Une approche a priori pour l'identification du scripteur en reconnaissance optique de l'écriture arabe

Sami Gazzah, Najoua Essoukri Ben Amara

Une approche a priori pour l'identification du scripteur en reconnaissance optique de l'écriture arabe

Sami Gazzah[1] - Najoua Essoukri Ben Amara[2]

[1]sami.gazzah@laposte.net

[2]Ecole Nationale d'Ingénieurs de Sousse
Najoua.BenAmara@eniso.rnu.tn/Najoua.BenAmara@enim.rnu.tn

Résumé : *La reconnaissance optique de l'écriture arabe multi-scripteurs hors ligne est une tâche non triviale à cause des particularités morphologiques de ce script. Les variations inter et intra-scripteur sont accentuées par la nature calligraphique de l'écriture arabe. Nous pensons que l'identification du style du scripteur serait d'un apport non négligeable à la reconnaissance optique hors-ligne multi-scripteurs. La complexité du multi-scripteurs serait ramenée à des problématiques mono-scripteur.*

Dans ce travail, nous faisons état des principaux problèmes liés à l'écriture manuscrite arabe hors ligne multi-scripteurs et nous proposons une approche d'identification du scripteur, basée sur un jeu de primitives structurelles et globales. Le jeu de primitives retenu a été optimisé moyennant les algorithmes génétiques. Deux classifieurs ont été testés : les machines à vecteurs de support et les réseaux de neurones de type perceptrons multicouches. Les meilleurs taux enregistrés sont de 94.73% sur une base de 120 échantillons d'un texte d'une lettre préalablement défini de manière à assurer la représentativité des différentes formes internes de chacun des caractères arabes. 60 scripteurs ont participé à la collecte des données.

Mots Clef : Identification du scripteur, ondelettes, les perceptrons multicouches, les machines à vecteurs de support.

1 Introduction

Généralement parlant, la reconnaissance de l'écriture manuscrite hors ligne représente le cas le plus complexe à résoudre dans le domaine de la reconnaissance optique des caractères (OCR-Optical Character Recognition), étant donné les différentes variabilités des écrits et l'absence d'informations dynamiques relatives aux différentes formes considérées. Ces difficultés repoussent d'ailleurs à long terme le traitement automatique de documents manuscrits quelconques et focalisent plutôt les travaux actuels sur les applications spécifiques tel que le tri postal ou la reconnaissance des montants littéraux de chèques [CEH]. Afin de réduire la complexité des tâches, la tendance actuelle est de doter le processus de reconnaissance d'une source d'informations supplémentaire telles que la prise en compte des propriétés structurelles de la langue [KAM 02], du lexique associé au vocabulaire utilisé ou encore de l'identification du style du scripteur [PAQ 00,GAZ 06].

Les travaux en reconnaissance de l'écriture manuscrite arabe hors ligne sont relativement nombreux, cependant la majorité des succès réalisés a été enregistrée dans le cas d'environnements contraints (nombre limité de scripteurs et/ou vocabulaire de test restreint). Ce domaine reste encore très actif en recherche étant donné la complexité de la tâche. Nos différents travaux en reconnaissance de l'écriture manuscrite arabe hors ligne [ESS 04], montrent la complexité de la tâche notamment dans un contexte multi-scripteurs. Les variations inter et intra-scripteur sont accentuées par la nature calligraphique du script arabe. Nous pensons que l'identification du style du scripteur serait d'un apport non négligeable, qui ramènerait la complexité du multi-scripteurs à des problématiques mono-scripteur. A notre connaissance aucune étude, en OCR Arabe, n'a abordé le problème d'identification des caractéristiques intrinsèques à chaque scripteur dans les rares cas où un environnement multi-scripteurs a été considéré.

Dans ce travail, nous proposons une contribution à l'identification du scripteur dans un contexte de reconnaissance de l'écriture arabe multi-scripteurs hors ligne. L'approche adoptée est de type *a priori*,

l'identification du scripteur s'opèrerait en amont du module de la reconnaissance (FIG. 1).

Dans la section suivante, nous donnons un bref aperçu sur les principaux travaux développés sur l'identification du scripteur. Dans la section 3, nous décrivons l'approche d'identification développée. Les résultats enregistrés sont adressés dans la section 4.

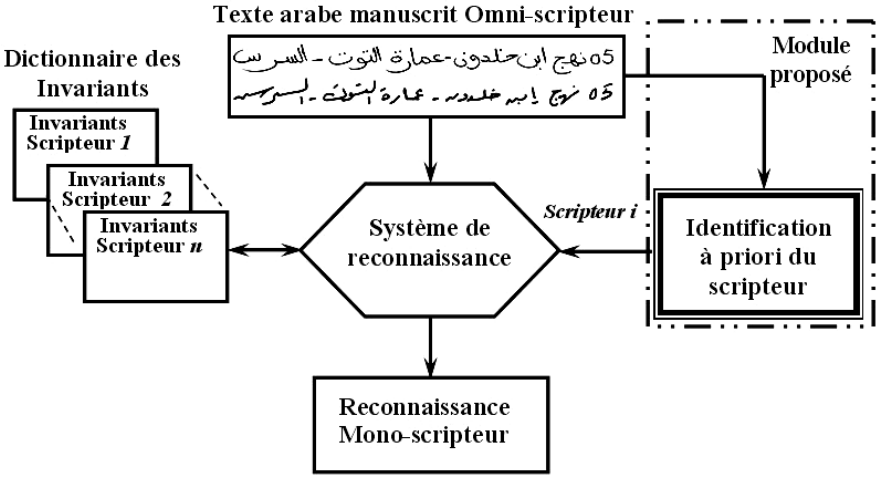
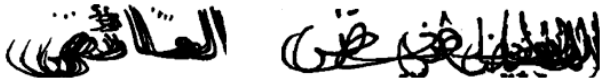


FIG. 1 - Approche *a priori* d'identification du scripteur

2 Aperçu sur les travaux en identification du scripteur

Les premiers travaux sur l'identification du scripteur relèvent du domaine de la biométrie [SAID 00, ZHU 00, SRI 02]. Les nouvelles investigations proposent d'autres horizons d'exploration de cette modalité, telles que l'utilisation de l'information extraite *a priori* sur l'identité du scripteur, pour la recherche d'information visuelle dans une base d'images de documents manuscrits [BEN 03], ou encore pour ramener la complexité de reconnaissance des textes arabes multi-scripteurs à des problématiques mono-scripteurs [GAZ 05a].

L'identification du scripteur est basée sur la différentiation entre les caractéristiques morphologiques qui sont intrinsèques à l'écriture d'un scripteur de celles qui changent d'un scripteur à un autre. Dans la figure 2, nous considérons la superposition d'un mot écrit six fois par le même scripteur (FIG. 2a) et celle produite par six scripteurs différents (FIG. 2b). Cette figure illustre les variations intra et inter-scripteurs et montre une stabilité relative de l'écriture pour le même scripteur (invariants du scripteur) [NOS 99].



(a) (b)

FIG. 2 - Illustration de la variabilité intra et inter scripteurs par la superposition (a) d'un mot écrit 6 fois par le même scripteur et (b) d'un mot écrit par 6 scripteurs différents.

Par ailleurs, les travaux sur le style du scripteur sont abordés selon deux approches : l'identification et la vérification. L'identification du scripteur consiste à reconnaître l'écriture d'un scripteur parmi un ensemble d'écritures multi-scripteurs, tandis que la vérification du scripteur consiste à déterminer si deux documents manuscrits ont été écrits par le même scripteur ou par deux scripteurs différents. L'identification du scripteur suppose généralement que le système a appris au préalable des échantillons d'écritures de chaque candidat. Le cas de la vérification peut être considéré comme une discrimination entre deux classes (scripteur authentique/autre scripteur) [HEU 03]. Plusieurs travaux ont été menés dans ce domaine dans le cas du latin et du chinois [ZHU 00, WAN 03].

Le tableau 1 donne une sélection de travaux développés dans ce domaine.

3 Une approche d'identification du scripteur en OCR arabe

Nous avons développé un module d'identification du style du scripteur qui se placerait en amont d'un système OCR arabe, ce qui ramènerait la complexité du multi-scripteurs à des problématiques mono-scripteurs (FIG. 1). Le schéma du module proposé suit le schéma classique d'un système de reconnaissance de formes : acquisition, prétraitement, extraction des primitives, classification (ou apprentissage selon la phase du traitement). Nous considérons dans cette section les étapes d'extraction des caractéristiques et de classification.

3.1 Extraction du jeu de primitives

Nous avons cherché à déterminer un jeu de primitives qui caractérise le mieux le style d'écriture de chaque scripteur et qui le différencie des autres scripteurs de la base. Les différents tests effectués ont conduit à la prise en compte de deux types de primitives : structurelles et globales, l'entité de base considérée étant la ligne de texte [GAZ 05a].

Références	Base de tests	Primitives	Classification	Performances
[ZOU 05]	• 500 textes arabes : 25 pages écrites par 20 scripteurs.	• Filtres de Gabor	• Distance euclidienne.	92,8%
[SCH 04]	• Apprentissage : 300 pages écrites par 100 scripteurs • Tests : 150 autres pages	• Carte de Kohonen pour la classification des Contours des graphèmes.	• K plus proches voisins • Distance de Hamming.	97 %
[BEN 03]	• Base 1 écrite par 88 scripteurs. • Base 2 écrite par 36 autres scripteurs	• Extraction des graphèmes (du 1er, 2e et 3e niveau) qui sont définis comme des groupes d'invariants.	• Recherche par le contenu graphique dans l'ensemble des documents de référence, • Mesure de similarité entre chaque document et la requête.	95.45% avec la Base 1 et 93.3% avec la Base 2
[WAN 03]	• 34 Caractères Chinois isolés écrits 16 fois chacun par 25 scripteurs	• Primitives directionnelles extraites du contour des caractères	• Distance euclidienne	96,12 %
[SRI 02]	• Une lettre copiée 3 fois par chaque scripteur (1000 scripteurs)	• Mesure de la pression sur le stylo. • Mesure des mouvements de l'écriture. • Mesure des mouvements élémentaires. •Direction, hauteur	• Réseau de neurones de type PMC	98 %
[SAID 00]	• 1000 pages écrites par 40 scripteurs	Filtres de Gabor	• K plus proches voisins • Distance euclidienne	96%

TAB. 1 – Sélection de travaux développés sur le style du scripteur.

1. Primitives structurelles :

L'étude morphologique sur la base de test a montré que la hauteur de la ligne de texte, l'espace moyen entre pseudo mots, l'inclinaison des hampes ainsi que les caractéristiques dimensionnelles des points diacritiques, représentent des primitives discriminantes, qui marquent bien le style d'un scripteur, nous les avons retenues pour la caractérisation du style de l'écriture (FIG.3).

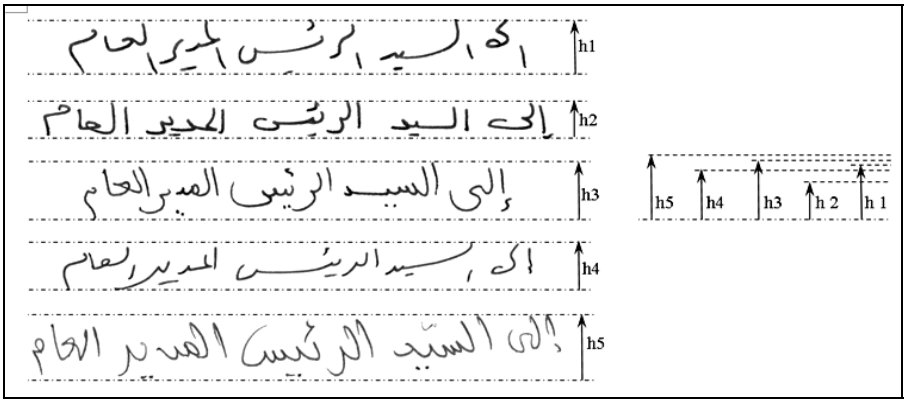
2. Primitives globales

Nous avons opté pour la description de la texture de l'écriture par la mesure de l'entropie de chacune des lignes de texte et par la considération de caractéristiques issues de l'application de la transformée en ondelettes. En effet, nos expérimentations en reconnaissance de fontes arabes [GAZ 03] nous ont incités à tester les ondelettes pour la caractérisation de l'écriture. Les primitives choisies sont l'écart type et la moyenne des matrices associées aux images de l'approximation, aux détails horizontal, vertical et diagonal.

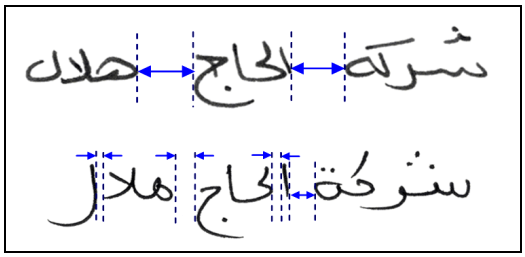
Ainsi, au total le jeu de primitives comporte 20 caractéristiques : 12 de type structurel et 8 de type global. Nous avons exploré, par la suite, les algorithmes génétiques pour opérer la sélection des primitives les plus discriminantes et non redondantes. L'application de l'algorithme d'optimisation a permis de retenir uniquement 12 caractéristiques parmi les 20. Les primitives retenues sont présentées dans le tableau suivant [GAZ 06].

F1	hauteur de la ligne
F3	Moyenne des inclinaisons des hampes.
F4	Ecart type des inclinaisons des hampes.
F5	Moyenne de l'épaisseur des hampes.
F7	Moyenne du ratio hauteur/largeur du rectangle englobant un point diacritique.
F9	densité de pixels dans un rectangle englobant un point diacritique.
F11	Moyenne de la matrice de l'image d'approximation issue des ondelettes.
F12	Ecart type de la matrice de l'image d'approximation issue des ondelettes.
F14	Moyenne de la matrice du détail horizontal.
F17	Ecart type de la matrice du détail vertical.
F18	Moyenne de la matrice du détail diagonal.
F20	Entropie.

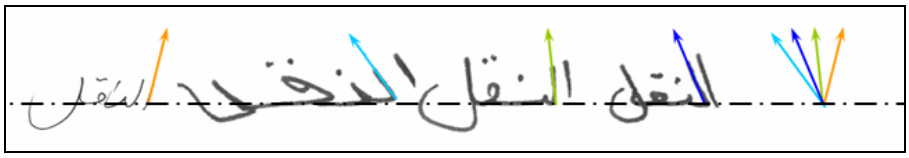
TAB. 2 – Primitives retenues extraites sur chaque ligne de texte.



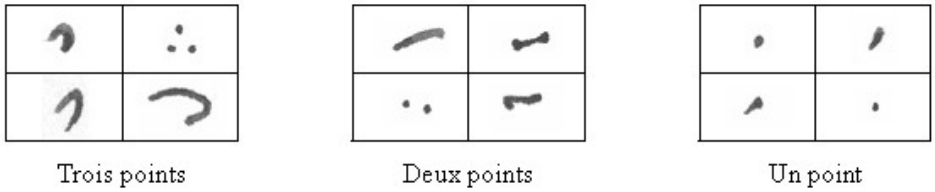
(a)



(b)



(c)



(d)

FIG. 3 - Illustration des variations structurelles de l'écriture, (a)Hauteur ligne de texte, (b) Espaces entre pseudo mots, (c) Inclinaisons de l'écriture, (d) Exemples de formes de points diacritiques.

3.2 Classification

Il s'agit de déterminer l'identité du scripteur étant donné le jeu de primitives préalablement extrait. Pour cela, nous avons implémenté deux classifieurs : les machines à vecteurs de support (SVM-Support Vectors Machines) et les réseaux de neurones de type PMC-Perceptrons multicouches.

1. Classification par les PMCs

Les réseaux de neurones de type PMC sont parmi les techniques les plus couramment utilisées. Ils ont prouvé une grande efficacité notamment en classification [EGM 02]. Nous avons retenu une architecture modulaire constituée de 60 réseaux PMCs [GAZ 06]. Chaque réseau est constitué d'une couche d'entrée à 12 neurones (correspondants aux 12 primitives retenues), d'une couche cachée ayant un nombre de neurones variant, selon l'identité du scripteur, de 10 à 14 (ce nombre est déterminé pendant la phase d'apprentissage) et d'une couche de sortie à un seul neurone correspondant au scripteur authentique/ou autre scripteur.

2. Classification par les SVMs

Les machines à vecteurs de supports proposent une méthode de classification supervisée qui se base sur le principe de la minimisation du risque structurel. Cette approche a été initialement utilisée pour l'optimisation de l'hyperplan linéaire de discrimination entre deux classes. Ensuite, l'utilisation des fonctions noyaux a

permis de projeter les données non linéairement séparables dans un espace augmenté afin de les rendre linéairement séparables. Les SVMs ont été notamment appliquées avec succès à différents problèmes en reconnaissance de formes [BUR 98, AYA 04].

Les différents tests effectués nous ont conduit au choix d'une architecture SVM dont le noyau est une fonction à base radiale [GAZ 05b]. L'architecture adoptée est modulaire constituée de 60 modules SVMs différents, chacun étant associé à un style de scripteur, le but est d'établir un modèle propre à chaque scripteur.

4 Expérimentations et résultats

Dans cette section, nous définissons d'abord notre base de tests, ensuite nous développons les résultats obtenus avec chacun des deux classifieurs implémentés.

4.1 Description de la base de données

Nous avons rédigé une lettre qui comporte un texte arabe composé de 505 caractères, 15 chiffres et 6 signes de ponctuation [GAZ 05a]. Le choix du contenu de la lettre a été fait de manière à assurer la représentativité des différentes formes internes (au nombre de quatre au maximum) de chacun des caractères arabes. 60 scripteurs différents ont participé à la collecte des données. Chaque scripteur a écrit la lettre 3 fois, ce qui a donné un total de 180 pages en format A4, les deux tiers ont servi pour l'apprentissage, le reste a été utilisé pour les tests. Un stylo noir, une feuille blanche et un support d'écriture ligné, ont été fournis à chaque scripteur ; le problème d'accolements de lignes de texte n'étant pas considéré dans ce travail. Les différents échantillons de textes ont été numérisés en niveaux de gris à une résolution de 300 dpi.

4.2 Résultats enregistrés

Les deux architectures implémentées, SVM et PMC, ont été entraînées par les mêmes vecteurs de primitives (dont les valeurs des composantes ont été préalablement normalisées) avec une même stratégie d'apprentissage. Pour chaque scripteur nous avons fait apprendre au réseau qui lui est associé, aussi bien les bonnes que les mauvaises réponses (parmi l'ensemble de la base d'apprentissage). Cependant nous avons constaté, dans les deux cas de classifieurs, que la grande disproportion entre le nombre de vecteurs primitives associé au scripteur authentique et celui des autres scripteurs, a conduit à un système biaisé, avec un taux de rejet assez élevé. Pour remédier à ce problème nous avons retenu la solution suivante : le réseau associé à un scripteur donné, apprend tous les vecteurs de primitives du scripteur présumé se trouvant dans la base d'apprentissage, cependant il apprend un seul vecteur de primitives pour chacun des autres scripteurs de la base. Ce vecteur est choisi, dans la base d'apprentissage, de manière tout à fait aléatoire. Cette mesure a permis d'améliorer les performances globales de chacun des classifieurs et le taux de rejet a chuté considérablement.

Le tableau 3 donne les performances, par scripteur, enregistrées dans le cas des 2 classifieurs.

Les meilleurs résultats obtenus sont de 94.73 % pour les réseaux PMCs et de 93.76 % pour le cas des SVMs. L'analyse des résultats montre que les PMCs sont relativement plus performants que les SVMs tant au niveau des résultats par scripteur qu'au niveau performances globales. Cependant, d'autres mesures sont en cours de considération afin de tester d'autres architectures SVMs multi-classes.

Scripteur	Taux(%) SVM	Taux (%) PMC	Scripteur	Taux(%) SVM	Taux (%) PMC	Scripteur	Taux(%) SVM	Taux (%) PMC	Scripteur	Taux(%) SVM	Taux (%) PMC
1	93.12	94.16	16	97.74	97.66	31	97.57	93.47	46	90.93	95.12
2	92,05	96.73	17	92.01	96.04	32	96.48	96.04	47	92.17	91.16
3	99.19	96.74	18	93.22	96.96	33	94.88	97.44	48	92.95	94.63
4	93.44	97.20	19	97.74	94.63	34	86.95	96.25	49	94.17	93.01
5	93.93	93.26	20	96.51	96.03	35	96.27	96.37	50	94.85	92.06
6	94.48	94.19	21	94.63	92.29	36	92.54	95.10	51	98.14	98.14
7	96.26	97.20	22	91.74	91.59	37	93.78	93.94	52	92.31	95.10
8	89.8	92.99	23	91.12	92.29	38	95.57	96.97	53	90.89	96.03
9	91.12	92.99	24	98.6	97.67	39	96.73	99.07	54	93.71	93.47
10	91.56	92.74	25	91.61	98.37	40	93.49	95.37	55	96.5	92.54
11	97.67	96.27	26	88.08	93.24	41	94.88	95.35	56	91.45	90.44
12	93.93	96.73	27	91.84	92.54	42	92.68	96.28	57	95.66	94.63
13	90.19	96.03	28	91.59	93.46	43	98.37	97.90	58	95.11	93.72
14	89.72	88.08	29	91.06	91.36	44	91.08	92.77	59	93.44	94.87
15	93.22	91.36	30	89.16	94.63	45	96.2	90.42	60	99.76	99.30

TAB. 3 – Performances par scripteur enregistrées par les 2 classifieurs SVM et PMC.
Taux minimum, Taux maximum, obtenus par classifieur

5 Conclusion

Dans ce travail, nous avons présenté une contribution à l'identification du style du scripteur dans un contexte de reconnaissance de l'écriture arabe multi-scripteurs hors ligne. Différentes expérimentations ont été conduites sur une base de 180 échantillons d'une lettre écrite 3 fois par 60 scripteurs. Au niveau de la caractérisation, nous avons retenu la ligne de texte comme entité de base. Deux types de primitives ont été retenus : des primitives globales issues des ondelettes mettant en évidence les variations de la texture de l'écriture et des primitives structurelles décrivant les variations topologiques du style du scripteur.

Deux classifieurs ont été implémentés : les SVMs et les réseaux de neurones de type PMC. Les premiers résultats obtenus sont encourageants. Les meilleurs taux sont de 94.73%, ont été enregistrés avec un vecteur attribut de 12 primitives et un classifieur de type PMC.

Etant donné l'importance des enjeux considérés, différents tests sont en cours de validation, en vue de réaliser de meilleures performances.

Références

[AYA 04] AYAT N. D., Sélection Automatique de Modèle dans les Machines à Vecteurs de Support : Application à la Reconnaissance d'Image de Chiffres Manuscrits, *PhD thesis, Ecole de Technologie Supérieur / Université de Québec*. 2004.

[BEN 03] BENSEFIA A., PAQUET T., HEUTTE L., Information Retrieval Based Identification. *7th International Conference on Document Analysis and Recognition Edinburgh/Scotland,* 2003, pp. 946-950.

[BUR 98] BURGES C., Tutorial on support vector machines for pattern recognition, *Data Mining and knowledge Discovery 2(2)*, 1998, pp. 121–167.

[CEH] CEHEUX G.R, Traitement automatique de l'écriture et du document, présentation du domaine, www.infress.enst.fr/~elc/GRCE/presentation/Grce2.html.

[EGM 02] EGMONT-PETERSEN M., DE RIDDER D., and HANDELS H., Image processing with neural networks – a review, *Pattern recognition journal*, Published by *Elsevier Science Ltd*. 35- 2002, pp. 2279-2301.

[ESS 04] ESSOUKRI BEN AMARA N., Contribution à la Reconnaissance Optique de l'Ecriture Arabe, Rapport de synthèse, *Habilitation universitaire - Université Tunis El Manar*, 2004.

[GAZ 03] GAZZAH S., ESSOUKRI BEN AMARA N., Etude et Caractérisation des Fontes Arabes, *International Conference on Image and Signal Processing, Marroc*, 2003, Vol. 2 , pp. 650-659.

[GAZ 05a] GAZZAH S., ESSOUKRI BEN AMARA N., Neural Networks and Support Vector Machines Classifiers for Writer Identification using arabic script, *The second International Conference on Machine Intelligence*, *Tozeur/Tunisie*, 2005, pp. 1001-1005.

[GAZ 05b] GAZZAH S., ESSOUKRI BEN AMARA N., Writer Identification using SVM Classifier and Genetic Algorithm for Optimal futures selection, *International Arab Conference on Information Technology, Amman / Jordan* , 2005, pp. 461-466.

[GAZ 06] GAZZAH S., ESSOUKRI BEN AMARA N., Writer Identification Using Modular MLP Classifier and Genetic Algorithm for Optimal Features Selection, T*he Third International Symposium on Neural Networks*. *Chengdu/China*, May 29-31, 2006. (Accepté)

[HEU 03] HEUTTE L., Analyse et Reconnaissance de l'écriture : de Nouvelles Perspectives en Traitement Automatique de Documents Manuscrits, *Rapport de synthèse, Habilitation à Diriger les Recherches*, *Université de Rouen/France*. 2003.

[KAM 02] KANOUN S., ALIMI A., ENNAJI A., LECOURTIER Y., Reconnaissance de Mots Arabes par Approche Affixale, *Colloque International Francophone sur l'Ecrit et le Document*, *Hammamet/Tunisie*, 2002, pp. 21-30.

[ZOU 05] AL-ZOUBEIDY L. M., AL-NAJAR H. F., Arabic Writer Identification For Handwriting Images, *International Arab Conference on Information Technology, Amman*, 2005, pp. 111-117.

[NOS 99] NOSARY A., HEUTTE L., PAQUET T., LECOURTIER Y., Defining writer's invariants to adapt the recognition task». *5th International Conference on Document Analysis and Recognition*, *Bengalore / India*, 1999, pp.765-768.

[PAQ 00] PAQUET T., NOSARY A., HEUTTE L., LECOURTIER Y., Apprendre l'Ecriture du Scripteur pour Adapter la Reconnaissance, *12éme Congrès Francophone AFRIF-RFIA Paris/France*, 2000, Vol. 3, pp337-346.

[SAID 00] SAID H.E.S., TAN T.N., BAKER K.D., Personal Identification Based on Handwriting, *Pattern Recognition*, 2000, vol. 33, pp149-160.

[SCH 04] SCHOMAKER L., BULACU M., FRANKE K., Automatic Writer Identification Using Fragmented Connected-Component Contours, *9th International Workshop on Frontiers in Handwriting Recognition, Tokyo/Japan*, 2004, pp. 185-190.

[SRI 02] SRIHARI S. N., H.CHA S., ARORA H., LEE S., Individuality of handwriting, *Journal of Forensic Sciences*. 2002, Vol.47, No.4, pp. 856-872,

[WAN 03] WANG X., DING X., LIU H., Writer Identification Using Directional Element Features and Linear Transform, *7th International Conference on Document Analysis and Recognition, Edinburgh/ Scotland*, 2003, pp.942-945.

[ZHU 00] ZHU Y., TAN T., WANG Y., Biometric Personal Identification Based on Handwriting, *15th International Conference on Pattern Recognition - Barcelona*, 2000, Vol. II pp. 2797-2800.