Identification de scripteurs pour l'écriture arabe par une approche locale

DJEDDI Chawki, SOUICI-MESLATI Labiba
Laboratoire LRI, Université Badji Mokhtar, BP 12, 23000, Annaba, Algérie.
djeddi_2008@yahoo.fr

Résumé : La plupart des travaux de recherche dans le domaine de l'identification de scripteurs se sont orientés vers des documents écrits en script latin et peu d'études ont été effectuées sur des documents en arabe. Dans cet article, nous nous intéressons à l'identification hors-ligne de scripteurs utilisant l'écriture arabe. Nous proposons une approche locale en mode dépendant du texte où on cherche les formes invariantes et propres à l'écriture de chaque scripteur. Ces formes sont extraites par un découpage de l'écriture suivi d'une classification des formes obtenues qui sont ensuite organisées dans une base de référence pour permettre d'identifier l'auteur d'un document inconnu dans un processus de Pattern Matching. Les résultats obtenus sont de l'ordre de 93.33% en Top 1 et de 100% en Top 2 sur une base de 30 scripteurs.

Mots clés : Ecriture arabe manuscrite, Identification de scripteur, Imagettes, Invariants du scripteur, Mode dépendant du texte, Pattern matching.

Introduction

L'écriture est un acte personnel: chaque scripteur est caractérisé par son écriture, par la reproduction de détails et d'habitudes inconscientes. Plusieurs validations scientifiques de l'individualité de l'écriture ont été effectuées [08, 17] et l'identification de scripteurs a fait l'objet de plusieurs travaux de recherche car elle a une large variété d'applications telles que la sécurité au niveau des d'activités financières par la vérification de signatures [11], l'authentification des auteurs de documents tels que les actes de ventes dans les cours de justice, par exemple. Elle peut aussi être intégrée pour authentifier les auteurs dans des systèmes de navigation et d'interrogation de bases de documents manuscrits anciens numérisés [03]. L'identification de scripteurs est utilisée dans des systèmes de reconnaissance hors ligne de textes manuscrits exploitant le principe d'adaptation à l'écriture à reconnaître [09].

Dans cet article, nous nous intéressons au problème de l'identification d'une personne en se basant sur son écriture. L'identification de scripteur est la tâche qui consiste à identifier, parmi un ensemble de scripteurs connus du système l'auteur du document lu en entrée [11]. Si n'importe quel texte peut être employé pour établir l'identité de l'auteur, la tâche est dite indépendante du texte. Autrement, si un auteur doit écrire un texte prédéfini particulier (tel que sa signature) pour s'identifier ou pour vérifier son identité, la tâche est dite dépendante du texte. L'identification de scripteur peut être effectuée en temps réel (en ligne), où les informations temporelles et spatiales de l'écriture sont disponibles, ou en différé (hors-ligne), quand seulement une image de l'écriture est disponible [03].

La méthode proposée dans ce travail est basée sur une approche locale en effectuant l'extraction des formes spécifiques et invariantes [03, 09, 12, 15] d'un scripteur. Ces formes spécifiques sont trouvées pendant une phase d'apprentissage à partir d'un processus de découpage de l'image en imagettes [12, 15] suivi par une classification séquentielle de ces imagettes.

Toutes les approches qui ont été proposées pour l'identification de scripteur sur des documents arabes manuscrits sont fondées sur des caractéristiques texturales et structurelles (voir section 2). Malgré le fait qu'elles donnent d'assez bons résultats, nous avons choisi de caractériser l'écriture par des caractéristiques locales : graphèmes et imagettes car ce type de caractéristiques n'a jamais été considéré dans le cas de l'identification de scripteurs de documents arabes.

Les graphèmes et les imagettes sont des formes propres à l'écriture d'un scripteur, ils sont extraits à partir de cette dernière et ont la particularité de pouvoir décrire un échantillon d'écriture qu'il soit de grande ou de petite taille. Ce choix de caractéristiques a été motivé par notre étude de certains travaux où des résultats très encourageant ont été atteints pour l'écriture manuscrite latine. Ali Nosary [09] et Ameur Bensafia [03] utilisent des graphèmes issus de la segmentation de mots pour caractériser les scripteurs, alors que Audrey Seropian [13], Siddiqi et al [15] utilisent des imagettes pour modéliser le scripteur.

Nous avons structuré cet article de la manière suivante : dans la première partie (section 2), un état de l'art de l'identification de scripteurs sur des documents arabes est présenté. Dans la seconde partie (sections 3, 4, 5, 6), nous décrivons l'architecture de notre système d'identification et nous montrons comment extraire les formes invariantes et propres à l'écriture d'un scripteur afin de les utiliser pour caractériser ce dernier. Nous présentons comment les similarités entre les formes enregistrées sont utilisées pour différencier entre les scripteurs. Une troisième partie de l'article (section 7) est consacrée à la présentation des résultats des

expérimentations effectuées et nous terminons l'article par une conclusion et plusieurs perspectives d'extensions futures.

1. Etat de l'art de l'identification de scripteur sur des documents arabes

La recherche dans le domaine de l'identification de scripteurs a donnée lieu à de nombreuses études pour les scripts occidentaux, à l'opposé du script arabe qui a été étonnamment peu étudié jusqu'à présent.

Récemment des travaux qui traitent le script arabe ont été proposés, les méthodes développées dans ces travaux utilisent des caractéristiques qui se classent en caractéristiques *texturales* où le contenu du document est vu comme une image non comme une écriture et en caractéristiques *structurelles* qui s'attachent à décrire les particularités de l'écriture. Certains travaux se sont orientés vers une combinaison entre ces deux types de caractéristiques.

La première étude dans le domaine de l'identification de scripteurs pour le script arabe remonte à l'année 2005 où Al Zoubeidy et al [02] ont proposé une méthode globale utilisant un filtre de Gabor multi canal et un calcul de matrice de co-occurrence. Un taux d'identification de l'ordre de 92.8% a été atteint avec la distance euclidienne sur une base de 500 textes arabes (à raison de 25 pages par scripteur).

Gazzah et al [05, 06] proposent une approche d'identification de scripteur, en mode dépendant du texte, basée sur un jeu de primitives structurelles décrivant les variations topologiques du style du scripteur et des primitives globales mettant en évidence les variations de la texture de l'écriture issues de l'application des ondelettes. L'ensemble de primitives a été optimisé moyennant les algorithmes génétiques, les meilleurs résultats obtenus sont de l'ordre de 94.73% pour les réseaux neuronaux de type PMC (perceptron multicouches) [06] et 93.76% pour les SVM (support vector machines) [05] sur une base de 180 échantillons appartenant à 60 scripteurs.

Dans d'autres travaux, Gazzah et al [07] proposent une approche globale en mode dépendant du texte en explorant l'écriture par l'analyse de la texture avec des ondelettes 2D utilisant le Lifting Schème. Une évaluation comparative entre les caractéristiques extraites de la texture par neuf transformations en ondelettes différentes a été effectuée. Des expériences ont été effectuées en utilisant un perceptron multicouches sur une base de 180 échantillons de texte de 60 scripteurs différents. L'identification du scripteur était correcte dans 95.68% des documents considérés.

Al Dmour et al [01] présentent une technique d'extraction de caractéristiques basée sur la combinaison de mesures statistiques et spectrales, les auteurs extraient deux types de caractéristiques issues de la texture, les premières sont obtenues en utilisant les filtres de Gabor à multiples canaux, les secondes à partir de matrices de co-occurrence en niveau de gris. Les caractéristiques les plus discriminantes avaient été choisies en utilisant un module hybride basé sur les algorithmes génétiques et les SVM. Quatre classifieurs ont été testé: SVM, KPPV (K plus proches voisins), un classifieur linéaire discriminant et un autre utilisant la distance euclidienne pondérée. Un taux d'identification de 90% a été atteint sur une base de 20 scripteurs.

Shahabi et al [13] déterminent la performance des caractéristiques utilisées dans [11] sur une base de données de Farsi manuscrit. Les auteurs proposent une approche globale, basée sur la texture de l'image, en mode indépendant du texte. Un ensemble de caractéristiques sont extraites à l'aide des filtres de Gabor multi canaux et les matrices de co-occurrence. Les expériences effectuées sur une base de 25 scripteurs donnent un taux d'identification de 88% en Top 1 et 92% en Top 3.

Dans d'autres travaux, Shahabi et al [14] proposent une approche en mode dépendant du texte, où ils choisissent parmi les filtres de Gabor ceux qui sont appropriées à la structure des textes Farsi manuscrits. Ils utilisent une nouvelle méthode d'extraction de caractéristiques qui se base sur les moments et les transformations linéaires. Deux niveaux ont été considérés dans leurs séries de tests : le niveau mot et le niveau texte. Le taux d'identification correcte est de l'ordre de 45% en Top 1 et 80% en Top 5 au niveau mot, il atteint 82,5% en Top 1 et 100% en Top 5 au niveau texte sachant que 40 scripteurs ont été considérés.

Bulacu et al [04] ont présenté un travail portant sur la combinaison de caractéristiques multiples pour l'identification de scripteur indépendamment du texte. Des fonctions de distributions de probabilités (PDF) sont extraites indépendamment du contenu textuel fournis par les échantillons d'écriture. Les auteurs ont effectués une analyse de la combinaison des caractéristiques texturales et allographiques et ont montré qu'en fusionnant ces caractéristiques, les performances atteintes sont améliorées. Des expériences ont été effectuées sur la base IFN/ENIT [10] et des taux d'identification de l'ordre de 84% en Top 1 et 99% en Top 10 ont été atteints.

2. Architecture du système proposé

Notre système d'identification est basé sur trois étapes principales : prétraitements, extraction de caractéristiques et décision. Une base de référence des scripteurs est créée en effectuant l'extraction des caractéristiques de l'écriture de chaque scripteur. Le scripteur d'un document inconnu est alors identifié ultérieurement pendant une

étape de décision. L'architecture générale de notre système est présentée sur la Figure 1 et les différentes étapes sont décrites dans les prochaines sections.

3. Prétraitements

3.1. Binarisation

La première étape à laquelle est soumise une image de texte manuscrit est la binarisation de cette image. Les images considérées n'étant pas trop bruitées, nous avons mis en oeuvre un seuillage global. Une fois l'image du texte binarisée, nous procédons à l'étape d'extraction des mots.

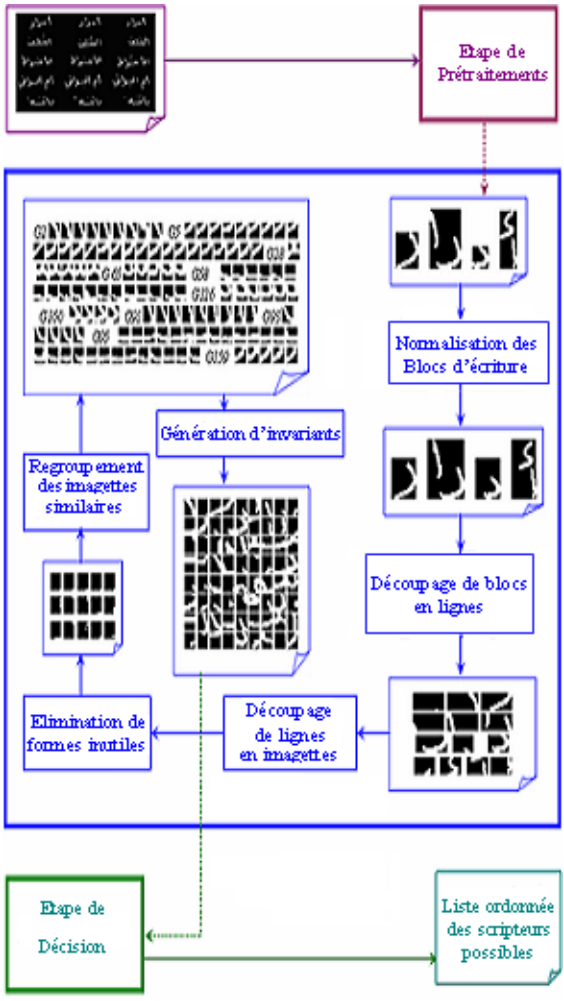

Figure 1 : Architecture générale du système proposé.

3.2. Segmentation en mots

Dans cette étape, notre but est d'isoler chaque mot du texte manuscrit et d'appliquer un découpage spécifique décrit dans la section 5. Nous procédons à une analyse des projections horizontales et verticales de la page d'écriture, ce qui permet d'obtenir la liste des mots d'une page.

3.3. Fragmentation

Une fois que l'image d'un mot est récupérée, nous procédons à sa fragmentation en un ensemble de blocs d'écritures disjoints. Cette opération est basée sur l'analyse des projections verticales et horizontales du mot, ce qui permet d'obtenir un partitionnement de l'image du mot en blocs complètement déconnectés (voir Figure 2).

Figure 2 : L'ensemble des blocs composant le mot «أدرار»

4. Extraction de caractéristiques

Notre méthode d'extraction de caractéristiques consiste à normaliser les blocs d'écriture (cf. 5.1) puis à appliquer un découpage en imagettes de taille N×N sur tous blocs de l'écriture (cf. 5.2). Une fois les imagettes obtenues, on procède à l'élimination des imagettes qui contiennent peu d'information (cf. 5.3) puis un algorithme de regroupement (cf. 5.4) est appliqué sur ces imagettes pour les regrouper dans des classes différentes. Enfin, nous pouvons obtenir les formes caractérisant le scripteur en appliquant un algorithme de génération des formes invariantes du scripteur (cf. 5.5).

4.1. Normalisation des blocs d'écriture

Pour chaque bloc d'écriture dont la largeur et la hauteur ne sont pas un multiple de N on ajoute des lignes vides en haut et des colonnes vides à gauche (Figure 3).

Figure 3 : Normalisation des sous mots du mot «أدرار»

4.2. Découpage de l’écriture

Le découpage est utilisé pour extraire les formes inhérentes au scripteur, donc c'est une partie importante du processus [12, 15]. Après avoir découpé l’image du mot en un ensemble de blocs d’écriture disjoints, nous effectuons le découpage de chaque bloc d’écriture en carrés de taille N×N, l’origine verticale étant ainsi propre à chaque bloc d’écriture.

Figure 4 : Découpage de l’écriture.

L’image d’un bloc d’écriture est découpée de façon régulière selon les lignes et sur chaque ligne un carré peut être décalé vers la gauche pour optimiser le partitionnement de l’écriture. Notons que nous déplaçons un carré vers la gauche pour trouver le premier pixel noir afin de réduire la division d'un trait d’écriture dans deux carrés différents et les carrés sont ainsi mieux placés. La méthode est illustrée par la Figure 4.

4.3. Elimination des formes inutiles

Dans cette étape, nous procédons à l'élimination de certaines formes qui contiennent très peu d’informations et qui sont en fait trop communes à tous les scripteurs. La Figure 5 présente quelques formes inutiles et communes à tous les scripteurs.

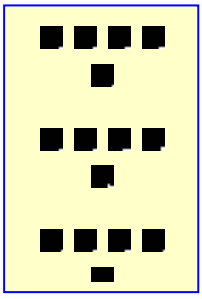

Figure 5 : Quelques formes inutiles et communes à tous les scripteurs

4.4. Regroupement des imagettes

Notre but est de regrouper les imagettes ayant des formes similaires. Pour notre classification, nous avons choisi un algorithme classique de regroupement séquentiel [03, 09, 15]. L'algorithme est simple et ne demande pas un choix préalable du nombre de classes. Nous avons apporté une modification sur cet algorithme. Cette modification consiste à ne mettre un élément dans un groupe que s'il est proche de tous les éléments du groupe au sens de la mesure de corrélation qui se définit comme suit [03, 09, 15]:

$$S(X,Y)=\frac{n_{11}\times n_{00}-n_{10}\times n_{01}}{\sqrt{(n_{11}+n_{10})\times(n_{01}+n_{00})\times(n_{11}+n_{01})\times(n_{10}+n_{00})}}$$

n_{ij} est le nombre de pixels pour lesquels les deux images binaires de même taille X et Y ont la valeur de correspondance suivante :

X(k)=i, Y(k)=j, pour k = 1.. L×H où L est la largeur de l'image et H sa hauteur.

La Figure 6 montre quelques groupes obtenus sur une page d'écriture d'un scripteur de notre base.

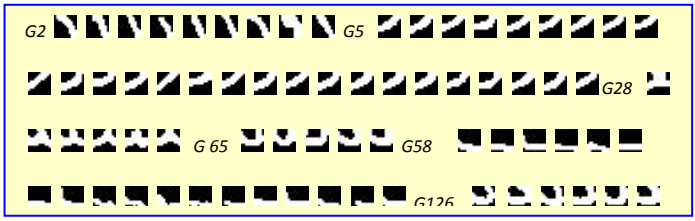

Figure 6 : Quelques groupes obtenus sur la page d'écriture d'un scripteur

4.5. Générations des invariants du scripteur

Après avoir regroupé les imagettes morphologiquement similaires dans différentes classes, nous allons retenir une seule imagette pour chaque classe, celle qui est la meilleure représentante de son groupe. Dans une classe, nous calculons la mesure de similarité [15] de chaque élément par rapport à tous les autres éléments de la même classe. L'élément le plus proche de tous les autres est choisi en tant que représentant. Nous obtenons ainsi une représentation pour des formes propres à l'écriture du scripteur considéré.

Soit *D* un document manuscrit, on a $D = \{x_i / x_i = Représentant\ (C_i)\}$. Quelques formes invariantes d'un scripteur de notre base sont montrées sur la Figure 7.

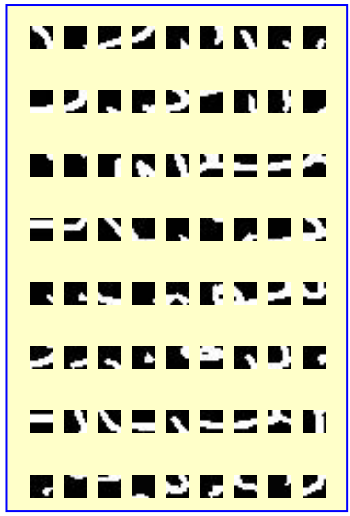

Figure 7 : Quelques formes invariantes d'un scripteur

5. Décision

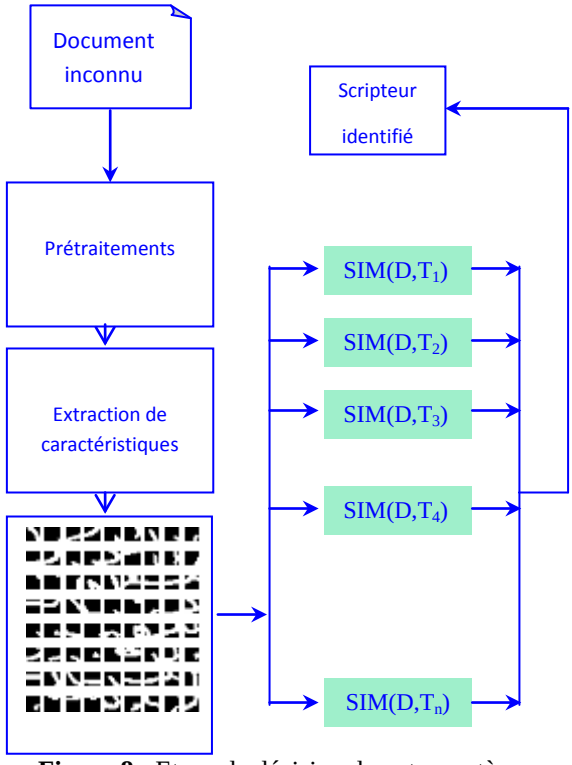

Figure 8 : Etape de décision de notre système.

Décider quel est l'auteur de l'échantillon de l'écriture présentée au système revient à comparer cet échantillon avec tous ceux qui constituent la base de référence. Cette comparaison se fait entre les caractéristiques du document D à identifier et celles d'un document T de la base de référence. Le scripteur du document inconnu D sera finalement le scripteur du document de la base de référence le plus similaire au document inconnu (au sens de la mesure que allons définir) soit donc :

$$\text{Scripteur(D)} = Arg \max_{Ti \in Base}(SIM(D,Ti))$$

Sachant que SIM(D,T_i) est la mesure de similarité entre D le document non spécifié et le document de référence T_i, tous les deux représentés par leurs caractéristiques. On définit la mesure de similarité entre T_i et D par :

$$SIM(D,T) = \frac{1}{Card(D)} \sum_{i=1}^{Card(D)} Max(sim(xi, yj)_{yj \in T}$$

xi, yj étant les imagettes des documents D et T respectivement ; et sim(xi, yj) est la même mesure de similarité que nous avons employée afin de comparer les imagettes pendant la phase de regroupement. Notons que, selon cette mesure, deux documents manuscrits seront d'autant plus proches que la mesure de similarité sera proche de 1. La Figure 8 décrit l'étape de décision de notre système.

6. Expérimentations et résultats

Nous avons effectué deux séries de tests, dans la première, un échantillon d'écriture est représenté par l'ensemble des imagettes en lesquelles se décompose l'écriture d'un scripteur (cf. 7.2), et dans la deuxième série nous exploiterons la stabilité de chaque écriture en caractérisant le scripteur par ses formes invariantes (cf. 7.3).

6.1. Description de la base d'images utilisées

Pour l'évaluation de notre approche d'identification de scripteurs, nous avons mené un certain nombre d'expérimentations en utilisant la base d'images de noms manuscrits de wilayas algériennes qui a été construite au sein du laboratoire LRI (Annaba, Algérie) et qui a fait l'objet de travaux sur la reconnaissance de mots [16]. Pour le vocabulaire considéré (48 noms de wilayas), un modèle de formulaire pré-imprimé constitué de trois pages (16 mots/page) a été conçu. chaque mot devant être recopiés trois fois par chacun des 100 scripteurs impliqués dans la construction de cette base. Les documents ont été numérisés sur 256 niveaux de gris avec une résolution de 300 points par pouce. Nous avons aléatoirement choisi les documents écrits par 30 scripteurs pour l'évaluation des performances de notre système. Pour

chaque scripteur, nous avons retenu une page d'écriture, les deux tiers ont servi pour l'apprentissage et le reste a été utilisé pour les tests

6.2. Caractérisation des scripteurs par des imagettes

Nous avons procédé à une série de tests en utilisant une caractérisation des documents de la base de tests et de la base de référence qui s'appuie sur l'ensemble des toutes les imagettes. Les résultats obtenus illustrent la pertinence des imagettes comme caractéristiques discriminantes dans l'identification de scripteurs. En effet nous sommes parvenus à identifier le bon scripteur dans 96,66% des cas présentés au système en Top 1. Les résultats obtenus sont décrits dans la Figure 9. L'inconvénient de cette représentation est qu'elle est très coûteuse en temps d'exécution.

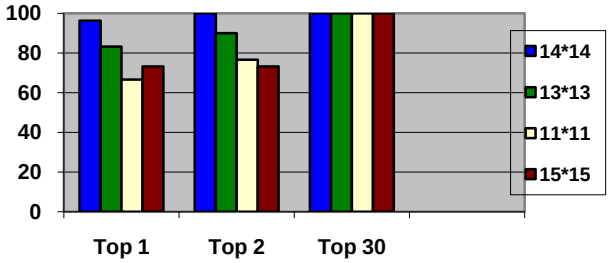

Figure 9 : Taux d'identification en utilisant des imagettes

6.3. Caractérisation des scripteurs par des invariants

Dans cette série de tests, nous nous sommes intéressés à la pertinence de la représentation des textes de référence et de test par des formes invariantes. Les résultats obtenus sont illustrés par la Figure 10. Nous sommes parvenus à identifier le bon scripteur dans près de 93.33% (Top 1) et 100% (Top 2) des cas présentés au système.

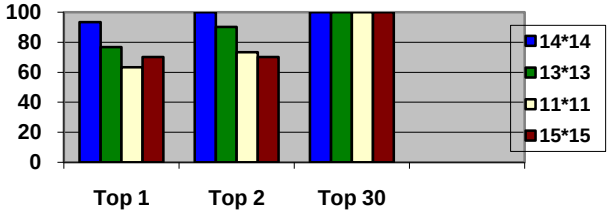

Figure10 : Taux d'identification en utilisant les invariants

Conclusion et perspectives

Nous avons présenté une méthode efficace pour l'identification de scripteur dans les documents arabes manuscrits. La méthode est basée sur l'extraction des formes invariantes que le scripteur emploierait dans son écriture. Les taux réalisés d'identification sont très prometteurs et valident l'hypothèse de l'existence de formes invariantes et propres à la main d'un scripteur au sein de son écriture arabe manuscrite.

Comme perspectives d'avenir nous comptons développer l'approche proposée afin qu'elle soit appliquée en mode indépendant du texte. Il est aussi possible d'effectuer l'intégration du module d'identification de scripteur que nous avons développé dans un système de reconnaissance hors-ligne de l'écriture arabe exploitant le principe d'adaptation à l'écriture à reconnaître.

Références

[1] Maes, P. (1994). Agents that reduce work and information overload. Communications of the ACM, 37, 30–40.

[2] Goldberg, D.E. (1989). Genetic algorithms in search, optimization andmachine learning. Addison-Wesley.

[3] Salton, Gerard (ed.). (1971). *The SMART Retrieval System – Experiments in Automatic Document Processing*. Englewood Cliffs, NJ: Prentice-Hall. 97, 131, 150

[4] A. Al-Dmour, R. Abu Zitar, "Arabic writer identification based on hybrid spectral-statistical measures", Journal of Experimental & Theoretical Artificial Intelligence, Volume 19, December 2007, pp : 307 – 332.

[5] L. M. Al-Zoubeidy et H. F Al-Najar, "Arabic writer identification for handwriting images", International Arab Conference on Information Technology, Amman, 2005, pp.111-117.

[6] A. Bensefia, "Analyse de documents manuscrits : Identification et vérification de scripteur ", Doctorat de l'université de Rouen, France, 2004

[7] M. Bulacu, L. Schomaker, A Brink, "Text-Independent Writer Identification and Verification on Offline Arabic Handwriting". ICDAR 2007, Volume 2, 2007, pp : 769-773.

[8] S. Gazzah, N. Essoukri Ben Amara. "Writer Identification using SVM Classifier and Genetic Algorithm for Optimal futures selection", International Arab Conference on Information Technology, Amman, 2005, pp. 461-466.

[9] S. Gazzah, N. Essoukri Ben Amara. “Writer Identification Using Modular MLP Classifier and Genetic Algorithm for Optimal Features Selection”, Lecture Notes in Computer Science, International Symposium on Neural Networks, China, Vol. 3972, 2006, pp. 271-276.

[10] S. Gazzah, N. Essoukri Ben Amara. “Arabic Handwriting Texture Analysis for Writer Identification Using the DWT-Lifting Scheme”. ICDAR 2007, Volume 2. Septembre 2007, pp.1133-1137.

[11] R. Huber, A. Headrick, “Handwriting Identification: Facts and Fundamentals”, 1999, CRC Press.

[12] A. Nosary, "Reconnaissance Automatique de Textes Manuscrits par Adaptation au Scripteur", Thèse de Doctorat de l'Université de Rouen, 6 Janvier 2002.

[13] M. Pechwitz, S. Snoussi Maddouri, V.Märgner, N. Ellouze, H. Amiri, “IFN/ENIT-Database of handwritten arabic words”, CIFED’2002, Colloque International Francophone sur l’Ecrit et le Document, Hammamet, Tunisie, 2002, pp. 129-136

[14] R. Plamondon, G. Lorette, "Automatic signature verification and writer identification–the state of the art". Pattern Recognition, volume 22, 1989, pp : 107 – 13.

[15] H. Said, T. Tan, K. Baker, “Personal identification based on handwriting”. Pattern Recognition, 2000, pp : 149 – 160.

[16] A. Seropian, “Analyse de Document et Identification de Scripteurs”, Doctorat de l’université de Toulon et du Var, France, 18 décembre 2003.

[17] F. Shahabi Nejad, M. Rahmati, “Comparison of Gabor-based features for writer identification of Farsi/Arabic handwriting”. In Proc. of 10th IWFHR, La Baule, France, 2006, pp : 545 - 550,

[18] F. Shahabi Nejad, M. Rahmati, "A New Method for Writer Identification and Verification Based on Farsi/Arabic Handwritten Texts", ICDAR 2007, Volume 2, Septembre 2007 pp : 829 – 833.

[19] I. A. Siddiqi, N.Vincent, “Writer Identification in Handwritten Documents”. ICDAR 2007, Volume 01, 2007, pp : 108-112.

[20] Souici-Meslati L, Sellami M., " A hybrid neuro-symbolic approach for arabic handwritten word recognition ", JACIII, Journal of Advanced Computational Intelligence and Intelligent Informatics, FujiPress, Japan, Volume 10, No. 1, 2006, pp. 17-25.

[21] S. N. Srihari. H. Arora, S. H. Cha and Sangjik Lee, "Individuality of handwriting" Journal of Forensic Sciences, Volume 47, no. 4, 2002, pp. 1 - 17.