RECONNAISSANCE D'ECRITURE MANUSCRITE ARABE PAR RESEAUX DE NEURONES

A. Menasria et A. Bennia
Département d'Electronique, Université Mentouri, Constantine, Algérie
E-mail: az_menasria@yahoo.com
Fax: 031-819110

ABSTRACT

Nous proposons un système de reconnaissance d'écriture arabe dédié à la lecture automatique des montants littéraux de chèques libellés en langue arabe. Dans ce travail, nous présentons un nouveau jeu de primitives pour la caractérisation des mots du montant. Le système développé s'articule autour de quatre modules distincts. Un module d'acquisition, un module de pré-traitements, un module d'extraction des primitives et un module de reconnaissance (classification et décision). Ce dernier est un classificateur neuronal. Les résultats obtenus sur les bases de données utilisées sont prometteurs.

MOTS CLES

Réseaux de neurones, caractéristiques structurales, caractéristiques géométriques, classification.

I. INTRODUCTION

La nature de l'écriture arabe pose de nouveaux problèmes au système de reconnaissance automatique conventionnel (des problèmes qui ne sont pas rencontrés lors de la reconnaissance des caractères latins), comme on peut trouver des mots qui ne sont pas séparés tout le long d'une ligne (la plupart des caractères arabes d'un mot sont reliés le long d'une grande ligne), l'absence de séparation des caractères dans un mot écrit en arabe, et en plus la présence fréquente des ligatures les rend difficilement adaptée surtout aux approches de reconnaissance basées sur la segmentation des caractères [1].

Les progrès réalisés ces dernières années dans le domaine de la reconnaissance de l'écriture manuscrite arabe sont dues en grande partie à l'utilisation d'approches statistiques [1]-[5]. Parmi celles-ci, deux techniques ont plus particulièrement été mises à contribution. Il s'agit d'une part des approches basées sur les modèles de Markov cachés [6]-[9] et d'autre part des modélisations basées sur les modèles connexionnistes [2],[8],[10]-[12]. Parmi les travaux récents qui utilise ces approches en trouve des travaux qui ont proposé un état sur les principaux travaux "markoviens" développés en reconnaissance de l'arabe [13]-[14]. D'autres travaux ont proposé un système de reconnaissance de mots manuscrit arabes basé sur des modèles connexionnistes [8], [12]-[15].

Dans ce travail, nous nous intéressons à l'élaboration d'un système de reconnaissance de l'écriture à vocabulaire limité car l'application visée derrière cette architecture est la classification des montants littéraux des chèques postaux. Pour cela, on a adopté l'approche globale qui se base sur la reconnaissance de tout le mot en le considérant comme une seule entité sans être obligé de passer par la phase de segmentation en caractères. Cette dernière qui est un vrai problème complexe à résoudre car elle provoque une multitude de difficultés, et de confusion de choix des points de segmentation.

Dans ce que suit nous détaillons les différentes phases adoptées pour la réalisation de ce système. Les étapes d'un système de reconnaissance d'écriture comportent les étapes suivantes : acquisition, pré-traitement, extraction des caractéristiques. Une base d'apprentissage est élaborée et un

classificateur est développé pour prendre une décision.

Dans ce qui suit, nous allons décrire brièvement l'opération d'acquisition et les opérations de pré-traitements Dans la section 3, nous présenterons la procédure adoptée pour l'extraction des primitives qui sera plus au moins détaillée. Dans la section 4, nous abordons le classificateur à base de réseaux de neurones et nous conclurons notre article en présentant les résultats obtenus dans notre travail.

II. ACQUISITION ET PRE-TRAITEMENT

Un montant de chèque écrit en arabe ne diffère des autres que par le vocabulaire utilisé. Quelque soit ce montant, il est limité à un vocabulaire de 48 mots, voir Tableau 1.

Les opérations de pré-traitement préparent le fichier de l'image pour les étapes suivantes du processus de reconnaissance : binarisation, lissage et extraction des contours extérieurs [16]-[18].

Mot	code	Mot	code
عشر	1	مليونا	25
عشرة	2	مليونان	26
احد	3	مليارا	27
آلاف	4	مليران	28
ألف	5	مليار	29
ألفا	6	عشرون	30
ألفان	7	سبعة	31
أربعة	8	سبعمائة	32
أربعمائة	9	سبعون	33
أربعون	10	سنتيم	34
دنانير	11	سنتيمات	35
دينار	12	ستمائة	36
اثنا	13	ستون	37
اثنان	14	ستة	38
جزائري	15	ثلاثة	39
خمسة	16	ثلاثمائة	40
خمسمائة	17	ثلاثون	41
خمسون	18	ثمانية	42
مائتا	19	ثمانمائة	43
مائتان	20	ثمانون	44
مائة	21	تسعة	45
ملايين	22	تسعمائة	46
ملايير	23	تسعون	47
مليون	24	و	48

Tableau.1: Vocabulaires des montants littéraux de chèques.

III. DETECTION ET CHOIX DES PRIMITIVES

Un des problèmes fondamentaux de la reconnaissance de l'écriture est de déterminer quelles caractéristiques employer pour avoir le bon résultat de la classification, surtout si l'approche appliquée de reconnaissance est globale comme dans notre cas. Les caractéristiques que nous avons

choisies dans notre système sont les suivantes:

A .Caractéristiques structurelles

Nous avons choisi un ensemble de caractéristiques structurelles pour décrire les différents mots de notre lexique, qui sont

- Le nombre de composantes connexes.
- Le nombre de boucles.
- Le nombre de hampes.
- Le nombre de jambages.
- Le nombre de points diacritiques en haut.
- Le nombre de points diacritiques en bas.

1. Le nombre des composantes connexes

La limitation du lexique est un avantage dans la reconnaissance des montants des chèques ; chacun des 48 mots de notre lexique est composé de : une, deux, trois, ou quatre composantes connexes.

Nous avons remarqué que la majorité des scripteurs respectent en général deux règles dans l'écriture des mots arabes ; qui sont :

- Les présences des signes diacritiques.
- La séparation des composantes connexes (figure 1).

Ces avantages ont été le soutien de la décomposition du vocabulaire des 48 mots en quatre sous ensembles selon le nombre de composantes connexes. Ces quatre sous-ensembles sont :

- Mots composées de quatre composantes connexes : جزائري , مليارات ,أربعمائة , أربعون .
- Mots composées de trois composantes connexes: ثلاثمائة , مائتان , ثمانون , ثلاثون ,عشرون ,أربعة , اثنان , دناتير , دينار , مليارا , ملونان , آلاف , ألفان , ثمانمائة .
- Mots composées de deux composantes connexes :**ستون** , ستمائة , تسعون , تسعمائة , ثمانية , ثلاثة , ألفا , ألف , احد , عشرة , خمسمائة , خمسون , مائتا , مائة , ملايين , مليار , مليون , اثنا , سبعمائة , سبعون , سنتيمات .
- Mots composées de quatre composantes connexes : خمسة , ستة , سبعة , عشر , سنتيم , و .

Par la suite, la détection du nombre de composantes connexes nous sert comme clé de référence au sous vocabulaire concerné.

2. Le nombre de boucles pour chaque composant connexe

Notons que chaque composante connexe contient une ou deux boucles au maximum.

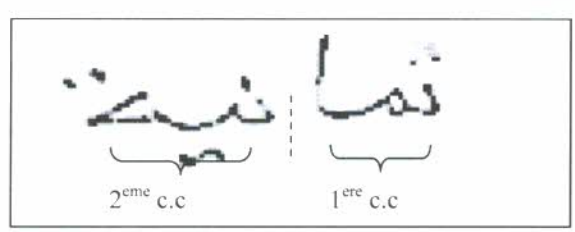


Figure 1 : Mot 'Thamania' contient deux composantes connexes.

Nous proposons une nouvelle méthode d'extraction des boucles basées sur la procédure d'étiquetage

(connexité figure 2) des zones blanches de l'image.

Figure 2: Extraction des boucles par la méthode d'étiquetage.

3. *Le nombre de hampes et jambages pour chaque composant connexe*

La méthode que nous avons utilisée pour détecter les hampes, les jambages et la médiane est basée sur l'analyse de l'histogramme horizontal. Ce dernier permet de mettre aisément en évidence la zone médiane du mot car la contribution des lettres sans hampe ni jambage y est déterminante. On cherche dans un premier temps la ligne de base où la ligne de l'écriture. On calcule la somme $S(i)$ suivante :

$$S(i) = \sum h(j) \tag{1}$$

L'indice i correspondant à la ligne de base où la somme $S(i)$ est maximum, est noté M. Dans la plupart des cas, cette ligne d'indice i se trouve à l'intérieur de la zone médiane du mot, même si elle est parfois plus près d'un bord de la zone que de l'autre.

Dans un deuxième temps, on recherche dans la partie supérieure à la ligne d'indice M ainsi que dans la partie inférieure, les indices des minima de l'histogramme respectivement m_h et m_b, figure 3. Dans le cas idéal, ces deux minima délimitent la zone médiane. Mais, en pratique, l'histogramme est souvent étalé et dessine une lente décroissance autour de la zone médiane. On recherche donc plutôt les maxima de la dérivée de la fonction h afin d'obtenir l'indice pour lequel la variation est la plus importante.

La ligne de séparation entre les hampes et la zone médiane est obtenue au maximum de la dérivée de la fonction h compris entre m_h et M, tandis que la ligne de séparation entre la zone médiane et les jambages est obtenue au maximum de la valeur absolue de la dérivée de h compris entre M et m_b [19].

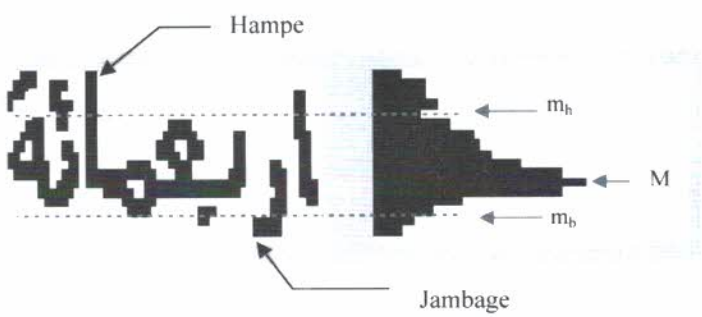


Figure 3: Localisation des hampes, jambages et médiane dans le mot 'Arbaoumaa'.

4. Le nombre de points diacritiques

L'écriture arabe est riche en points diacritiques. L'écriture de ces points est, généralement, respectée par les scripteurs car ces derniers permettent de distinguer entre les caractères ayant le même corps principal. Ces points peuvent être un seul point, deux, ou trois, figure4. De plus ces points sont en dehors de la zone d'information principale (zone médiane) et ils sont simples à détecter. En contre partie, les points simples sont de taille assez faible, ce qui les rend sensibles aux bruits d'acquisition. Les points multiples par contre, sont des formes complexes, car ils présentent un regroupement de points d'un même caractère. Nous détectons les points par leurs dimensions dans le mot.

A titre d'exemple, le nombre de points diacritiques de chaque composant connexe pour la figure 4 est illustré dans le tableau 2.

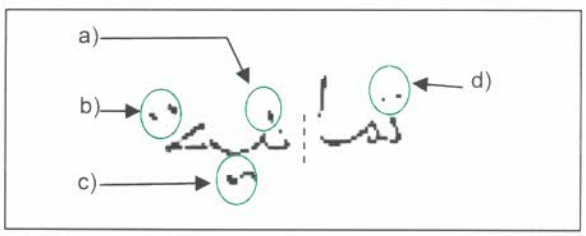


Figure 5: Détection de diacritique -a) un point simple -b) deux points simples -c) deux points liés -d) Trois points liés.

	Code	c.c1	c.c2	c.c3	c.c4
1 point bas	Pb	0	0	0	0
1 point haut	Ph	0	1	0	0
2 points bas	Dpb	0	1	0	0
2 points hauts	Dph	0	1	0	0
3 points hauts	Tph	1	0	0	0
2x2 points hauts	2Dph	0	0	0	0
2x2 points bas	2Dpb	0	0	0	0

Tableau 2: Représentation des points diacritiques de chaque composant connexe de la figure 5.

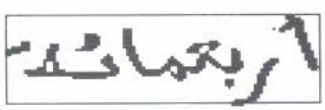
B .Caractéristiques géométriques

1-Histogrammes de directions(HD)

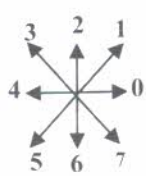
Nous avons décidé de mettre en oeuvre la technique des histogrammes de directions afin de coder le contour extérieur de la composante connexe. L'extraction de ces caractéristiques est donc réalisée sur l'image des contours. Cette dernière est disponible dans notre système au cours de la phase de pré-traitement.

Un suivi de contour est alors réalisé de manière à obtenir les directions entre pixels successifs. Ces dernières sont codées par l'intermédiaire du code de Freeman [16]. Les occurrences de ces directions sont alors dénombrées pour chaque zone individuellement. Nous présentons dans figure 6, un exemple de cette technique d'extraction. Pour ces caractéristiques, la normalisation s'effectue en divisant chaque compteur de direction d'une zone par le nombre de pixels de contours présents dans cette zone. Le vecteur caractéristique associé à cet espace de représentation contient alors huit composantes par zone [18], [20].
Dans notre application nous avons choisi le vecteur caractéristique des histogrammes de direction. L'extraction de ces caractéristiques est réalisée sur l'image des contours de chaque composante connexe.

a)



b)



c)

	cc1	cc2	cc3	cc4
0	0.0536	0.0139	0.1595	0.2129
1	0.0357	0.3627	0.0925	0.1207
2	0.2702	0.1291	0.1457	0.0999
3	0.1869	0.0139	0.1031	0.0354
4	0.0607	0.0556	0.1141	0.2041
5	0.0125	0.2116	0.0911	0.0242
6	0.2500	0.1838	0.1647	0.1849
7	0.1304	0.0294	0.1293	0.1180

Figure 6: Caractéristique de l'histogramme de direction.
a) Image originale.
b)Le contour extérieur de chaque composant connexe.
c) Matrice de direction normalisée.

2- Les descripteurs de Fourier (DF)

Le choix des descripteurs de Fourier dans notre système de reconnaissance est dû à leurs propriétés d'invariabilité et de stabilité [18]. En effet, à partir de la fonction du contour d'une image, on peut générer un ensemble complet de nombres complexes dits descripteurs de Fourier ou harmoniques. Un point s'y déplaçant génère un signal monodimensionnel complexe $z(n)$ où n représente l'abscisse curviligne du contour [16] *Figure 7*.

$$z(n) = x(n) + j \cdot y(n) \qquad (2)$$

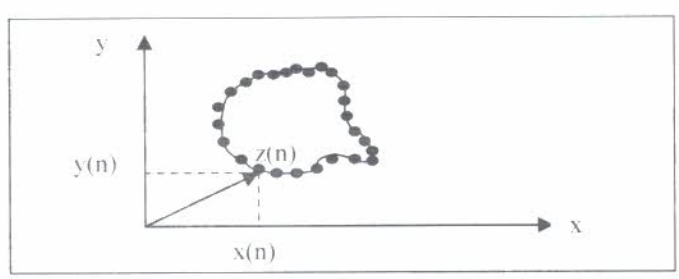


Figure 7 : Représentation complexe d'un contour.

Où x représente la ligne (abscisse), y la colonne (ordonnée) pour n variant de 0 à $N-1$.
Compte tenu de la nature périodique (de période N) de cette suite, on peut la représenter en utilisant la transformation de Fourier discrète (DFT) donnée par l'équation suivante :

$$Z(k)=\sum_{n=0}^{N-1} z(n)\cdot e^{-2j\pi kn/N} \quad \text{pour} \quad 0\le k\le N-1. \tag{3}$$

Les coefficients $Z(k)$ $(k=0,1,....,N-1)$ désignent les descripteurs de Fourier du contour. Après application de la transformée de Fourier rapide (TFR) le contour est décrit par les coefficients (descripteurs) de Fourier que l'on visualise (en module) sous forme de *raies*.
La reconstruction du contour peut être réalisée par l'application de la transformée de Fourier rapide inverse en utilisant l'équation (4) aux coefficients $Z(k)$.

$$z(n)=(1/N)\cdot\sum_{k=0}^{N-1} Z(k)\cdot e^{2j\pi kn/N} \quad \text{pour} \quad 0\le n\le N-1. \tag{4}$$

On peut également effectuer une opération de filtrage, par exemple en supprimant certains coefficients. Après transformée inverse, on obtient un contour fermé qui approxime plus ou moins bien le contour initial, *Figure 8.*
Le vecteur caractéristique est composé de 16 coefficients, ce dernier est suffisant pour la reconstruction du contour du caractère comme montré par *Figure 8.*

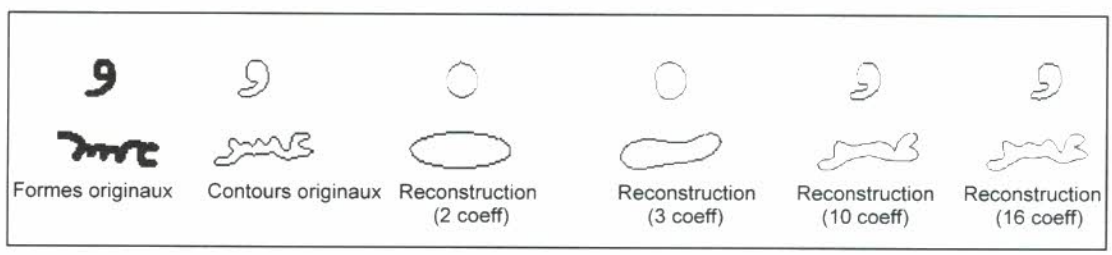


Figure 8 : Reconstruction de la forme originale du contour extérieur à partire de 16 coefficients.

L'algorithme utilisé pour le calcul des seize coefficients est :

POUR chaque caractère FAIRE
- Lecture du modèle de caractère à partir du fichier.
- Extraction des N points ordonnés de contour extérieur.
- Représentation complexe $Z(n)$ de ce contour (*équation 2).*
- Calcul de la TFR du signal complexe (*équation 3).*
- Sélection des 16 premières harmoniques.

FIN FAIRE

C. Constitution du vecteur de caractéristiques

Pour chaque trait de composante connexe le nombre et un vecteur de 37 primitives sont obtenus (13 caractéristiques structurelles et 24 caractéristiques géométriques). Ce vecteur gardera cette

cardinalité pour toutes les autres composantes connexes. Le format de vecteur caractéristique est illustré dans la figure 9.

	Caractéristique	Code
Structurelles	Une boucle	u_b
	Deux boucles	D_b
	Hampe	H
	Jambage	J
	Hampe et jambage	H_J
	Médiane	M
	Un point haut	Ph
	Un point bas	Pb
	Deux points haut	Dph
	Deux points bas	Dpb
	Trois points haut	Tph
	2 fois deux point haut	2Dph
	2 fois deux points bas	2Dpb
Géométriques	- Histogrammes de direction : 8 caractéristiques -Coefficients de Fourier : 16 coefficients	

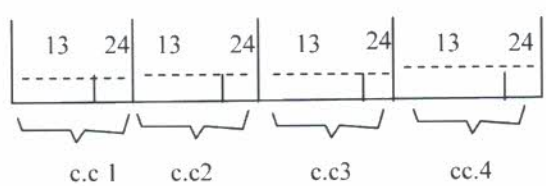


Figure 9: Présentation du vecteur des primitives.

V. CLASSIFICATION

Le classificateur que nous avons utilisé est un perceptron multicouches rétro-propagation de gradient d'erreur à une couche cachée [21]-[22]. Les 144 primitives qui représentent les 37 caractéristiques possibles pour chacune des 4 composantes connexes développées précédemment sont les entrées du réseau. La couche cachée est composée de 19 neurones. Les classes à discriminer sont les 48 qui représentent les mots du vocabulaire des montants littéraux, d'où le choix de 48 neurones pour la couche de sortie. La fonction d'activation des neurones est la fonction sigmoïde unipolaire [23].

A. Apprentissage

La base d'apprentissage représentant 48 mots différents et pour chaque mot nous avons pris 5 images de prises de vues différentes. Donc, pour l'apprentissage nous avons utilisé une base

constituée de 240 images. Après avoir effectué plusieurs tests pour fixer les caractéristiques adéquates du réseau, nous avons procédé à la phase d'apprentissage en utilisant un nombre d'itérations considérable pour minimiser l'erreur, ce qui nous a mené au graphe de la figure 10 où l'erreur a atteint la valeur de 0.0101.

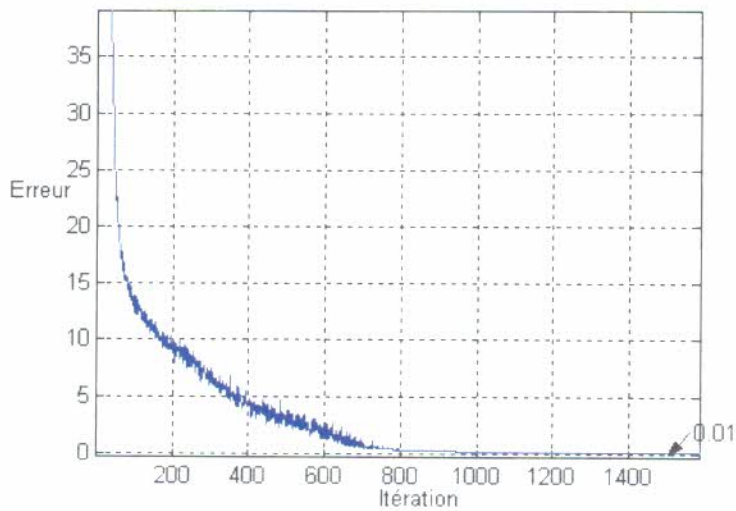


Figure 10: Evolution de l'erreur en fonction du nombre d'itérations.

B. Reconnaissance

Pour tester la fiabilité de notre système, nous avons effectué des tests sur la base d'apprentissage elle-même, nous avons obtenu des taux de reconnaissance de 100% pour un seuil de rejet égal à 0.001. Ensuite, nous avons effectué les tests de reconnaissance sur une base de test de 144 mots. Les résultats sont illustrés aussi dans le tableau 3.

	Taux de Reconnaissance (en %)	Taux de Rejet (en %)
Base d'apprentissage	100 %	0 %
Base de test	82 %	18 %

Tableau 3 : Taux de reconnaissance sur les ensembles d'apprentissage et de test.

VI. Conclusion

Dans cet article, nous avons proposé un prototype d'un système de lecture automatique de montants littéraux manuscrits de chèques libellés en langue arabe. Le travail s'articule essentiellement autour de deux parties. Une partie d'extraction des primitives les plus pertinentes et une partie de classification.

Nous avons présenté dans la première partie un nouveau jeu de primitives basé sur les

caractéristiques structurelles et géométriques. Dans la deuxième partie, nous avons choisi un réseau de neurones multicouches entraînés par l'algorithme de retro-propagation du gradient qui est un modèle connexionniste. Cependant, les résultats obtenus et malgré les efforts et les travaux réalisés dans le domaine de la reconnaissance d'écriture manuscrite arabe aucun système n'est jugé fiable à 100%. Aussi, c'est un domaine qui reste ouvert aux propositions et aux expérimentations.

VII. REFERENCES

[1] N. Benamara and N. Ellouze, *A Robust approach for Arabic printed character segmentation*, Proc. 3rd International Conference on Document Analysis and Recognition (ICDAR'95), pp.865-868, Montreal, Canada, 1995.

[2] C. Olivier, H. Miled, K. Romeo and Y. Lecourtier, "Segmentation and coding of arabic handwritten words", Proc. 13th International Conference on Pattern Recognition (ICPR'96), pp. 264-268, Vienne, Autriche, 1996.

[3] S. Snoussi-Maddouri, H. Amiri, A. Belaid and C. Choisy.*Combination of local and global vision modelling for Arabic handwritten word recognition*, International Workshop Frontier in Handwriting (IWFHR'02),Canada.2002.

[4] O. Hachour. *Reconnaissaance hybride des caractères Arabes imprimé*, JEP-TALN 2004, Traitement Automatique de l'Arabe, Frés, 2004.

[5] W. Kammoun and A. Ennaji. *Reconnaissance de textes Arabes à vocabulaire ouvert*, Laboratoire Perception, Système, Information (PSI) FRE-CNRS 2645, Université de Rouen 76821 Mont Saint Aignan Cedex, France.

[6] Y. Al-Ohalia, M. Cheriet and B. C. Suena, "Databases for recognition of handwritten Arabic cheque", *Pattern Recognition*, 36 (2003), pp. 111-121.

[7] D. .S. Alceu. J. R. Britto, R. Sabourin, F. Bortolozzi and C. Y. Suen. *The recognition of handwritten numeral strings using a two-stage HMM-based method*, IIJDAR (2003) 5: 102-117.

[8] S. Snoussi-Maddouri, H. Amiri, A. Belaid et C. Choisy. *Combination of local and global vision modelling for Arabic handwritten words recognition*, in : Eighth International Workshop on Frontiers in Handwriting Recognition - IWFHR'02, Ontario, Canada, 2002.

[9] A. S. Britto. R. Sabourin. F. Bortolozzi and C. Y Suen,. *A two-stage HMM-based systems for recognizing handwritten numeral strings*, Proceedings of the International Conference on Document Analysis and Recognition (ICDAR'01), pp. 396-400, Seattle, USA, September 2001.

[10] J. H. Kim, K. K. Kim, C. P. Nadal, C. Suen.*A methodology of combining HMM and MLP classifiers for cursive word recognition*, In proceedings of ICPR'2000, vol. 2, pp. 319-322, Barcelona-Spain,2000.

[11] A. Koerich, Y. Leydier, R. Sabourin, C.Y. Suen,. *Système hybride de reconnaissance de mots manuscrits sur un grand vocabulaire utilisant des réseaux neuronaux et des modèles de Markov Cachés*, CIFED. Volume X- n° X/2002.

[12] M. Blumenstein and B. Verma. *A Segmentation Algorithm used in Conjunction with Artificial Neural Networks for the Recognition of Real-World Postal Addresses*, ICCIMA'97, Australia.

[13] N. Benamara, A. Belaïd and N. Ellouze.*Utilisation des modèles markoviens en reconnaissance de l'écriture arabe : Etat de l'art*, Colloque International Francophone sur l'Ecrit et le Document (CIFED'00), Lyon, juillet 2000.

[14] H. Miled, M. Cheriet, C. Olivier and Y. Lecourtier. *Modélisation markovienne de l'écriture arabe manuscrite: une approche analytique*, Proc. Colloque international francophone sur l'écrit et le document (CIFED'98), pp. 50-59,Québec, Canada, 1998.

[15] S. Snoussi-Maddouri, A. Belaid, C. Choisy and H. Amiri,. *Modèle perceptif neuronal à vision globale-locale pour la reconnaissance de mots manuscrits arabes*, in : Conference Internationale Francophone sur l'Ecrit et le Document - CIFED 2002,
[16] A. Belaîd et Y. Belaîd. *Reconnaissance des formes : méthodes et applications*, InterEditions, France, 1992.
[17] T. Pavlidis, Structural pattern recognition, Springer-Verlag, 1977.
[18] M. Kunt, *Reconnaissance des formes et analyse de scènes,* Presses Polytechniques et Universitaires Romandes, Lausanne, 2000.
[19] P. Leroy. Re*connaissance d'écriture manuscrite dynamique par approche descendante – caractérisation du style de l'écriture et application*, Thèse de Doctorat, Université de Rennes 1, avril 1997.
[20] E. Marisa. *Automatic recognition of handwritten dates on brazilian bank cheques*, Thèse de Doctorat, Ecole de Technologie Supérieure, Université du Québec, 2003.
[21] A. Koerich. *Large vocabulary off-line handwritten word rocognition*, Thèse de Doctorat, Ecole de Technologie Supérieure, Université du Québec, 2002.
[22] M. Cheriet and C. Y. Suen. *Un système neuro-flou pour la reconnaissance de montants numériques de cheques arabe*, *Pattern Recognition letters* 14(1993), pp. 1009-1017.
[23] E. Davalo et P.Naim, *Des réseaux de neurones* , Editions Eyrolles, France 1993.