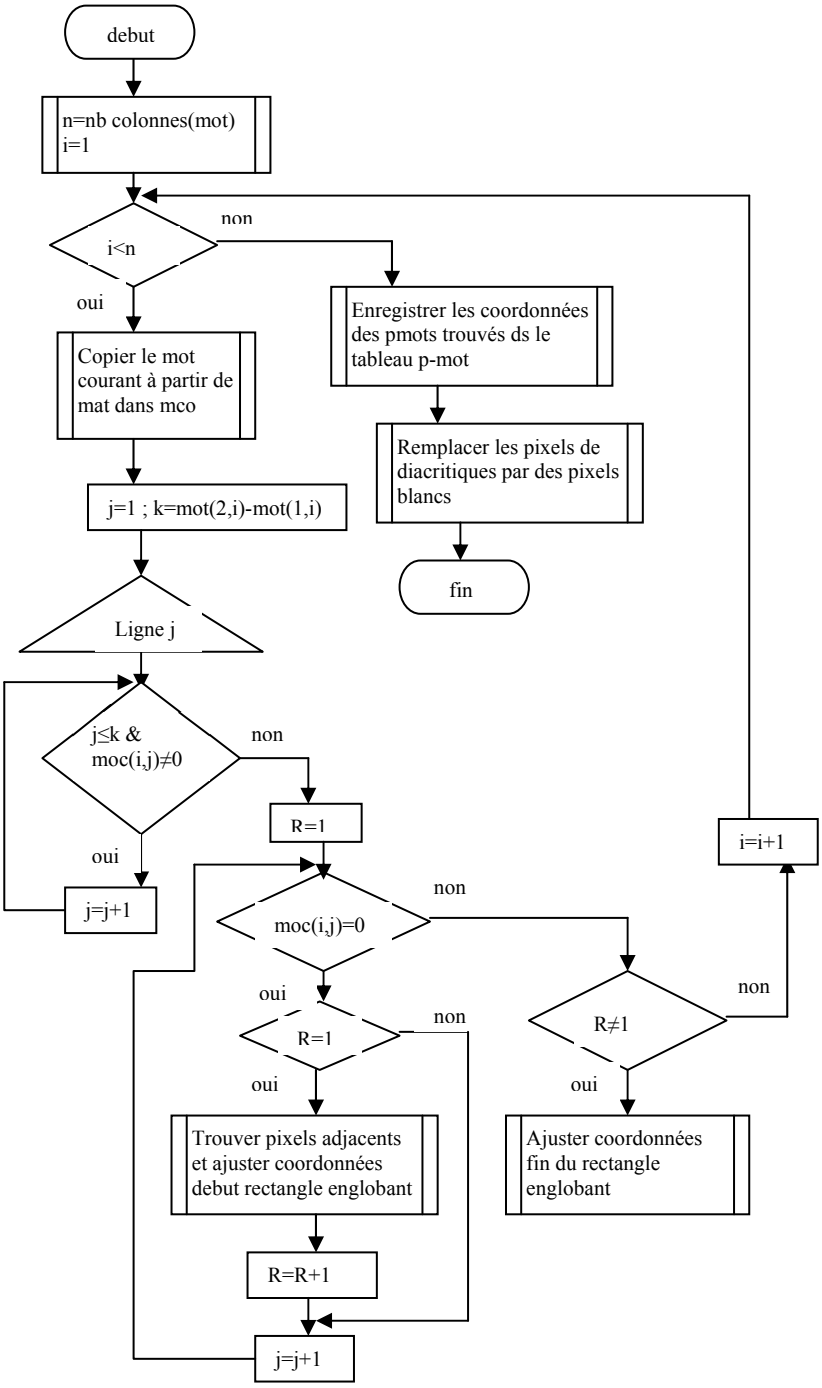
Fonction decoup



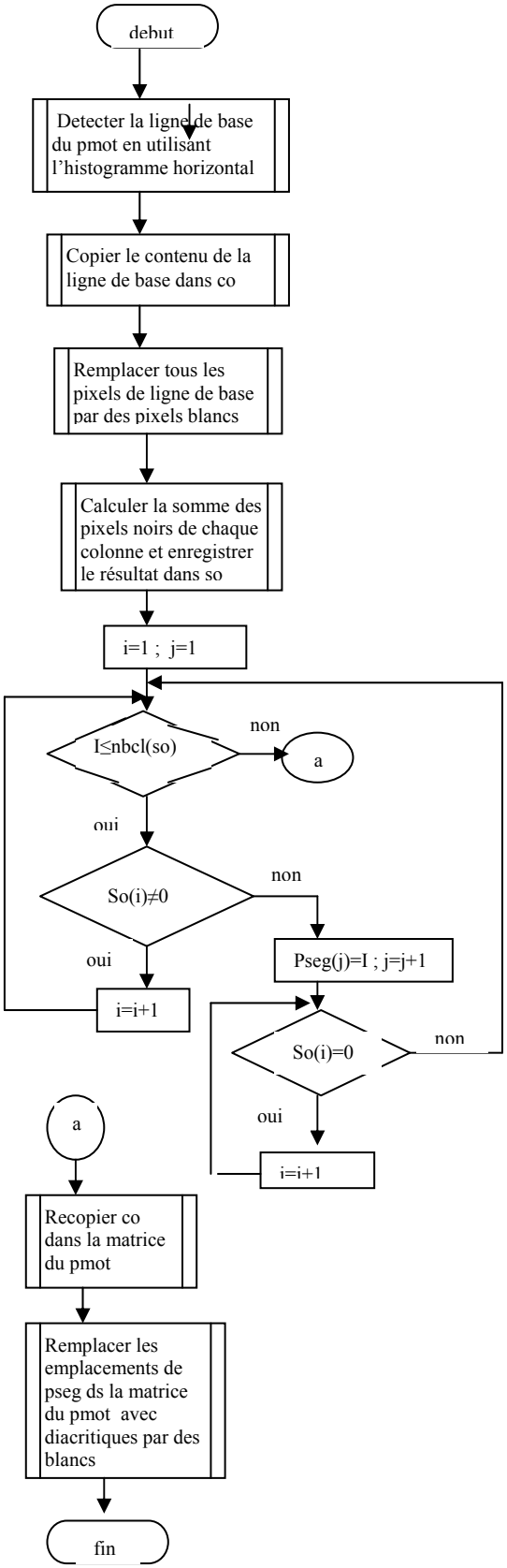
Fonction segment (debl, finl,nbcl)



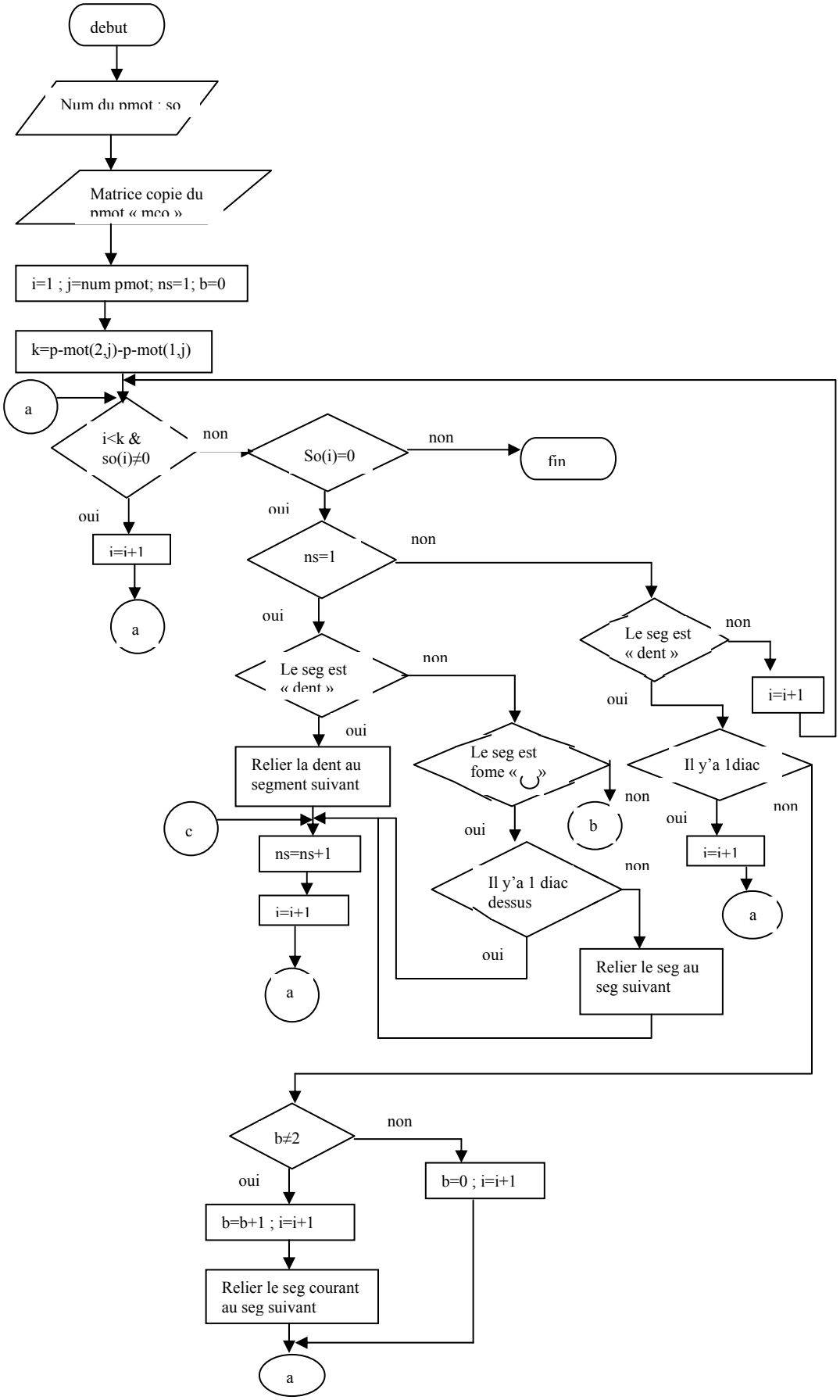
Fonction decxx (mot,debl, finl)

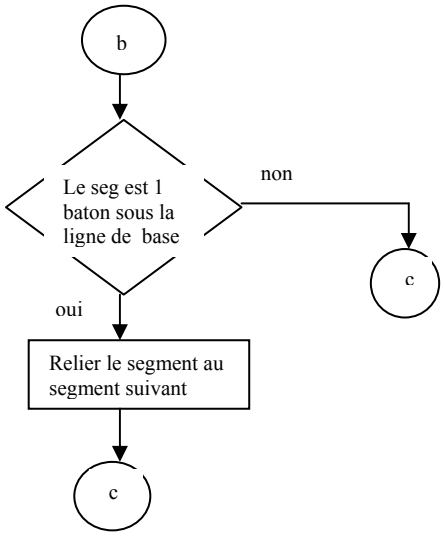


Fonction segpmot (p-mot)

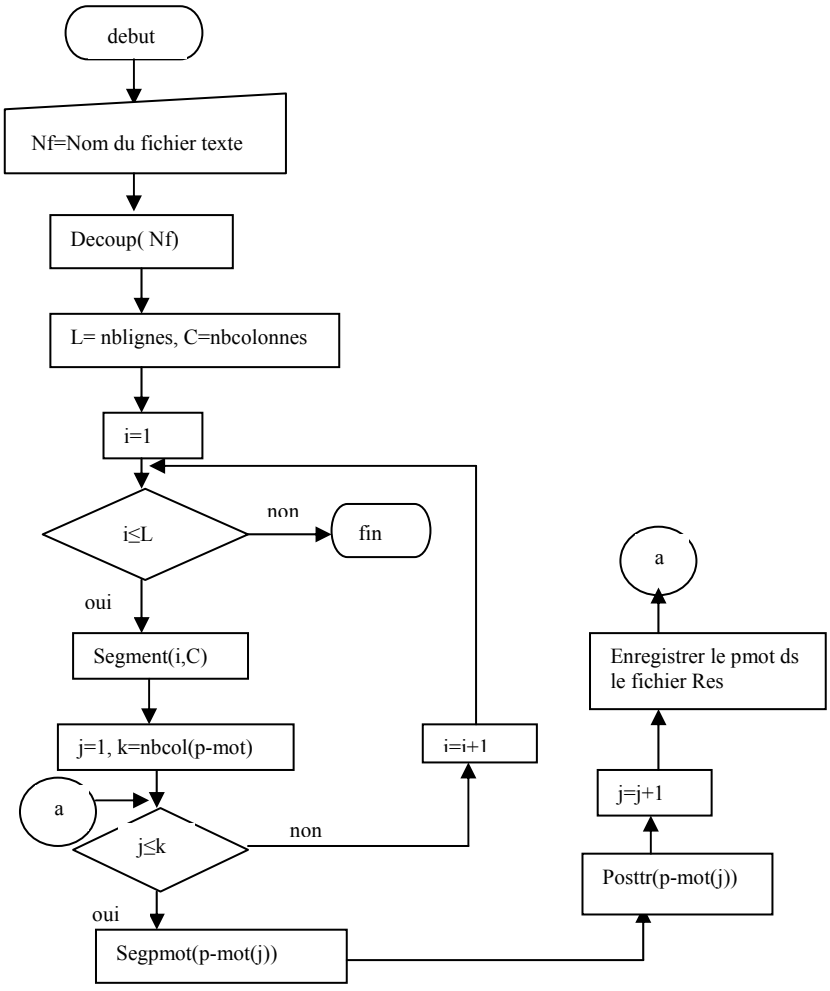


Fonction posttr(p-mot)





Programme principal



V-6- RESULTATS EXPERIMENTAUX

L'algorithme proposé a été testé sur trois pages de textes arabes écrits dans douze fontes différentes. Le texte se compose de 581 mots et de 2293 caractères. Les résultats de segmentation obtenus sont classés dans le tableau ci dessous. La majorité des cas d'échec dans la segmentation sont soit de type sous_segmentation, or dans certaines fontes où les caractères ne pouvant s'attacher aux autres caractères tels que (ر, ز, و) peuvent toucher les caractères qui les suivent (exemple :تعاون les caractères و et ن se touchent). Dans ce cas les deux derniers caractères sont considérés en tant qu'un seul segment. Un autre cas d'échec a été constaté dans les fontes "almas " et "zomorod", mais dans ce cas c'est une sur_segmentation du caractère « شـ », à cause des points diacritiques qui s'écrivent sur chaque dent ; il est confondu avec " نـ " répété trois fois. Des échantillons de résultats des expérimentations sur les différentes fontes sont détaillés dans la partie annexe.

Fonte	Nombre de mots	Nombre de caractères	Taux de bonne segmentation	Taux d'échec
Akeek	581	2293	99.52%	0.48%
Almas			99.00%	1.00%
Arabic transparent			100%	0.00%
Arabic simplified			100%	0.00%
Al-kharashi56			99.69%	0.31%
Buriyadah			99.52%	0.48%
Hillal			100%	0.00%
Koufi modern			99.65%	0.35%
Times new roman			100%	0.00%
Ramadan			99.22%	0.78%
Riadh			100%	0.00%
zomorod			98.69%	1.31%

Tableau V-2-: Résultats des expérimentations

V-7- CONCLUSION

Dans ce chapitre, un algorithme de segmentation des caractères imprimés multifontes a été proposé. Des résultats de bonne segmentation proches du 100% ont été obtenus.

Par ailleurs l'algorithme reste encore sensible à certains cas de type ligatures verticales et accolement entre caractères. Or dans certaines fontes deux caractères peuvent être ligaturés verticalement, dans certains autres deux caractères ou plus se succédant peuvent s'accoler. Pour ce dernier cas des auteurs tel que MAHMOUD dans [Mahmoud 94] résout ce problème en détectant les accolements et sépare les caractères. Des cas sur-segmentation ont aussi été constatés comme c'est le cas du caractère "شـ" dans la fonte "almas" où les points sont disposés au dessus de chaque dent qui est confondue avec trois "ـنـ" successifs.

Nous espérons enfin pouvoir dans l'avenir trouver les solutions convenables à ces problèmes dans des recherches avenirs, et pouvoir contribuer à l'avancement des recherches dans le domaine de la segmentation des caractères arabes.

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.

Dans le cas de notre étude, nous nous sommes intéressés aux méthodes de segmentation qui ont prouvé leur performance du point de vue taux de bonne reconnaissance. Cependant les problèmes majeurs influençant la recherche en AOCR sont, le manque de normalisation des calligraphies des caractères arabes, l'absence d'études approfondie relative à la classification des fontes du point de vue calligraphie et corps et l'absence d'outils tels que dictionnaires, bases de données et statistiques se rapportant à l'écriture arabe. La résolution des ces problèmes serait d'un apport considérable, tant au niveau simplification de la tache de l'AOCR, qu'aux nivaux validation et portabilité des produits réalisés.

Par ce travail nous espérons avoir couvert une grande partie concernant le domaine de recherche en segmentation des caractères arabes, et pouvoir contribuer à l'évolution des recherches, malgré que les efforts de nos jours s'intensifient dans ce domaine et chaque jour de nouveaux articles sont publiés, traitant du sujet.

Nous espérons dans l'avenir pouvoir intégrer un groupe de recherche dans ce domaine pour pouvoir mettre l'épreuve notre algorithme en l'intégrant dans un système de reconnaissance de caractères et le voir contribuer à l'avancement des recherches dans le domaine de reconnaissance des caractères arabes.

Références bibliographiques

[Abdeazim 89] H.Y. Abdelazim, M.A. Hashish : « Interactive font learning for arabic OCR». Proc. 1ST Kuwait computer conference, pp 463-486 Kuwait, 1986.

[Abdelazim 90] 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.

[Abuhaiba 93] 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.

[Aissaoui 94] 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.

[Aissaoui 96] 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.

[Aissaoui 97] 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.

[Al-Badr 94] B. Al-Badr , R.M. Haralick : « Symbol recognition without prior segmentation ». Conference SPIE-EI 1994.

[Al-Badr 95] B. Al-Badr , S.A. Mahmoud : « Survey and bibliography of Arabic optical text recognition ». Signal processing , vol. 41, pp. 49-77, 1995.

[Al-Badr 95b] 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.

[Al-Emami 90] 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.

[Alimi 94] 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.

[Alimi 95] A.M. Alimi O.A. Ghorbel : « The analysis of error in an on-line recognition system of Arabic handwritten characters». IEEE. Proc. 3rd International conference on document analysis and recognition (ICDAR'95), pp. 890-893, Montreal, Canada, 1995.

[Almuallim 87] 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.

[Alqaisy 85] E.K. Alqaisy, H.L. Naser : « Recognition of Arabic numerals using probabilistic functions ». Proc. Computer processing and transmission of Arabic language workshop, Kuwait, 1985.

[Al-Tuwaijri 95] M. Altuwaijri, M. Bayoumi : «A new thinning algorithm for Arabic characters using self-organizing neural network ». Proc. IEEE, pp. 1824-1827 , 1995.

[Al-Yousefi 88] H.S. Al-Yousefi, S.S. Udpa : « Recognition of handwritten Arabic characters ». Proc. SPIE 32nd Annual international technical symposium on optical and optoelectronic applied science and engineering, vol. 974, pp. 330-336, San Diego, CA, 1988.

[Al-Yousefi 92] 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.

[Amat 96] J.L. Amat, G. Yahiaoui : « Techniques avancées pour le traitement de l'information». Edition CEPADUES 1996.

[Ameur 93] 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.

[Ameur 94] 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.

[Ameur 97] 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.

[Amin 82] A. Amin : « Machine recognition of handwritten Arabic words by the IRAC II system ».Proc. of the 6th international joint on pattern recognition, Munich, 1982.

[Amin 89] 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.

[Amin 91] 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.

[Amin 92] A. Amin, H.B. Al-Sadoun : « A new segmentation technique of Arabic text ». IEEE. Proc. 11th IAPR, pp. 441-445, The Hague, The Netherlands, 1992.

[Amin 94] A. Amin, H.B. Al-Sadoun : « Handprinted Arabic character recognition system ». IEEE. Proc. Of the 12th International conference on pattern recognition, pp. 536-539, 1994.

[Amin 96] A. Amin, H.B. Al-Sadoun , S. Fisher : « Handprinted A rabic character recognition system using an artificial network ». Pattern recognition, vol. 29, No 4, pp. 663-675, 1996.

[Anigbogu 92] J. Anigbogu : « Reconnaissance de textes imprimés mutifontes à l'aide de modèles stochastiques et métriques ». thèse de doctorat, Université de Nancy I, 1992.

[Ayat 00] 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.

[Azmi 01] R. Azmi, E. Kabir : « A new segmentation technique for omnifont Farsi text ». Pattern Recognition letters. No 22, pp 97-104, 2001.

[Belaid 97] A.Belaid : «Analyse de documents: de l'image à la représentation par les normes de codage». Cours de l'INRIA 1997.

[Benamara 95] 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.

[Benamara 96] N. Benamara, A. Belaid : « Une méthode stochastique pour la reconnaissance de l'écriture arabe imprimée ». Forum de la recherche en informatique, Tunis, Tunisie, 1996.

[Benamara 98] N. Benamara, A. Belaid, N. Ellouze : « Modélisation pseudo bidimentionnelle pour la reconnaissance des chaînes de caractères arabes imprimées». Proc. 1er Colloque International francophone sur l'écrit et le document (CIFED'98), pp. 131-140, Quebec, Canada , 1998.

[Benamara 99] 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.

[Benamara 00] N. Benamara, A. Belaid, N. Ellouze : « Utilisation des modèles Markoviens en reconnaissance de l'écriture arabe : état de l'art ». Proc. 3ème Colloque International francophone sur l'écrit et le document (CIFED'00), 2000.

[Bennasri 99] A. Bennasri, A. Zahour, B. Taconet : « Extraction des lignes d'un texte manuscrit arabe ». Vision interface 99, Trois-Rivières, Canada, 19-21 mai 1999.

[Bouslama 97] F. Bouslama : «Arabic character recognition by fuzzy techniques» Proc. 5th European congress on intelligent techniques and soft computing, Aachen, Gerrmany, 1997.

[Bouslama 99] 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.

[Bulmenstein 98a] 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.

[Bulmenstein 98b] 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.

[Bulmenstein 99] 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.

[Burges 92] 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.

[Burrow 04] P. Burrow : « Arabic handwriting recognition ». Master of science thesis. School of Informatics, university of Edinburg, England, 2004.

[Bushofa 97] 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.

[Casey 95] R.G. Casey, E. Lecolinet : « Strategies in character segmentation : A survey ». IEEE. Proc. 3rd international conference on document Analysis and recognition (ICDAR'95), pp. 1028-1033, Montreal, Canada, 1995.

[Casey 96] 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.

[Coüasnon 96] B. Coüasnon : « Segmentation et reconnaissance de documents guidées par la connaissance a priori : application aux partitions musicales». Thèse de doctorat de l'université de Rennes I, France, 1996.

[Dunn 92] C.E. Dunn, P.S.P. Wang : « Segmentation of merged characters by

neural networks and shortest path ». pattern recognition , vol. 27, No 5, may 1994.

[El-Dabi 90] 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.

[Elgammal 01] 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.

[El-Khaly 90] 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.

[El-Ramly 89] S.H. El-Ramly, M.A. El-Hamlaway : « A new font for Arabic character simplifies recognition procedure ». Proc. 11th National computer conference, pp. 396-401, Dhahran, Saudi Arabia, 1989.

[El-Sheikh 88] T. El-Sheikh, R. Guindi : «Computer recognition of Arabic cursive scripts ». Pattern recognition, vol. 21, No 4, pp. 293-302, 1988.

[El-Sheikh 90] T. El-Sheikh, S.G. El-Taweel : « Real time Arabic handwritten character recognition ». Pattern recognition , vol. 23, No 12, pp. 97-105, 1990.

[Etemad 94] K. Etemad, D. Doermann, R. Chellappa : « Page segmentation using decision integration and wavelet packets ». International conference on pattern recognition, 1994.

[Fahmy 01] 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.

[Fakir 93] M. Fakir, C. Sodeyama : « Machine recognition of Arabic printed scripts by dynamic programming matching ». Transaction on informatics systems, vol. 76, No 2, pp. 235-242, 1993.

[Fehri 94] 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.

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

[Gillies 99] A. Gillies, E. Erlandson, J. Trenkle, S. Schlosser : « Arabic text recognition system ». Proc. Of the symposium on document image understanding technology, Annapolis, Marryland, 1999.

[Goraine 92] H.Goraine, M. Usher, S. Al-Emami : « Off-line Arabic character recognition ». IEEE. Computer society, vol. 25, No 7, pp. 71-74, 1992.

[Goraine 94] 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.

[Ha 96] T.M. Ha, G. Kaufmann, H. Bunke : « Text localization and handwriting recognition». Technical report, university of Berne, 1996.

[Hachour 04] O. Hachour : « Reconnaissance hybride des caractères arabes imprimés ». Traitement automatique de l'arabe (JEP-TALN 2004), Fès, 20 Avril 2004.

[Hadj-Hassen 91] F. Hadj-Hassen : « Printed Arabic text recognition ». Arabian Journal of Engineering science, vol. 16, No 4, 1991.

[Hamid 01] A. Hamid, R.Haraty : « A neuro-heuristic approach for segmenting handwritten Arabic text». IEEE. Proc. Of International conference on computer systems and applications (ACS), pp. 110-113, 2001.

[Haralick 94] 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.

[Hassibi 94] 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.

[Hu 93] 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.

[Jambi 93] 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.

[Kavianifar 99] M. Kavianifar, A. Amin : « Preprocessing and structural feature extraction for a multi-fonts Arabic/Persian OCR ». Proc. 5th international conference on document analysis and recognition (ICDAR'99), pp. 213-216, 1999.

[Kermi 99] S. Kermi : « Classifieur neuronal base connaissances, application à la reconnaissance des caractères arabes isolés manuscrits ». Thèse de magister, université Badji Mokhtar, Annaba, Algerie 1999.

[Khella 92] F. Khella : « Analysis of hexagonally sampled images with application to Arabic cursive text recognition ». PhD. Thesis, University of Bradford, England 1992.

[Kosawat 03] K. Kosawat : « Méthodes de segmentation et d'analyse automatique de textes Thaï ». Thèse de doctorat, université de Marne-La-Vallée, France 2003.

[Kozima 93] H. Kozima : « Text segmentation based on similarity between words ». Proc. 31st annual meeting of the association for computational linguistics, pp. 286-288, Columbus, OH, USA 1993.

[Kurdy 93] 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.

[Lallican 00] P.M. Lallican, C. Viarp-Gaudin, S. Knerr : « From off-line to on-line handwriting recognition ». Proc. 7th workshop on frontiers in handwriting recognition, pp. 303-312, Amsterdam 2000.

[Lecolinet 91] 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.

[Lecolinet 93] E.Lecolinet, O. Barett : « Cursive word recognition : Methods and strategies ». In NATO/ASI , Fundamentals in handwriting recognition, Bonas, France June 21-july 3, 1993.

[Lippmann 87] R.P. Lippmann : « An introduction to computing with neural nets ». IEEE, ASSP magazine, April 1987.

[Mahjoub 96] 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.

[Mahjoub 98] 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.

[Mahmoud 94] S.A. Mahmoud : « Arabic character recognition using Fourier descriptors and character contour encoding ». Pattern recognition, vol. 27, No 6, pp. 815-824, 1994.

[Masmoudi 02] S.T. Masmoudi, N.E. Benamara, H. Amiri : « Segmentation stage of a PHMM-based model for off-line recognition of Arabic handwritten city names». IEEE. International conference on systems, Man and cybernetics, SMS 2002, vol. 4, 6-9 October 2002.

[Mao 00] S. Mao, T. Kanungo : « Empirical performance evaluation of page segmentation algorithms ». Proc. SPIE on document recognition and retrieval, vol. 3967, pp. 303-314, 2000.

[Miled 96] 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.

[Miled 98] 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.

[Miled 01] H. Miled , N.E. Benamara : « Planar Markov modeling for Arabic writing recognition : advanced state ». Proc. 6th International conference on document analysis and recognition (ICDAR'01), 2001.

[Motawa 97] 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.

[Nawaz 03] S.N. Nawaz, M. Sarfaz, W.G. Al-Khatib : « An approach to off-line Arabic character recognition using neural networks ». IEEE. Proc. 10th international conference on Electronics, Circuits and Systems, ICECS 2003, vol. 3, pp. 1328-1331, December 2003.

[Olivier 96] 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.

[Parhami 81] B. Parhami, M. Taraghi : « Automatic recognition of printed farsi texts ». Pattern recognition, vol. 14, No 1, pp. 1-6, 1981.

[Sarfaz 03] M. Sarfaz, S.N. Nawaz, A. Al-Khuraidly :«Off-line Arabic recognition system ». IEEE Proc. Of the 2003 International conference on geometric modeling and graphics (GMAG'03), 2003.

[Sari 03] T. Sari, L. Souici, M. Sellami : « Off-line handwritten Arabic character segmentation algorithm : ACSA ». IEEE Proc. 8th international workshop on frontiers in handwriting recognition (IWFHR'02), 2002.

[Sayre 73] K.M. Sayre : « Machine recognition of handwritten words : a project report » . Pattern recognition, vol. 5, pp. 213-228, 1973.

[Seymore 99] 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.

[Souici 97] 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.

[Steinherz 99] 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.

[Tang] 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.

[Tappet 90] 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.

[Tolba 90] M.F. Tolba, F. Sheddad : « on the automatic reading of printed Arabic characters». IEEE proc. International conference on systems, Man and cybernet, pp. 496-498, Los Angeles, 1990.

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

[Trenkle 97] 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.

[Trenkle 01] 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.

[Tsang 00] I.R. Tsang : «Pattern recognition and complex systems». Thèse de doctorat, université d'Anterwerpen, 2000.

[Zahour 91] 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.

[Zahour 98] 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.