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