

République algérienne 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

Examineurs: Dr BILAMI Azzeddine Maître de conférence université de Batna

Dr CHIKHI Salim Maître de conférence université de Constantine

Soutenu 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

Abréviations	1	
Introduction	2	
Plan de lecture du mémoire	3	
Chapitre I	LA RECONNAISSANCE DE L'ECRITURE	4
I-1- Introduction	4	
I-2- Différents aspects de l'OCR	4	
I-2-1- Reconnaissance En-ligne et Hors-ligne	5	
I-2-2- Reconnaissance globale ou Analytique	7	
I-3- Problèmes liés à l'OCR	9	
I-4- Organisation générale d'un système de reconnaissance	10	
I-4-1- Phase d'acquisition	10	
I-4-2- Phase de prétraitements	10	
I-4-3- Phase de segmentation	13	
I-4-4- Phase d'analyse ou d'extraction des caractéristiques	13	
I-4-5- Phase de classification	15	
I-4-6- Phase de post-traitement	20	
I-5 Conclusion	20	
Chapitre II	L'OCR ET L'ARABE	22
II-1- Introduction	22	
II-2- Calligraphie et typographie arabe	22	
II-2-1- Caractéristiques de l'écriture arabe	22	
II-2-2- Alphabet arabe : données graphiques	39	
II-2-3- Conséquences techniques des caractéristiques morphologiques de l'arabe	39	
II-2-4- Notions de typographie arabe	40	
II-2-4-1- Définition de la notion de fonte	40	
II-2-4-2- Styles de calligraphies arabes	40	

II-3- Avancées en OCR arabe	42
II-3-1- Prétraitements	44
II-3-2- La segmentation	44
II-3-3- Extraction des primitives, classification	46
II-3-4- Post-traitement	47
II-4- Conclusion	48
Chapitre III ETAT DE L'ART DE LA SEGMENTATION	54
III-1- Introduction	54
III-2- Segmentation de la page	54
III-3- Segmentation d'un bloc de texte en lignes	54
III-4- Segmentation des lignes en mots	55
III-5- Segmentation des mots en caractères	55
III-5-1- Organisation des méthodes	55
III-5-2- Techniques de dissection pour segmentation	57
III-5-3- Segmentation basée reconnaissance	60
III-5-4- Stratégies mixtes (sur-segmentation)	62
III-5-5- Stratégies holistiques	62
III-6- Conclusion	63
Chapitre IV SEGMENTATION DES MOTS ARABES EN CARACTERES	64
IV-1 Introduction	64
IV-2- Etat de l'art de la segmentation des mots arabes en caractères	64
IV-2-1- Introduction	64
IV-2-2- Décomposition de la page	64
IV-2-3- Segmentation des mots	65
IV-2-3-1- Première Approche	65
IV-2-3-2- deuxième approche	65
IV-2-3-3- Troisième approche	66
IV-2-3-4- Quatrième approche	67
IV-2-3-5- Cinquième Approche	68
IV-2-4- Enumération de certains travaux de segmentation de mots arabes en caractères	68
IV-3- Etude de l'existant	76

IV-4- Choix de l'approche et des algorithmes	76
IV-5- Etude détaillée de quelques algorithmes segmentant les mots arabes imprimés en caractères	77
IV-5-1- algorithme proposé dans [Benamara 95]	77
IV-5-2- algorithme proposé dans [Gillies 97]	82
IV-5-3- algorithme proposé dans [El-Gammel 2001]	85
IV-5-4- algorithme proposé dans [Azmi 2001]	89
IV-6- Choix d'une méthode pour l'implémentation	93
IV-7- Conclusion	93
Chapitre V CONTRIBUTION A LA SEGMENTATION DES MOTS ARABES IMPRIMES EN CARACTERES	94
V-1 Introduction	94
V-2- Aquisition et pré-traitement	94
V-2-1- Pré-traitements	95
V-2-2- segmentation du texte en lignes	96
V-2-3- Calcul de l'épaisseur du trait	96
V-2-4- Détection de la ligne de base	96
V-3- L'Algorithme de segmentation	97
V-3-1- Phase de segmentation des lignes en mots	97
V-3-2- Phase de segmentation des pseudo-mots en caractères	97
V-3-3- Phase de post-traitement	99
V-4- Structure du programmes	100
V-5- Organigrammes de l'algorithme	103
V-6- Résultats expérimentaux	109
V-7- Conclusion	110
Conclusion et perspectives	111
Annexe	112
Références Bibliographiques	124