Proposition d'une approche hybride pour le tri postal multilingue

Toufik Sari, Mokhtar Sellami

Proposition d'une approche hybride pour le tri postal multilingue

T. Sari – M. Sellami

Laboratoire LRI

Université Badji Mokhtar – Annaba

BP-12, 23000 Annaba Algérie

{sari,sellami}@lri-annaba.net

Résumé : *Dans cet article, nous proposons une solution globale au problème de tri du courrier postal dans les pays où coexiste l'usage simultané de l'arabe et d'autres langues d'origine latine (français, anglais, …). L'étude effectuée nous a permis de développer une architecture générale du système. Nous avons également commencé à implémenter certains nouveaux modules et à intégrer ceux déjà existants. Les modules d'identification du script de l'adresse et la localisation des informations pertinentes sont de nouveaux modules que nous avons développés. Dans la version actuelle du système, on s'intéresse aux noms de villes et pays. L'identification du script est nécessaire, le courrier interne, externe entrant ou sortant peut être rédigé en arabe, en latin ou bien avec les deux simultanément.*

Mots-clés *: Tri automatique du courrier, Identification du script, Segmentation explicite, Réseaux de neurones, Méthodes structurelles.*

1 Introduction

La lecture d'adresses postales est considérée comme l'une des applications des plus convaincantes de la reconnaissance de formes. Il s'agit de systèmes hors-ligne de type omni-scripteur à vocabulaire de grande taille (plusieurs milliers de mots). L'existence d'une redondance importante d'informations entre le nom de ville et le code postal peut être exploitée afin d'améliorer les taux de performances. Une grande quantité de courrier est traitée chaque jour par les services postaux dans le monde. A titre d'exemple, en 1997 le service postal américain a dû traiter environ 630 millions de lettres par jour [SRI92, SRI99]. Avec cette grande quantité de courrier, l'utilisation de systèmes automatiques de tri de courrier est donc primordiale. Des systèmes capables de reconnaître des caractères isolés sont déjà installés dans de nombreux bureaux de poste dans le monde, faisant partie des machines de tri de courrier. Une fois que le système a identifié le code postal sur l'enveloppe, le routage automatique du courrier peut être effectué. Des systèmes plus sophistiqués sont même capables de lire complètement l'adresse du destinataire. Les conditions requises pour cette application sont : une bonne vitesse (10 lettres/s), une bonne capacité (70 millions de lettres/jour) et un taux d'erreur réduit (moins de 1% pour l'adresse). Le texte à extraire peut comprendre : le code postal, le nom de ville, le numéro et le nom de rue, la boîte postale, le code cedex etc. Ces données peuvent comprendre soit du texte imprimé, soit du texte manuscrit isolé, soit du texte cursif. L'architecture fonctionnelle d'un système du tri du courrier est comporte: un module d'acquisition de l'image, un module de localisation du champ adresse, un module de lecture du texte et enfin un module de détermination de la destination du courrier.

La localisation du champ adresse englobe, [SRI93, BER94, BEN00, BEN96, ARC96, DOW94, SRI99], une étape de pré-traitements: seuillage et binarisation, segmentation en lignes et en mots [MAH99, TAK02, COH94] et une étape de segmentation-détection du numéro de voie, nom de voie, numéro d'appartement, boite postale, nom de ville, nom du pays, etc. Dans plusieurs systèmes de tri postal, on s'intéresse uniquement à un nombre limité d'informations (BP/nom de ville/nom de voies, etc.). Pour cela, il est nécessaire de localiser cette information particulière. La technique la plus usitée dans ce contexte est la recherche par mots-clés ou *Keyword-Spotting* [BER94]. Dans cette technique, un modèle est généré pour le mot cherché (mot-clé), et un autre pour tous les autres mots pouvant être présents dans le texte (mots-non-clés). Après la segmentation du texte en mots, chaque mot est aligné sur les deux modèles, et étiqueté mot-clé ou mot-non-clé selon que la probabilité d'alignement du modèle du mot-clé est respectivement supérieure ou inférieure à la probabilité d'alignement du modèle des mots-non-clés. Pour l'efficacité d'une telle technique, on a besoin d'un module de reconnaissance de mots, qui soit précis et robuste à la présence de bruit dans les images et aux variations que peuvent subir les mots. On fait appel en général, à des méthodes statistiques, en l'occurrence, les modèles de Markov cachés HMM [GIL92]. La détermination du destinataire, i.e. client final ou point de distribution local, consiste à reconnaître les unités localisées telles que caractères, mots, ponctuations, etc. et à interpréter les symboles reconnus en se basant sur des bases de données. Dans le cas de l'écriture

manuscrite cursive, ou non cursive, et imprimée dégradée ou de mauvaise qualité, les informations contextuelles comme le nombre de caractères constituant le code postal, la ligne où se trouve l'information utile, aident beaucoup le module d'interprétation. Il faut distinguer le traitement de l'imprimé du manuscrit. Dans le premier cas, on peut commencer par une segmentation explicite en caractères, tandis que dans le cas du manuscrit et/ou cursive, on utilise plutôt une approche globale par mot où l'appui d'un lexique est indispensable [PLA00]. Une adresse manuscrite est généralement saisie directement sur l'enveloppe donnant une image à fond uniforme. Un algorithme de seuillage peut suffire [OTS93]. Par contre, les adresses imprimées sont rarement éditées directement sur les enveloppes. On imprime d'abord l'adresse sur papier (généralement blanc), ensuite, on le colle sur l'enveloppe. L'image acquise dans ce cas possède plusieurs niveaux de couleurs dont la segmentation ne peut se faire par un seuillage uniquement [BOC02]. Pour la reconnaissance des caractères et mots proprement dite, toutes les techniques de reconnaissance de formes peuvent être appliquées. Néanmoins les méthodes statistiques kppv, HMM, réseaux neuronaux etc. semblent être les plus exploitées. Srihari dans [SRI93], a implémenté plusieurs classifieurs : structurel, hiérarchique, neuronal et *template matching* individuellement et en combinaison pour la reconnaissance des caractères, ensuite a développé un HMM pour la reconnaissance des mots. Srihari et al. [SRI99] ont testé la technique d'entropie conditionnelle pour la reconnaissance des caractères. Dans [BER94], les modèles HMM lettres sont combinés en modèles HMM mots pour la reconnaissance des noms de voies dans les adresses postales. Arcas-luque et Dérieux [ARC99] ont développé plusieurs techniques (LRAC, ENOC, MLOC, ...) reconnaissant les caractères segmentés à partir de mots constituants les adresses. Ces modules sont ensuite combinés par un Perceptron multi-couches. Toutes les techniques citées plus haut sont des modules de systèmes commercialisés et opérationnels dans plusieurs pays dans le monde.

2 Situation au Maghreb Arabe

Dans la situation actuelle, trois problèmes importants sont posés dans l'automatisation du tri postal Maghrébin: i) les enveloppes et colis postaux vendus sur le marché ne respectent pas des normes précises, ii) le format des adresses n'est pas encore standardisé et iii) l'existence d'adresses en arabe et en latin en plus de celles écrites en combinant des deux. La Fig.1 montre un exemple d'adresse non standard.

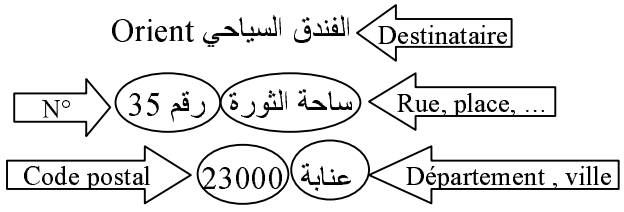

Fig.1 – Exemple d'adresse non standard en deux langues.

Pour le premier problème, la poste et les fabricants d'enveloppes doivent établir des normes bien précises, type ISO, pour les enveloppes. En ci concerne les colis postaux et le format des adresses une compagne de sensibilisation pourrait inciter les clients de la poste à respecter les consignes de rédaction de leurs adresses et préparation de leurs colis. Nous nous intéresserons dans le présent travail à présenter une proposition d'architecture globale et quelques idées pour résoudre ce troisième problème. Nous proposons de saisir les adresses selon le modèle suivant voir Fig.2:

<DESTINATAIRE>
<NUMERO-VOIE><TYPE-VOIE><SEPARATEUR><NOM-RUE>/<BOITE-POSTALE>
[<LOCALITE>]
<NOM-VILLE><SEPARATEUR><CODE-POSTAL>
[<PAYS>]

Fig.2 – Modèle standard d'adresse postale.

L'application de ce modèle sur l'adresse de la Fig.1 donne une adresse normalisée Fig.3:

Orient الفندق السياحي
35 ، ساحة الثورة
23000 عنابة

Fig.3 – Adresse standard

A l'heure actuelle, par exemple en Algérie, les documents administratifs, formulaires, notes de services, chèques bancaires et postaux, adresses postales ainsi que les pages Web sont rédigés dans l'une des deux langues : l'Arabe et le français imprimées ou manuscrits, et dans bien des cas avec les deux en même temps. Les adresses postales sont l'exemple typique d'une telle habitude. Ce problème est dû à deux principales raisons: i) une bonne partie de la population est francophone. ii) un nombre important de rues, monuments, places, entreprises et autres possèdent des appellations françaises ce qui pose le problème de leur translittération. Par exemple, au lieu d'écrire "*Saint Augustin*[1]", on traduit d'abord le mot "*saint*" par "القديس" et on translittère ensuite le nom propre "*Augustin*" par "أوغستين" ce qui donne en arabe :"القديس أوغستين " . Mais on peut avoir également une adresse dans laquelle toute la chaîne est écrite en arabe et uniquement les noms propres en lettres latines. Afin d'éviter d'écrire des adresses en deux langues, les francophones préfèrent translittérer arabe-français. Au lieu par exemple d'écrire : "*Rue* الأمير عبد القادر", on écrit plutôt : "*Rue Emir Abdelkader*". L'identification du script d'un texte donné est une tâche très complexe et primordiale au traitement par OCR. Avant qu'un système OCR ne puisse opérer sur une image de texte, il doit d'abord différencier la nature du script (imprimé/manuscrit) ensuite identifier le script lui-même (arabe, latin/roman, chinois, ...). Les techniques existantes se basent essentiellement sur les propriétés des

[1]En latin *Aurelius Augustinus*, Théologien latin, Père de l'Église (Tagaste, auj. Souq-Ahras, 354 — Hippone, 430 Algérie).

scripts traités. Ces propriétés peuvent être des statistiques globales extraites à partir de blocs de textes: taux de présence de diacritiques, de concavités et convexités, jambages, hampes, nombre de lignes de base, … [KAN02, TAN98, FAN98]. D'autres techniques analysent plutôt les mots et composantes connexes: rapport entre la largeur et la hauteur, la densité en points, … [ELG01, HOC97, SPI97a]. On trouve également des techniques qui combinent les propriétés décrites ci-dessus [ELG01].

La plupart des techniques d'identification du script développées opèrent sur des textes imprimés, peu d'entre elles traitent du manuscrit [HOC97, SPI97b] exploitant un corpus de textes assez large. Elles sont très efficaces sur des paragraphes mais échouent radicalement sur des mots isolés. Dans notre cas, le problème de discrimination du script est plus complexe. On peut trouver un mot en arabe à l'intérieure d'une ligne complète en latin ou inversement ce qui rend la tâche d'identification plus ardue.

3 Architecture globale du système de tri postal

L'architecture globale de notre système est illustrée par la Fig.4:

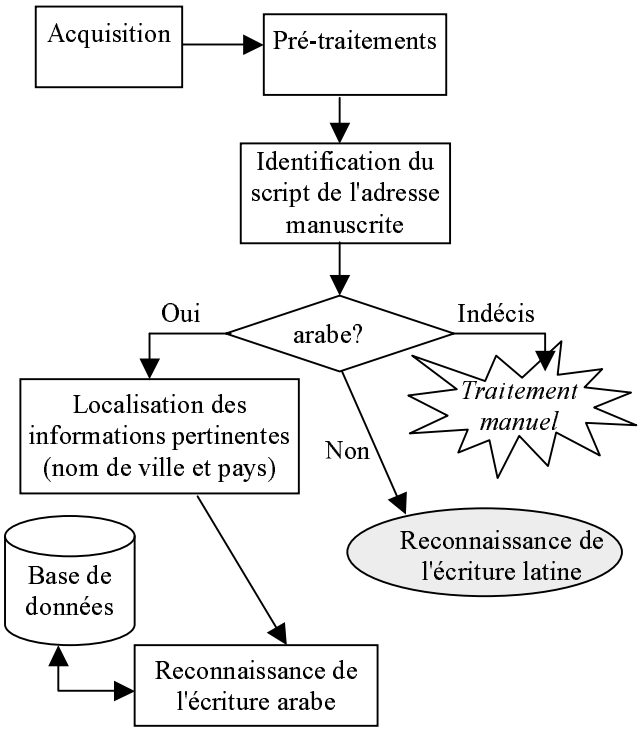

Fig.4 – Architecture globale du système de tri du

3.1 Acquisition et pré-traitements

Les adresses sont acquises en niveaux de gris à 200 DPI. Chaque image subit un filtrage de moyenne avec un élément structurant [3X3], suivi d'un seuillage local médian ce qui résulte en image binaire (0:fond et 1:écriture) [BEN96].

3.2 Identification du script de l'adresse

La zone code postal n'est pas traitée par la version actuelle du système. Plusieurs techniques efficaces existent pour la segmentation et reconnaissance des chiffres manuscrits et imprimés [PL00, LU96]. Deux approches sont possibles pour le traitement des adresses postales: la première consiste à traiter globalement la chaîne "nom ville, code postal" et effectuer la reconnaissance directement sur cette chaîne. La motivation pour une telle approche est que le champ nom-ville possède une relation de bijection avec le champ code-postal, i.e. chaque code-postal désigne un nom-ville unique et inversement un nom-ville donné possède un code-postal unique. Donc, la chaîne entière nom-ville+code-postal, en arabe ou en français, est unique. Par conséquent, si elle est traitée comme telle éviterait de devoir différencier l'écriture des chiffres et d'identifier également la langue [MAH99]. Par contre, cela implique l'utilisation d'un lexique commun plus étendu pour l'arabe et le français et appliquer des méthodes de reconnaissance très robustes exploitant des caractéristiques générales et discriminantes [BAZ99]. Ceci peut être efficace pour l'écriture imprimée [BAZ99] mais pas pour le manuscrit. La deuxième démarche que nous avons retenue consiste à identifier en premier lieu le texte ensuite effectuer la reconnaissance de la chaîne des lettres ainsi extraite.

L'écriture arabe est riche en diacritiques, à titre d'exemple à l'exception du nom de la ville algérienne (معسكر) et de l'Égypte (مصر) qui ne possèdent pas de diacritiques, tous les autres, noms de villes algériennes et noms de pays arabes, s'écrivent avec au moins une diacritique (cf. 3.4). Ces dernières peuvent venir en haut ou en bas du corps principal des lettres (ب ت ث ن …). Les textes en latin, en général, sont très pauvres en diacritiques. Nous proposons donc de calculer le nombre de diacritiques (ND) dans toute l'adresse et s'il est supérieur au tiers du nombre total de composantes connexes (NCC/3) alors le texte de l'adresse sera classé comme '*texte à tendance arabe*'. La présence de quelques mots en latin, respectivement en arabe, n'influe pas sur la suite de ce travail puisque dans une adresse en arabe, respectivement en latin, on n'écrit jamais le nom du pays ou le nom de la ville en latin, respectivement en arabe. Une composante connexe est classée comme diacritique si elle ne coupe pas la ligne de base et si elle est située en dessus ou en dessous de cette dernière [CHE98, SOU00]. L'extraction des composantes connexes est réalisée en calculant la largeur moyenne des espaces blancs entre composantes qui coupent la ligne de base dans toute l'adresse [MAH99]. Cette dernière étant supposée droite (voir Fig.5).

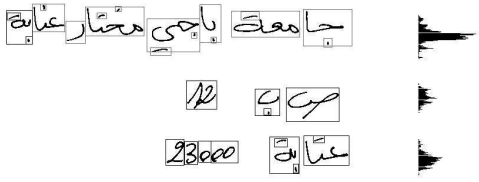

Fig.5 – Segmentation d'une adresse par analyse des espaces inter composantes.

3.3 Localisation et extraction des informations utiles

Le traitement se base sur le modèle de la figure 2, on extrait le texte de la dernière ligne qui correspond dans le cas d'un courrier destiné à l'étranger au nom d'un pays, tandis que pour le courrier interne ou international entrant contient le nom de ville et le code postal (en arabe ou en latin). Pour le courrier international sortant le texte de la dernière ligne sera ainsi acheminé vers le sous système de reconnaissance approprié. Si le texte du courrier interne ou externe entrant possède une tendance principale arabe on prend alors la chaîne se trouvant la plus à droite sinon on prend celle le plus à gauche dans la ligne correspondante.

3.4 Reconnaissance des noms des villes et noms des pays arabes

Nous proposons de combiner une vision locale avec une vision globale. Pour cela, nous utiliserons un réseau neuronal pour la reconnaissance des caractères segmentés et un classifieur structurel basé sur les primitives topologiques visuelles des mots.

3.4.1 Reconnaissance locale

On commence d'abord par isoler les diacritiques et selon les coordonnées de leurs centroïdes elles seront plus tard associées aux lettres. Ensuite, chaque mot est segmenté explicitement en lettre. Les points de segmentation sont identifiés comme étant les minima locaux sur les contours inférieurs dans la partie primaire [SAR02]. On segmente les mots sur les points de segmentation identifiés. Chaque segment primaire ainsi obtenu est caractérisé par C1 et C2 tel que:

C1: le rapport entre la largeur et la hauteur

C2: les deux principales directions du tracé selon le code de Freeman.

Nous utiliserons uniquement la caractéristique C1 pour les diacritiques. Deux réseaux neuronaux à interconnexion totale avec une seule couche cachée sont entraînés avec la fonction sigmoïde afin d'identifier respectivement les diacritiques et les segments primaires.

3.4.2 Reconnaissance globale

Ce module traite les mots en entier sans segmentation. Chaque mot est représenté par un ensemble de primitives topologiques visuelles. Nous avons retenu les caractéristiques suivantes : Nombre de sous-mots (NSM), Boucles (B), Hampes (H), Jambage (J), Diacritiques Hautes (DH), Diacritiques Basses (DB) (voir Fig.6).

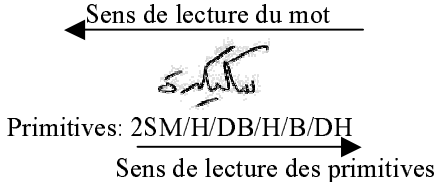

Fig.6 – Un mot arabe avec six primitives

A la différence de certains travaux [SOU00, CHE98], les primitives visuelles sont en nombre réduit et ne sont considérées dans l'ordre de leur apparition qu'en cas d'ambiguïté. Ces primitives sont traitées selon leur probabilité d'occurrence dans les lettres arabes. Pour cela, les mots du lexique (noms de villes et noms de pays) sont groupés en ordre décroissant du nombre de sous-mots. Les primitives choisies sont ordonnées ainsi[2]: DB, H, B, DH, J la plus à gauche étant la moins fréquente. Les mots à l'intérieur de chaque groupe sont ordonnés selon la présence des primitives (voir Fig.7).

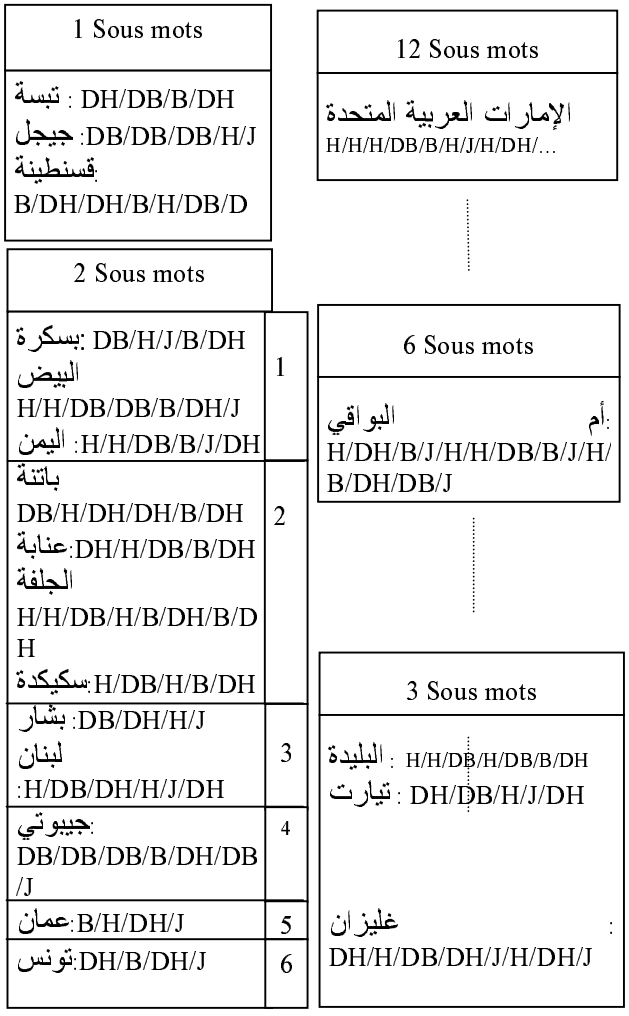

Fig.7 – Organisation du lexique

Si la sous-classe comporte plus d'un mot, on vérifie l'ordre des primitives afin d'identifier le mot exact. Pour l'exemple de la figure 6, il suffit de constater la différence de la deuxième primitive (H) par rapport au mot (الجلفة :H/H/DB/H/B/DH/B/DH). Donc, le mot en entrée est le mot (سكيكدة :H/DB/H/B/DH).

Les lettres candidates du module basé sur la vision locale sont concaténées pour former le mot. Ce dernier est comparé au mot candidat du module basé sur la vison

[2] 4 lettres arabes de base possèdent des DB, 5 H, 9 B, 13 DH, 17 J.

globale. Si les deux sous systèmes identifient le même mot, ce dernier sera retenu. Dans le cas d'un désaccord, l'adresse est rejetée.

4 Conclusion

Le module de segmentation des mots arabes a été déjà implémenté et testé sur de petites bases de données et les résultats sont prometteurs. Les modules de reconnaissance basés sur la vision locale et globale sont en cours d'implémentation et de validation. Les résultats préliminaires ne sont valables que sur des bases de données importantes et issues de courrier réel. Une convention avec Algérie Poste est en cours pour l'acquisition des images de courrier. Une fois la base des images construite et étiquetée, les tests approfondis pourront se faire afin de valider les performances des techniques proposées.

5 Bibliographie

[ARC96] Arcas-Luque G., Derieux F., Application postale de l reconnaissance du caractère manuscrit, *CNED'96*, 1996, p. 223-228.

[BAZ99] Bazzi I., Schwartz R., Makhoul J., An omnifont open-vocabulary OCR system for English and Arabic, *IEEE PAMI*, 1999, vol. 21, n°.6, p. 495-504.

[BEN96] Benyoub B., Une application industrielle de reconnaissance d'adresses, *CNED'96*, 1996, p. 93-100.

[BEN00] Bennasri A., Zahour A., Taconet B., Arabic script preprocessing and application to postal addresses, *ACIDCA'2000, Vision & Pattern Recognition*, Tunisia 2000, p. 74-79.

[BER94] Bertille J.M., Gilloux M., El Yacoubi A., Localisation et reconnaissance conjointes de noms de voies dans les lignes distribution des adresses postales, *SRTP/RD/Traitement automatique ligne distribution, 2ème semestre* 1994, *Transition n°7*, p. 16-25.

[BOC02] Bochnia G., Facon J., Segmentation du bloc adresse: application aux enveloppes postales complexes, *Actes CIFED'02*, 2002, p.245-254.

[CHE98] Cheriet M., Miled H., Olivier C., Visual aspect of Arabic handwriting recognition, *Vision Interface VI'98*, 1998, p. 263-270.

[COH94] Cohen E., Hull, J.J., Srihari S.N., Control structure for interpreting handwritten addresses, *IEEE PAMI*, 1994, vol. 16, n°. 10, p. 1049-1055.

[DOW90] Dowton A.C., Leedham C.G., Preprocessing and presorting of envelope imges for automatic sorting using OCR, *Pattern Recognition*, 1990, vol. 23, n°. 3/4, p. 347-362.

[ELG01] Elgammal A., Ismail M., Techniques for language Identification for Hybrid Arabic-English Document Images, *ICDAR'01*, 2001, p. 1100-1104.

[FAN98] Fan K., Wang L., Tu Y., classification of machine-printed and handwritten texts sing character block layout variance, *Pattern Recognition*, 1998, vol. 31, n°. 9, p. 1275-1284.

[GIL92] Gillies A.M., Cursive word recognition using hidden Markov models, *Proc. 5th US-Postal Service, advanced technology Conference*, 1992, p. 557-562.

[HOC99] Hochberg J., Bowers K., Cannon M., Kelly P., Script and language identification from handwritten document images, *IJDAR*, 1999, vol. 2, p. 45-52.

[KAN00] Kanoun S., Ennaji A., Lecourtier Y., Alimi A., une approche de discrimination arabe /latin, imprimé/manuscrit, *CIFED 2000*, Lyon, France, p. 121-129, Juillet 2000.

[KAN02] Kanoun S., Ennaji A., Lecourtier Y., Script and nature differentiation for arabic and latin text images, *IWFHR02*, 2002, p. 309-313.

[LU96] Lu Y., Shridar M., Character segmentation in handwritten words: an overview, *Pattern Recognition*, 1996, vol. 29, n°. 1, p. 77-96.

[MAH99] Mahadevan U., Srihari S.N., Parsing and recognition of city, state and ZIPcodes in handwritten addresses, *ICDAR'99*, 1999, p. 325-328.

[OTS79] Otsu N., A threshold selection method from gray-level histogram, *IEEE Transaction on image processing*, 1979, p.62-66.

[PLA00] Plamondon R. and Srihari S.N., On-line and off-line handwritten recognition: a comprehensive survey, *IEEE Trans. on PAMI*, 2000, vol. 22, n°. 22, p. 63-84.

[SAR02] Sari T., Souici L., Sellami M., Handwritten Arabic character segmentation system: ACSA, *Actes IWFHR02*, Canada, 2002, p. 452-457.

[SOU00] Souici L., Aoun ., Sellami M., Vers une architecture multi-classifieurs pour la reconnaissance de montants de chèques arabes, *Maghrebian Conference on Software Engineering and Artificial Inteligence MCSEAI'00*, 2000, Maroc, p. 125-133.

[SPI97a] Spitz A.L., Determination of the script and language content of document images, *IEEE PAMI*, 1997, vol. 19, n°. 3, p. 235-245.

[SPI97b] Spitz A.L., *Multi-lingual document recognition,* Handbook of Cha. Recog. and Doc. Image Anal., Eds. H. Bunke and P.S. Wang, World Scient., 1997, p. 259-284.

[SRI92] Srihari S.N., High-performance reading machines, *Proceeding IEEE*, 1992, vol. 80, n°. 7, p. 1120-1132.

[SRI93] Srihari S.N., Recognition of handwritten and machine-printed text for postal address interpretation, *Pattern recognition letters*, n° 14, 1994, p. 765-795.

[SRI99] Srihari S.N., Information theoretic analysis of postal address fields for automatic address interpretation, *ICDAR'99*, 1999, p. 309-312.

[TAK02] Takru K., Leedham C.G., Separation of touching and overlapping words in adjacent lines of handwritten text, ICPR'02, 2002, p. 496-501.

[TAN98] Tan T.N., Rotation invariant texture features and their use in automatic script identification, *IEEE Trans. on PAMI*, 1998, vol. 20, n°. 7, p. 751-756.