HAL
archives-ouvertes.fr

Modélisation Markovienne Planaire pour la reconnaissance de l'écriture arabe

Massmoudi Touj Sameh, Najoua Essoukri Ben Amara, Hamid Amiri

▶ **To cite this version:**

Massmoudi Touj Sameh, Najoua Essoukri Ben Amara, Hamid Amiri. Modélisation Markovienne Planaire pour la reconnaissance de l'écriture arabe. Jun 2004, 2004. <sic_00001175>

HAL Id: sic_00001175
https://archivesic.ccsd.cnrs.fr/sic_00001175

Submitted on 6 Dec 2004

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Modélisation markovienne planaire pour la reconnaissance de l'écriture arabe

Sameh Masmoudi Touj [1] – Najoua Essoukri Ben Amara [2] – Hamid Amiri [3]

[1,2,3] Laboratoire des Systèmes et de Traitement du Signal-ENIT

B.P 37 Belvédère 1002, Tunisie

[1,3] Ecole Nationale d'Ingénieurs de Tunis
[2] Ecole Nationale d'Ingénieurs de Monastir

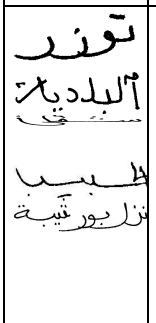
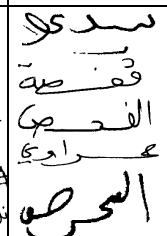
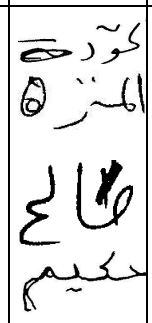
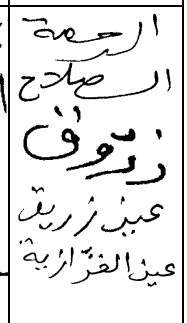
[1]Sameh.Masmoudi@isetgb.rnu.tn – [2]Najoua.BenAmara@enim.rnu.tn

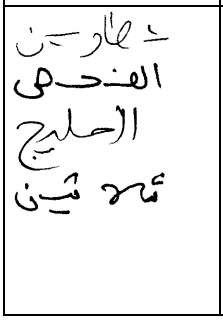
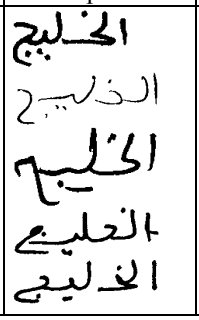
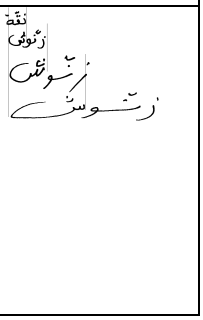
Résumé *: Dans ce papier, nous proposons une approche de reconnaissance de mots manuscrits arabes multi-scripteurs. La méthode adoptée repose sur l'exploitation de sources multiples d'information aussi bien au niveau de la description qu'au niveau de la classification. Une modélisation hybride de type planaire markovienne a été retenue permettant de suivre les variations horizontales et verticales de l'écriture. Cette modélisation s'appuie sur différents niveaux de segmentation : horizontal, naturel et vertical. Le processus de segmentation conduit à la décomposition de l'écriture en entités élémentaires, de morphologies relativement simples et spécifiques à chaque bande horizontale. Le choix de primitives de différents types s'impose alors afin d'assurer une description efficace. Différentes architectures de modélisation s'avèrent aussi indispensables. La classification est enfin réalisée grâce à un Modèle de Markov Caché de type planaire.*

Mots-clés *:* Reconnaissance de l'écriture arabe manuscrite, Modèles de Markov Cachés Planaires, segmentation, primitives variées.

1 Introduction

La reconnaissance de l'écriture arabe a connu un essor important durant cette dernière décennie [Feh 99, Ess 02]. Les différents travaux réalisés reposent sur des approches très variées (structurelles, statistiques, neuronales, géométriques..). Cependant, la reconnaissance de l'écriture arabe manuscrite reste toujours un problème ouvert à cause de sa grande variabilité. Nous retenons essentiellement des problèmes de discontinuité de l'écriture, d'inclinaison, de chevauchement et d'accolement de pseudo mots, de grande variabilité inter et intra scripteurs, de variations de dimensions des pseudo mots. Les signes diacritiques causent également problème [Ess 04]. Le Tableau 1 résume les principaux problèmes liés au manuscrit Arabe hors ligne.

Connexion de pseudo mots	Elongations variables	Ecritures rectifiées	Inclinaisons différentes
			

Ecritures discontinues	Aspect multi scripteurs	Variations des dimensions
		

TAB. 1 – Exemples de noms de villes tunisiennes illustrant différentes difficultés relatives au manuscrit Arabe hors ligne, d'après [Ess 04].

Par ailleurs, les Modèles de Markov Cachés ont déjà montré leur capacité d'absorption des variabilités de l'écriture [Ani 92]. Une architecture de type planaire a été adoptée afin de modéliser l'information spatiale

présente dans les différentes zones de variations logiques du script [Mil 01]. Une première étape de segmentation horizontale permet de délimiter les cinq bandes logiques de variation horizontale de l'écriture arabe. Ces zones correspondent respectivement aux diacritiques supérieurs et inférieurs, aux extensions supérieures et inférieures et à la zone médiane. Une deuxième étape de segmentation à deux niveaux (naturel et vertical), permet d'extraire les différentes entités ou graphèmes associés à chaque bande. Les résultats issus du processus de segmentation sont donc répartis dans les différentes bandes logiques de variation. Chaque classe de graphème présente des spécificités structurelles et morphologiques qui lui sont propres. Nous avons opté alors pour une caractérisation spécifiée, adaptée à la nature de l'information présente dans chaque niveau de variation. Ceci représente d'ailleurs l'une des orientations les plus recommandées en reconnaissance de l'écrit et particulièrement dans le cas de l'arabe [Ess 03]. Nous avons également opté pour des architectures de modèles appropriées aux variations de chaque bande horizontale. Pour cela, nous avons défini un modèle markovien analytique constitué d'une succession de modèles élémentaires capables de suivre les variations de la bande centrale. Des modèles markoviens droite gauche, utilisant un double jeu d'observations, ont été utilisés pour les autres bandes. Le processus de décision repose sur une classification au sens du maximum de vraisemblance.

Dans la section suivante, nous rappelons brièvement l'architecture globale du modèle retenu ainsi que la procédure de segmentation. La section 3 concerne le choix des primitives adoptées pour les différentes zones d'information. La description des architectures de modélisation est adressée dans la section 5. Nous terminons enfin par les résultats des expérimentations et la conclusion.

2 Modélisation adoptée et procédure de segmentation

La prise en considération des variations bi-dimensionnelles de l'écriture arabe, nous a conduit naturellement à l'utilisation de modèles stochastiques de type Markovien Planaire (PHMM-Planar Hidden Markov Models [Lev 92]). L'architecture retenue est formée de cinq MMC-Modèles de Markov Cachés horizontaux secondaires associés aux zones horizontales logiques de variation du manuscrit arabe (FIG. 1). Dans le sens vertical, un modèle markovien haut/bas modélise les transitions entre les bandes horizontales tout en tenant compte des différentes variations morphologiques du script arabe (absence de certaines zones). De ce fait, la segmentation représente une étape cruciale dans notre approche.

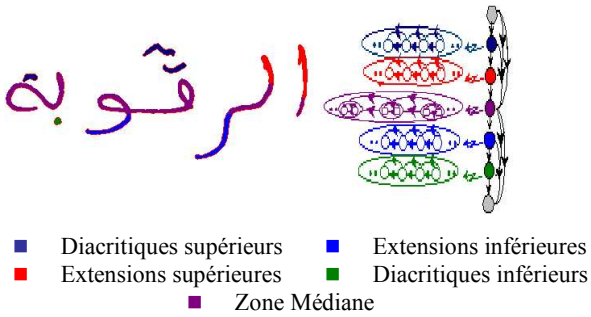


FIG. 1 – Architecture du PHMM développé correspondant au nom de ville «الرقوبة».

La procédure de segmentation est subdivisée en quatre étapes principales : une étape préliminaire de prétraitements, une étape de segmentation horizontale suivie par une étape de segmentation verticale qui concerne uniquement la zone médiane. La quatrième étape permet de calculer les durées chronologiques relatives aux positions des graphèmes associés aux extensions et aux diacritiques [Mas 03]. En fait, les bandes verticales délimitées dans la zone médiane sont utilisées comme des marques de référence pour le calcul de la durée entre les composantes dans les zones des diacritiques et des extensions. La Figure 2 donne une illustration du résultat obtenu dans le cas du mot « الرقوبة».

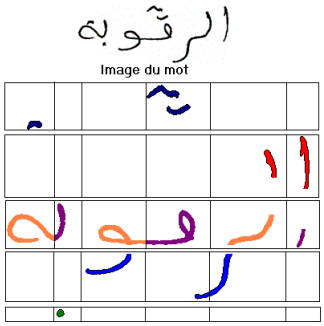


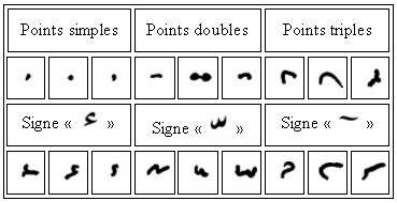
FIG. 2 – Exemple du résultat obtenu par le processus de segmentation.

3 Choix des primitives

A la sortie de l'étage de segmentation, nous obtenons un jeu de graphèmes de différents types. En effet, les entités dérivées de la segmentation de la zone des diacritiques, diffèrent de celles issues aussi bien de la zone des extensions que de la zone médiane. En outre, les entités obtenues dans chacune des zones horizontales, sont en nombre réduit et de formes simplifiées. Nous avons donc pu ramener la complexité du manuscrit AOCR à l'étude d'entités moins complexes morphologiquement. Le choix des primitives s'avère alors délicat, en effet, nous avons été amenés à choisir des caractéristiques appropriées à la nature topologique des différents graphèmes afin d'assurer une description efficace et fiable.

3.1 Zones des diacritiques

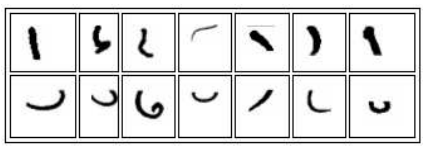
Les diacritiques font partie des deux bandes logiques supérieure et inférieure. Cependant les deux bandes se différencient par la nature des classes de diacritiques leur appartenant, les signes « ـّ » et « ـَ » ainsi que les points triples par exemple ne peuvent faire partie que de la zone des diacritiques inférieurs. Le tableau 2 représente quelques échantillons des formes de base des diacritiques issus de la phase de segmentation. Ces différentes formes se distinguent généralement par des critères de dimensions et de densité en pixels.



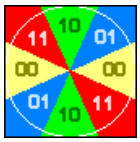
TAB. 2 – Exemples de graphèmes issus des zones des diacritiques

3.2 Zones des extensions

Les ascendants (hampes) et les descendants (jambages) sont extraits à la suite de la délimitation de la bande centrale (Tableau 3). L'étude du jeu d'entités associées aux extensions montre une grande variabilité morphologique tant au point de vue forme que dimensions, ce qui est dû essentiellement aux variabilités inter et intra-scripteurs accentuées par les variations de l'épaisseur de trait. Cependant, nous notons que ces différences morphologiques sont principalement de nature directionnelle, comme le montre le Tableau 3. Nous avons donc pensé à utiliser des caractéristiques de type directionnel par l'application de la Transformée de Hough standard. Cette technique, connue par sa robustesse et son invariance aux bruits, a été utilisée pour la détection des segments de droites dans des images. Cette transformation associe à chaque entité considérée une enveloppe contenant des informations directionnelles pertinentes. Nous avons utilisée l'approche développée au sein de notre équipe, par S. Touj [Tou 02]. Nous subdivisons ainsi chaque graphème en trois bandes horizontales de même hauteur et nous déterminons pour chaque bande la gamme de direction dominante. (FIG. 3).



TAB. 3 – Echantillons de graphèmes associés aux bandes des extensions.

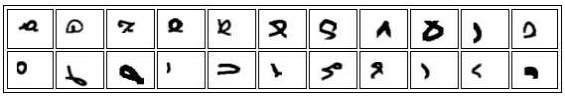


(a) (b)

FIG. 3 – (a) Gammes de directions pour la description des extensions (b) Exemple de graphème de la zone des extensions, le code extrait est : 10 10 00.

3.3 Zone médiane

La zone médiane est connue pour sa richesse en information, c'est le lieu des ligatures horizontales des caractères, des boucles et des caractères centrés. De ce fait, les composantes ou graphèmes issus de cette zone portent une information très variée (Tableau 4). Pour cela nous avons opté pour l'utilisation de sources d'information multiples : un ensemble de primitives structurelles fondés sur des indices visuels et un ensemble de descripteurs traduisant le comportement du processus de segmentation. Un vecteur de 8 primitives est alors extrait. Les caractéristiques sont relatives entre autres à la nature du point de segmentation droite et gauche, à la liaison éventuelle du graphème avec une extension supérieure ou inférieure et à la présence d'un signe diacritique au-dessus ou au-dessous du graphème.



TAB. 4 – Echantillons de graphèmes issus à la zone médiane

4 Modélisation

Nous définissons deux architectures de MMCs pour les modèles secondaires. Une architecture analytique pour la description de la bande centrale et une architecture droite gauche à double jeu d'observations pour les autres bandes horizontales. Le modèle vertical est enfin entraîné statistiquement sur toute la base d'apprentissage.

4.1 Modélisation de la zone médiane

Après la séparation des différentes bandes horizontales, la zone médiane apparaît comme une succession d'entités abstraites que nous appelons Pseudo Caractères (PCs). L'ensemble des PCs forme un alphabet spécifique à la zone médiane (TAB. 5). Etant donné la multitude de combinaisons de PCs associées à cette zone, une modélisation markovienne de type analytique semble être appropriée.

I	D	M	F	I	D	M	F	I	D	M	F
ا	-	-	ـا	ـسـ	ـسـ	ـسـ	ـس	ـل	-	-	ـل
ع	ـعـ	ـعـ	ـع	ـهـ	ـهـ	ـهـ	ـه	ـم	ـمـ	ـمـ	ـم
ه	ـهـ	ـهـ	ـه	و	-	-	ـو	ـي	ـيـ	ـيـ	ـي
ك	ـكـ	ـكـ	ـك	ـحـ	ـحـ	ـحـ	ـح	ـط	ـطـ	ـطـ	ـط
ص	ـصـ	ـصـ	ـص	ن	ـنـ	ـنـ	ـن	ى	ـىـ	ـىـ	ـى

TAB. 5 – Alphabet des pseudo caractères.

Un MMC est composé de deux processus stochastiques dont un est caché, l'autre est observable. Dans le cas de la zone médiane, le processus caché modélise le résultat de la segmentation grâce d'une part aux transitions à l'intérieur du PC et d'autre part aux transitions entre les PC successifs. Le processus observable correspond aux graphèmes obtenus qui sont classés pour correspondre à une des observations du modèle. Deux hypothèses simplificatrices ont été considérées à ce niveau pour réduire le nombre des états hypothétiques constituant le processus caché :

- Chaque PC peut être segmenté au maximum en deux parties : droite et gauche ce qui correspond à la limite de la sur-segmentation.
- Chaque graphème peut être formé au maximum de deux PC successifs ce qui correspond à la limite de la sous-segmentation.

Nous détaillons dans les sous sections suivantes les principales étapes d'estimation du modèle de la bande centrale.

4.1.1 Modèle du PC

Il s'agit d'un modèle droite-gauche à deux états selon les hypothèses formulées ci-dessus. Les probabilités de transitions entre les différents états sont estimées sur toute la base ce qui augmente considérablement le nombre d'échantillons par PC. Nous estimons deux principales probabilités Pr1(α) et Pr2(α) : Pr1(α) est la probabilité que le PC « α » soit segmenté en un seul graphème et Pr2(α) est la probabilité que le PC soit segmenté en deux graphèmes. Pr2(α) est directement déduite de Pr1(α), en effet, Pr2(α)= 1 - Pr1(α).(FIG. 4)

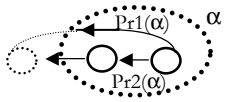


FIG. 4 – Modèle du Pseudo Caractère α.

4.1.2 Modèle de la bande centrale

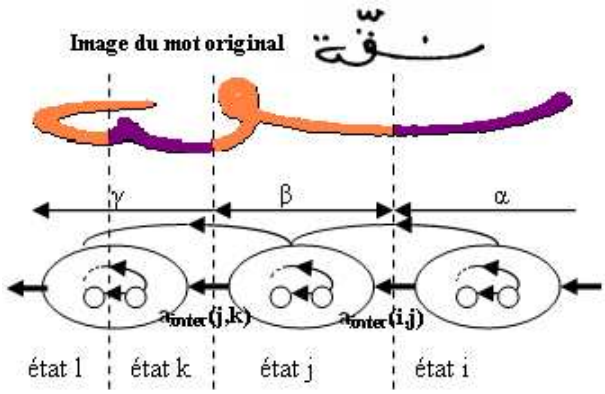
Un mot est une succession ordonnée de PCs. Le modèle de la bande centrale peut être vu comme étant une concaténation des modèles de PCs qui le composent. Pour cela, nous commençons par définir les modèles de bigrammes ; un bigramme étant la succession de deux PCs consécutifs. Les paramètres du modèle du bigramme sont calculés sur toute la base. Nous estimons les deux probabilités Pr1(α,β) et Pr2(α,β) qui correspondent respectivement à la probabilité que les deux PCs successifs α et β appartiennent tous les deux au même graphème et à la probabilité qu'ils soient séparés.

4.1.3 Estimation des paramètres du modèle de la bande centrale

Des calculs statistiques sont effectués sur une base de graphèmes étiquetés pour l'estimation des différents paramètres du MMC associé à la bande centrale [Mas 04b].

Les termes de la matrice A notés a_{ij} sont les probabilités de transition entre les différents états du modèle. Nous considérons que ces termes sont les produits de deux probabilités $a_{inter}(i,j)$ (la probabilité de transition entre les états) et $a_{intra}(j)$ (la probabilité de production de l'état q_j tel qu'il a été observé) (FIG. 5):

$$a_{ij} = a_{inter}(i,j) \,.\, a_{intra}(j).$$



$a_{ij} = a_{inter}(i,j) \,.\, a_{intra}(j) = Pr1(\beta) \,.\, Pr2(\beta,\gamma)$
$a_{jk} = a_{inter}(j,k) \,.\, a_{intra}(k) = Pr2(\gamma) \,.\, 1$

FIG. 5 – Exemple de calcul des probabilités de transitions.

Les termes de la matrice B représentent les probabilités d'observation des symboles. Ces derniers sont directement émis par les états du modèle final de la bande centrale. Les éléments de cette matrice sont estimés sur toute la base d'apprentissage.

4.2 Modélisation des zones des diacritiques et extensions

Pour la modélisation des zones des diacritiques et des extensions, nous avons opté pour des MMCs droite-gauche admettant un processus d'émission d'observations double. En effet, des observations sont d'une part, émises au niveau des états, elles correspondent aux vecteurs de primitives extraits des entités formant ces zones. D'autres observations sont émises au niveau des transitions, elles correspondent aux durées chronologiques associées à ces entités.

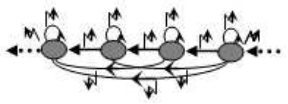


FIG. 6 – Architecture des MMCs des zones des diacritiques et des extensions.

4.3 Modèle vertical

C'est un modèle haut bas autorisant les transitions entre les différents super-états. L'apprentissage du modèle vertical se fait de manière statistique. Selon la modélisation adoptée, uniquement 12 transitions sont possibles entre les super états. Les probabilités de ces transitions calculées sur toute la base d'apprentissage permettent de pondérer les probabilités d'émission

obtenues pour chaque modèle secondaire ce qui nous permet de comptabiliser la probabilité totale d'émission d'un échantillon par le MMC planaire.

5 Expérimentations

Les tests ont été effectués sur un corpus de noms de villes tunisiennes qui sont extraits de la base de données IFN/ENIT [Pec 02]. Notre choix a été porté sur des noms de villes de fréquences variables pour montrer la capacité de l'approche adoptée à résoudre le problème de manque de données. Nous avons utilisé les échantillons de la sous base « a » (1167 échantillons d'un corpus de 25 noms de villes soit 8381 graphèmes) pour l'apprentissage et ceux de la sous base « b » (1180 échantillons) pour le test. Nous avons d'abord utilisé le modèle de la zone médiane pour la classification pour mettre en évidence l'apport de cette bande dans la descrimination entre les noms de villes. Le taux de reconnaissance obtenu sur la base de test est de 73%[Mas 04a]. Le taux moyen de reconnaissance obtenu pour les noms de ville les moins fréquents (moins de 25 échantillons par sous base) est passé de 44% en utilisant un MMC simple à 72% en utilisant un MMC analytique ce qui justifie le choix de l'architecture adoptée. Des tests utilisant le modèle planaire dans sa globalité ont donné des taux de reconnaissance de 88,7% ce qui montre la pertinence de l'information présente dans les bandes des diacritiques et des extensions.

6 Conclusion

Dans ce papier, nous avons proposé une approche de modélisation planaire qui se base sur les modèles de Markov cachés. L'utilisation de modèle hybride ainsi que des primitives de nature variée au niveau de la caractérisation, nous a permis de surmonter efficacement les problèmes majeurs de variations morphologiques de l'arabe manuscrit. La modélisation planaire retenue a permis de mettre en évidence les cinq bandes de variation horizontale de l'écriture arabe et de prendre en considération les différents types de variations morphologiques propres à ce script. Pendant la phase de segmentation verticale, les différentes bandes délimitées sont subdivisées en un ensemble de graphèmes. Ces graphèmes présentent des spécificités topologiques et morphologiques variées ce qui nous a conduit à adopter des techniques différentes pour leur description. Une technique de classification par des critères structurels a été jugée appropriée pour la caractérisation des différentes composantes diacritiques. Les orientations spécifiques des différents graphèmes des zones des extensions nous ont plutôt guidé à l'extraction de primitives de type directionnel. Pour ce faire nous nous sommes appuyés sur la transformée de Hough standard. Quand à la zone médiane, étant donné sa richesse en information et la complexité des dessins des graphèmes qui lui sont associés, nous avons opté pour la combinaison d'informations structurelles et topologiques qui reflètent le comportement du processus de segmentation. Les différentes suites d'observations issues des différentes zones de variation sont utilisées par la suite comme entrées pour entraîner les MMC secondaires correspondants. Un modèle vertical de type haut bas permet de corréler les résultats des différents modèles secondaires. Les premiers résultats obtenus sont encourageants.

7 Bibliographie

[Ani 92] J. ANIGBOGU: "Reconnaissance de textes imprimés multifontes à l'aide de modèles stochastiques et métriques". *Thèse de doctorat, Université de Nancy I*, 1992.

[Ess 02] N. ESSOUKRI BEN AMARA: "Sur la problématique et les orientations en reconnaissance de l'écriture arabe". *Colloque International Francophone sur l'Ecrit et le Document CIFED'02*, Hammamet, Tunisie Octobre 2002, pp1-10.

[Ess 03] N.ESSOUKRI BEN AMARA, F. BOUSLAMA: "Classification of Arabic script using multiple sources of information: State of the art and perspectives". *International Journal on Document Analysis and Recognition IJDAR*, 2003, pp195-212.

[Ess 04] N. ESSOUKRI BEN AMARA, N. ELLOUZE : "Overview and advances in Arabic Optical Character Recognition". *Asian Journal on Information Technology*, Vol 3, N°4, 2004.

[Feh 99] M.C. FEHRI : "Reconnaissance de textes arabes multifonte à l'aide d'une approche hybride neuro-markoviennes". *Thèse de doctorat, Université de Tunis II*, Tunisie, 1999.

[Lev 92] E. LEVIN, R. PIERACCINI : "Dynamic planar warping for optical character recognition". *Proc. International conference on acoustics, speech and signal processing ICASSP'92*,1992, pp. III-149-III-152.

[Pec 02] M. PECHWITZ, S. SNOUSSI MADDOURI, V. MÄRGNER, N. ELLOUZE, H.AMIRI, "IFN/ENIT: "Database of handwritten arabic words". *Colloque International Francophone sur l'Ecrit et le Document CIFED'02*, Hammamet, Tunisia, October 2002, pp. 129-136.

[Mas 03] S. MASMOUDI TOUJ, N. BEN AMARA, H.AMIRI : « Arabic handwritten words segmentation using a planar hidden Markov modelling », *Signals, Systems, Decision & information technology SSD'03*, Sousse, Tunisie 2003.

[Mas 04a] S. MASMOUDI TOUJ, N. ESSOUKRI BEN AMARA, H.AMIRI : « Approche analytique pour la modélisation de la bande centrale de l'écriture arabe manuscrite », *Congrès international de Signaux, Circuits & Systèmes SCS'04*, Monastir, Tunisie 2004.

[Mas 04b] S. MASMOUDI TOUJ, N. ESSOUKRI BEN AMARA, H.AMIRI : « Estimation des paramètres d'un Modèle de Markov Caché pour la modélisation de la bande centrale de l'écriture arabe manuscrite », *Quatrième Journées Scientifiques des Jeunes Chercheurs en Génie Electrique et Informatique GEI'2004* Monastir, Tunisie 15-17 Mars 2004.

[Mil 01] H. MILED N. ESSOUKRI BEN AMARA, "Planar Markov Modeling for Arabic Writing Recognition : Advancement State" . *International Conference on Document Analysis and Recognition, ICDAR'2001*.

[Tou 02] SO.TOUJ, SA.TOUJ, N. BEN AMARA, H.AMIRI, "Reconnaissance hors ligne de caractères Arabes isolés manuscrits ". *Colloque International Francophone sur l'Ecrit et le Document CIFED'02*, Hammamet, Tunisie Octobre 2002, pp. 169-176.