PAW-IFN/ENIT: une nouvelle base de pseudo-mots arabes pour une approche de reconnaissance pseudo analytique

Hanene Boukerma[1,2], Nadir Farah[2]

[1]Ecole Normale Supérieure d'Enseignement Technologique de Skikda, ENSET, 21000, Algérie.

[2]Université Badji Mokhtar, BP 2, 23200, Annaba, Algérie.

LABoratoire de Gestion Eléctronique du Document. {boukerma, farah}@labged.net

RÉSUMÉ. Le problème adressé par cet article est la construction d'une base annotée de pseudo-mots à partir d'une autre base annotée de mots arabes. Cette étape est indispensable pour notre système de reconnaissance pseudo analytique, car il n'existe pas, à notre connaissance, une base des images de pseudo-mots arabes. Dans notre démarche, la phase de segmentation en pseudo-mots constitue une étape fondamentale ; l'algorithme proposé a donné des résultats tout à fait satisfaisants avec une extension pour la résolution d'un cas particulier de la sous-segmentation. La base IFN/ENIT de noms de villes tunisiennes a été choisie comme cas d'application. La redondance importante au niveau de pseudo-mots qui constituent les mots de cette base est interprétée avec profit, ainsi le vocabulaire de 946 villes de la base IFN/ENIT a donné lieu à un vocabulaire de 759 pseudo-mots, soit une réduction de 187 entités à reconnaître. En s'appuyant sur ce nouveau vocabulaire, nous avons construit une nouvelle base d'image de pseudo-mots « PAW-IFN/ENIT » qui contient un total de 74104 images avec la position de la ligne de base de chaque pseudo-mot.

ABSTRACT. The problem addressed in this paper is the automatic generation of sub-words database from a labeled database of handwritten Arabic words. This stage is essential for our semi-global recognition system because, to the best of our knowledge, no attempts have been reported towards the development of database of handwritten Arabic sub-words. In our work, sub-words segmentation is the most important step; the proposed algorithm achieves satisfactory results with an extension for under-segmentation resolution. Experiments were performed using the IFN/ENIT database of 946 Tunisian town/village names. This database contains an important redundancy in sub-word level of their lexicon word, as well the number of unique sub-words in this word lexicon is 759 (lexicon reduction of 187 classes). Utilizing the generated sub-word lexicon, we construct a novel database of handwritten Arabic sub-words « PAW-IFN/ENIT » which contains a total of 74104 images with sub-word baseline information.

MOTS-CLÉS : Ecriture arabe manuscrite, Pseudo-mot, Base d'apprentissage, Vocabulaire, Segmentation en pseudo-mots.

KEYWORDS: Handwritten Arabic script, Sub-word, database, lexicon, sub-word segmentation.

1. Introduction

Les systèmes de reconnaissance de mots cursive sont généralement classés selon la nature de la modélisation mise en œuvre, on distingue ici deux approches principales : l'approche globale ou holistique [5] et l'approche analytique ou locale [8]. Chacune de ces deux approches a ses avantages et ses limites. Le principal avantage de la modélisation analytique est qu'elle est la seule envisageable pour une reconnaissance à vocabulaire ouvert. Cependant, la difficulté de cette approche est directement liée à la complexité de la segmentation. À titre indicatif, la segmentation d'un mot cursif latin en lettres est un problème déclaré insoluble depuis longtemps dans la communauté de la reconnaissance de l'écriture cursive latine [2]. Cependant, contrairement à l'écriture cursive latine, en arabe la présence des ligatures verticales, la diversité des formes de caractères et la variabilité de liaison entre caractères rendent cette tache de segmentation en lettres de plus en plus difficile. Par conséquent, la reconnaissance analytique de l'écriture arabe n'est pas triviale, particulièrement dans le cas du manuscrit. Quand à la modélisation globale, elle est peu discriminante pour les mots différents dont la forme est proche, ce qui limite cette approche à des applications à vocabulaire distinct et réduit [4].

Un autre niveau de modélisation spécifique à l'écriture arabe est le niveau pseudo-mot (ou PAW : Pieces of Arabic Words). Un PAW est une séquence de lettres liées, ce qui donne l'aspect de cursivité à l'écriture arabe, notons qu'un caractère isolé peut constituer un PAW à lui seul. La notion de PAW introduit une segmentation naturelle de l'écriture arabe et fait apparaître une nouvelle approche de reconnaissance appelée approche pseudo analytique. Cette approche, peu exploitée dans l'écriture arabe, offre une solution intermédiaire aux limites de deux approches analytique et globale : d'une part, elle évite le problème délicat de la segmentation en lettres lié à l'approche analytique. D'autre part, l'interprétation des PAWs plutôt que les mots conduit à une réduction de la taille et de la complexité du vocabulaire et ouvre la voie à l'exploration des vocabulaires plus étendus que ceux abordés par l'approche globale.

Dans le cadre de notre travail, nous avons choisi comme objectif principal la proposition d'un système de reconnaissance hors ligne de l'écriture arabe manuscrite capable de traiter un vocabulaire de grande dimension (environ 1000 mots). Pour reconnaitre un vocabulaire de taille pareille, nous somme particulièrement attirés par les performances prometteuses de l'approche pseudo analytique. La base IFN/ENIT v1.0p2 [7] de noms de villes tunisiennes a été utilisée pour le développement et l'évaluation de nos travaux menés sur l'écriture arabe [3]. Cette base contient en totale 26459 noms de villes dans un lexique de 946 villes, 115585 PAWs, et 212211 caractères. Une annotation (ground truth) des images de la base est faite automatiquement au niveau des lettres qui constituent les mots. L'annotation des PAWs n'est pas incluse dans les fichiers '.tru', la seule information disponible est le nombre de PAWs qui constituent le nom de ville. Il

nous est apparu dans l'étude du vocabulaire de l'IFN/ENIT, que les mots disposent d'une redondance importante au niveau des PAWs qui les constituent. De ce fait, la reconnaissance des PAWs plutôt que les mots conduit à une réduction de la taille et de la complexité du problème.

Dans cet article, nous proposons, pour implémenter l'approche de reconnaissance pseudo analytique choisie, un nouveau vocabulaire de PAWs et une nouvelle base d'images de PAWs que nous appelons « PAW-IFN/ENIT ». La partie reconnaissance pseudo analytique n'est pas décrite dans cet article. Nous nous concentrerons sur la partie segmentation en pseudo mot et construction de la base des images de pseudo mots. Le schéma synoptique de notre système est donné à la figure 1.

Le reste de cet article est organisé comme suit : dans la section 2, nous présenterons la procédure de définition du nouveau vocabulaire de PAWs. La section 3 est consacrée à l'étape de segmentation des images de mots en PAWs et à l'analyse des résultats de cette étape. La construction de la base PAW-IFN/ENIT sera décrite dans la section 4. Finalement, nos conclusions seront exposées dans la section 5.

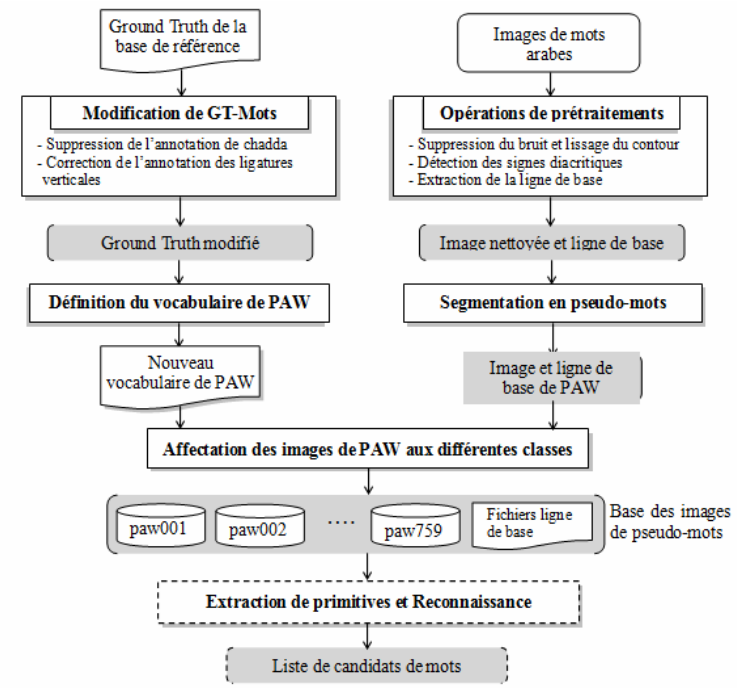


Figure 1. *Procédure de construction de la base de pseudo-mots arabes pour une reconnaissance pseudo analytique. L'étape indiquée par le rectangle pointillé n'est pas abordée dans cet article.*

2. Vocabulaire de pseudo-mots avec signes diacritiques

Comme nous l'avons évoqué précédemment, l'annotation des mots de la base IFN/ENIT est faite au niveau des lettres qui les constituent. Cette annotation est réalisée de telle sorte qu'une séquence de lettres contienne également l'information de la forme qui prend chacune des lettres au sein du mot (**B** pour Begin (début), **M** pour Middle (milieu), **A** pour Alone (isolé) et **E** pour End (finale)). Cette information a facilité la définition automatique du vocabulaire de PAWs, puisque, en arabe, un PAW peut être constitué soit par une séquence de lettres qui commence par une lettre écrite selon sa forme 'début de mot' (B) et termine par une lettre écrite selon sa forme 'fin de mot' (E), soit par une seule lettre 'isolée' (A). L'annotation des mots de la base de référence fournit également des informations concernant la présence des ligatures verticales (par exemple la ligature :لـ est annotée :'**haMlaB**' de telle sorte que la deuxième lettre de la ligature précède la première) et du signe diacritique *chadda* (**llL**), ce qui crée, dans notre vocabulaire de PAWs, des classes supplémentaires qui ne sont pas en réalité des classes de PAWs distinguées. À titre d'exemple, le tableau 1 illustre différentes images d'un même mot dont l'annotation correspondante se distingue à cause de la présence des ligatures et/ou de *chadda*.

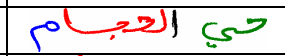
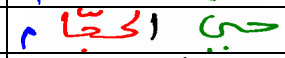
Nom de l'image	Image du nom de ville	Pr lig	Pr ch	Annotation du mot
aj18_027		0	0	haB|yaE|aaA|laB|haM|jaM|aaE|maA
af01_042		1	1	haB|yaE|aaA|haMlaB|jaMllL|aaE|maA
ai09_033		1	0	haB|yaE|aaA|haMlaB|jaM|aaE|maA
ai14_002		0	1	haB|yaE|aaA|laB|haM|jaMllL|aaE|maA

Tableau 1. *L'annotation des ligatures verticales et de chadda produisent différentes annotations d'un même mot, ce qui complique le processus de définition automatique du vocabulaire de PAWs (Pr. lig et Pr. ch: pour présence de ligature et de chadda respectivement).*

Pour éviter l'extraction de ces classes de PAWs supplémentaires, des modifications sont appliquées sur l'annotation de la base IFN/ENIT avant la définition du vocabulaire de PAWs. Ces modifications consistent à supprimer, dans l'annotation, la chaîne de caractères '**llL**' qui marque la présence de *chadda*, elles consistent également à corriger l'annotation des ligatures pour que la première lettre de la ligature (**B**) précède la deuxième (**M**). Par exemple, l'annotation de لـ : '**haMlaB**' devient '**laBhaM**'.

Il est également important de noter que, le plus souvent, l'annotation de la base décrit les mots d'une manière exacte y compris les fautes d'écriture introduites par les

scripteurs. Citons l'exemple du mot قليبـــيـــة الشرقية (postcode :8069) avec l'annotation : kaB|laM|yaM|baM |**yaM**|teE|aaA|laB|shM|raE|kaB|yaM|teE, qui a été mal écrit comme : قليبـــة الشرقية (image :*am15_040.bmp*) avec l'annotation :kaB|laM|yaM|baM|teE| aaA|laB|shM|raE|kaB|yaM|teE. Ceci produit l'apparition de classes de PAWs parasites ('**قلية**' dans l'exemple précèdent) qui seront incluses dans notre vocabulaire de PAWs même si leur fréquence d'apparition dans toute la base est parfois égale à 1 (voir section 4). Par conséquent, un même mot peut être composé de différentes séquences de classes de pseudo-mots. Cette remarque doit être prise en compte lors de la définition du vocabulaire de mots exprimés dans un alphabet de pseudo-mots.

Comme nous l'avons évoqué précédemment, le vocabulaire de la base IFN/ENIT dispose d'une redondance importante au niveau des PAWs qui constituent les mots (par exemple, le pseudo-mot *Alif* apparaît 722 fois dans les mots de ce vocabulaire).Ainsi, ce vocabulaire composé de 946 noms de villes a donné lieu à un vocabulaire de 759 PAWs, soit une réduction de 187 entités à reconnaître. Une démonstration du gain, en termes de réduction de la taille du vocabulaire à reconnaître, de la reconnaissance à base de PAWs par rapport à celle à base de mots est donnée par la figure 2.

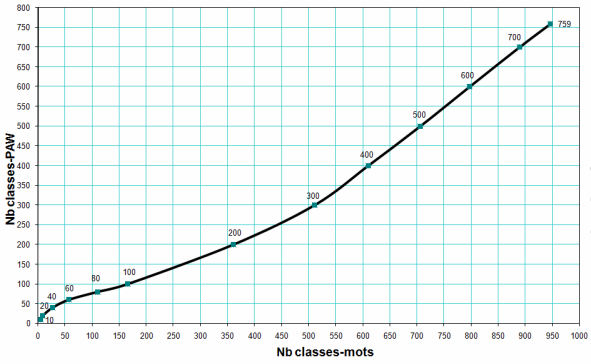


Figure 2. *Gain de la reconnaissance à base de pseudo-mots par rapport à celle à base de mots du vocabulaire de l'IFN/ENIT.*

3. Segmentation en pseudo-mots

En arabe, un PAW se compose à la fois de sa composante primaire et de ses composantes secondaires ou signes diacritiques. Après l'élimination des signes diacritiques d'un mot, les composantes connexes restantes présentent les composantes primaires des PAWs [6].La segmentation en PAWs peut être alors réalisée en commençant par extraire les composantes primaires des PAWs après une élimination totale des signes diacritiques du mot, puis on réaffecte les signes diacritiques à leurs composantes primaires correspondantes. L'algorithme utilisé pour la détection des signes diacritiques est présenté dans [3]. Le processus de réaffectation implémenté prend en compte le fait que les diacritiques se situent soit en dessous, soit au dessus de leurs composantes primaires.

Les diacritiques qui ne vérifient pas cette condition de recouvrement vertical seront affectés aux composantes primaires des PAWs selon un critère de proximité (Fig. 3).

Les signes diacritiques sont alors éliminés mais leurs images ainsi que leurs positions sont mémorisées pour exécuter le processus de réaffectation. Dans certains cas, il arrive que les signes diacritiques, généralement de taille importante par rapport à la taille du mot, ne soient pas supprimés. Ces cas nécessitent l'application d'un second filtre des signes diacritiques qui s'appuie sur la position de la ligne de base extraite. Ainsi, les petites composantes connexes qui se situent au dessus ou en dessous de la ligne de base et qui vérifient le test de la superposition verticale sont ajoutées à la liste des signes diacritiques avant de commencer la segmentation en pseudo-mots.

Figure 3. *Segmentation en PAWs correcte, les diacritiques sont réaffectés aux composantes primaires selon des critères de recouvrement vertical et de proximité.*

3.1. Résultats et analyse

Le processus de segmentation en PAWs proposé ne permet pas la résolution des problèmes de la *sur-segmentation* et de la *sous-segmentation*. Toutefois, la détection de la présence de ces deux problèmes est possible par une comparaison entre le nombre de PAWs extraits par cette méthode et le nombre de PAWs de l'annotation de l'image. Les images qui présentent ces types de problème seront exclues de la base d'apprentissage. Elles présentent 31.5%, 15.82%, 33.45% et 31.26% des sous ensembles '*A*', '*B*', '*C*' et '*D*' de la base IFN/ENIT respectivement. Ces résultats sont proches de ceux obtenus par A. AbdulKader [1]. Une analyse approfondie de ces images nous indique que :

– La sur-segmentation en PAWs est due principalement au phénomène de la levée de plume et aux artefacts d'acquisition. La sous-extraction des signes diacritiques qui permet aussi la production de la sur-segmentation (un signe diacritique est considéré comme étant un PAW) est résolue par l'application du second filtre des diacritiques.

– Le plus souvent, la sous-segmentation en PAWs est due aux PAWs qui se touchent alors qu'ils ne le devraient pas (fin de PAW relié avec le PAW suivant). Ce phénomène de connexion indésirable entre PAWs est plus ou moins facilement détectable lorsque les PAWs connectés présentent des jambes (des descendants) et se touchent en dessous de la ligne de base. Dans la section 3.2, nous présenterons une piste de réflexion pour apporter une solution au problème envisagé.

Les autres images pour lesquelles le nombre de PAWs extraits est égal à celui de l'annotation sont considérées comme des cas de segmentation valide. L'hypothèse que nous faisons ici est celle de considérer que la sur-segmentation et la sous-segmentation

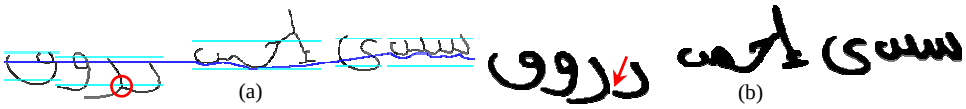
sont deux problèmes qui figurent conjointement rarement dans une même image avec la même fréquence. En parcourant alors l'image du mot de droite à gauche, les PAWs extraits sont attribués à une des 759 classes du vocabulaire selon l'annotation de l'image.

3.2. Amélioration de la segmentation en PAWs : résolution d'un cas particulier de la sous-segmentation

L'algorithme présenté ici traite le cas de succession des caractères avec jambes (des descendants) qui se touchent. Cette situation fait apparaître un point d'embranchement en dessous de la bande de base (voir Fig. 4.a). La solution proposée se base sur la détection de la ligne de base (l'algorithme présenté dans [3]) et l'extraction du squelette du mot (l'algorithme de ZHANG ET SUEN [9]), elle se fait en trois étapes :

1. Chercher les points d'embranchement qui se situent en dessous de la bande de base et qui ne correspondent pas à un point d'embranchement d'une boucle (dans la plupart des cas, l'algorithme de squelettisation génère, au niveau des boucles, deux points d'embranchement au lieu d'un seul point de croisement).
2. Détection du point de coupure pour dissocier les descendants connectés : partant du point d'embranchement détecté, un parcours du squelette est fait selon les cinq directions de Freeman F_4, F_3, F_2, F_1 et F_0 dans cet ordre, ce parcours est répété *n* fois où *n* est suffisamment grand pour s'éloigner du point d'embranchement. L'ordre du parcours choisi assure l'aboutissement à un point de coupure qui se situe dans la lettre la plus à droite de la zone de sous segmentation. Puis, à partir de la direction la plus fréquente parmi F_2, F_3 ou F_4 on applique, au niveau du point de coupure déterminé, une coupe verticale '|', diagonale '/' ou horizontale '–' respectivement, dont la hauteur égale à l'épaisseur du tracé.
3. Enfin, on applique un lissage du contour de l'image obtenue.

Les résultats préliminaires obtenus par cet algorithme sont tout à fait satisfaisants, un exemple est illustré par la figure 4.



(a) (b)

Figure 4. *Segmentation des descendants connectés.*

4. Construction de la base de pseudo-mots avec diacritiques

Chaque image binaire préalablement nettoyée de PAW est enregistrée sous un nom qui indique sa classe d'appartenance et son numéro dans cette classe. On sauvegarde également, dans la base de PAWs, des fichiers qui contiennent la position de la ligne de

base de chaque PAW. Cette information est nécessaire pour la segmentation de pseudo-mots en lettres et pour les systèmes de reconnaissance qui utilisent des primitives dépendantes de la ligne de base.

Analyse

La base de PAWs ('PAW-IFN/ENIT') est organisée en quatre sous ensembles de telle sorte que la source du contenu de chacun d'entre eux corresponde à un des quatre sous ensembles '*A*, '*B*', '*C*' ou '*D*' de la base IFN/ENIT. Nos quatre sous ensembles contiennent respectivement : 18070, 19064, 17782 et 19188 images de PAWs. A noter ici la non uniformité de la distribution de fréquence d'apparition des classes de PAWs dans la base totale[1]. À titre d'exemple, le pseudo-mot *alif*[2] apparaît 14175 fois, cependant, trois classes de PAWs ont une fréquence nulle, ces classes sont :

– Le PAW parasite فصا : '**faB**saMaaE' qui correspond normalement au pseudo-mot قصا: '**kaB**saMaaE' issu du nom de ville 'مركز **قصّا**ص' (postcode: '3013'). Dans toute la base IFN/ENIT, ce PAW a figuré une seule fois sur l'image : *ae07_033.bmp*. Cette dernière soufre d'un problème de sur-segmentation (levée de plume au niveau du caractère ص). De ce fait, ses PAWs n'ont pas été inclus dans notre base de pseudo-mots.

– Le PAW parasite : لمتر : 'laBmaM**taMraE**' qui correspond normalement au PAW لمنز : 'laBmaM**naMzaE**' issu du nom de ville '6 **المنزه**', (postcode: '2091'). Le PAW لمتر a figuré une seule fois sur l'image : *ce03_024.bmp*, qui présente également un problème de sur-segmentation en pseudo-mots (levée de plume au niveau du caractère م).

– Le PAW valide للجمي : 'laBlaMjaMmaMyaE' issu du nom de ville 'مركز **اللّجمي**', (postcode: '3067'). Dans toute la base IFN/ENIT, ce non de ville a figuré 6 fois sur les images : *bi17_007*, *ce00_008*, *ce98_012*, *ci00_009*, *cm25_023*, et *df57_037*. Malheureusement, toutes ces images présentent des problèmes de sur-segmentation due au phénomène de la levée de plume au niveau de la lettre ج, elles sont de ce fait exclues de la base de pseudo-mots.

Pour résoudre le problème de *PAW parasite : mal écrit par le scripteur, bien annoté* (les classes فصا 'faBsaMaaE' et لمتر 'laBmaMtaMraE'), l'annotation de l'IFN/ENIT doit inclure une information indiquant la présence de ce genre de problème. Dés lors, ces PAWs seront, soit exclus du vocabulaire, soit réaffectés à leurs PAWs originaux. Pour réaliser la réaffectation, une méthode à base de calcul de distance d'édition peut être utilisée afin d'établir une mesure de similarité entre la chaîne correspondant à la classe de PAW parasite et les chaînes qui correspondent aux autres classes de PAWs (par exemple, entre la chaîne '**f**aBsaMaaE :فصا ' et la chaîne '**k**aBsaMaaE :قصا ').

[1] Ce qui est tout à fait logique étant donné la non uniformité de la distribution de fréquence d'apparition des noms de villes de la base de référence IFN/ENIT.
[2] On note ici que les lettres : أ ، إ ، آ ، ا sont regroupées dans la même classe de pseudo-mot alif :paw001:aaA.

5. Conclusion

Convaincus de la supériorité de la notion de PAW en écriture arabe, l'approche pseudo analytique est devenue notre choix premier. Ce choix nous a amené au développement d'un nouveau vocabulaire et d'une nouvelle base de PAWs. Nous avons ainsi présenté, dans cet article, un processus complet de construction d'une base annotée des images de PAWs à partir d'une autre base annotée des images de mots. La disponibilité de l'annotation pour la base de référence est indispensable pour l'automatisation de la démarche proposée et permet une exploitation générale du travail présenté à n'importe quelle base de données. De ce fait, notre travail peut être considéré comme un outil qui apporte une solution au problème de manque de base de données de référence d'écriture arabe.

La présence de la redondance au niveau des pseudo-mots de mots arabes présente un phénomène naturel dans cette écriture. Pour le cas du vocabulaire de l'IFN/ENIT, un gain considérable en termes de réduction de la taille du vocabulaire à reconnaitre est obtenu : 759 classes de PAWs contre 946 classes de mots.

Notre base PAW-IFN/ENIT a été finalisée et complètement analysée, son exploitation à l'entrée d'un ou de plusieurs classifieurs présente l'objet d'une futur publication.

Bibliographie

[1] AdbulKader A., (2006). Two-tier approach for Arabic offline handwriting recognition. *In The Tenth International Workshop on Frontiers in Handwriting Recognition (IWFHR 10)*.

[2] Belaïd A. and Choisy Ch., (2006). Human reading based strategies for off-line Arabic word recognition. Summit on *Arabic and Chinese Handwriting Recognition*, *SACH'06*.

[3] Boukerma H. and Farah N., (2010). A Novel Arabic Baseline Estimation Algorithm Based on Sub-Words Treatment. *ICFHR'2010*. IEEE Computer Society, pp. 335 – 338.

[4] Farah N., Souici L. and Sellami M., (2006). Classifiers combination and syntax analysis for arabic literal amount recognition. *Engineering Application of Artificial Intelligence 19*.

[5] Madhvanath S. and Govindaraju V., (2001). The role of holistic paradigms in handwritten word recognition. *IEEE transaction on Pattern Analysis and Machine Intelligence*, Vol. 23.

[6] Mozaffari S., Faez K., Märgner V. and El-Aded H., (2007). Strategies for Large Handwitten Farsi/Arabic Lexicon Reduction. *ICDAR'07*. IEEE Computer Society.

[7] Pechwitz M., Snoussi Maddouri S., Märgner V., Ellouze N. and Amiri H., (2002). IFN/ENIT database of handwritten Arabic words. *CIFED'02*.

[8] Tay Y. H., Lallican P. M., Khalid M., Nnerr S. and Viard-Gaudin C. (2001). An analytical handwritten word recognition system with word-level discriminant training. In Proc. of *6th ICDAR*. Seattle, USA, pp. 726 – 730.

[9] Zhang T. Y., Suen C. Y., (1984). A Fast Parallel Algorithm for Thinning Digital Patterns. *Image Processing and Computer Vision*. Vol. 27, No. 3.