Base de données et compétitions - Outils de développement et d'évaluation de systèmes de reconnaissance de mots manuscrits arabes

H. El Abed, V. Märgner

Base de Données et Compétitions - Outils de Développement et d'Évaluation de Systèmes de Reconnaissance de Mots Manuscrits Arabes

Haikal El Abed – Volker Märgner

Braunschweig Technical University
Institut for Communications Technology (IfN)
Department of Signal Processing for Mobile Information Systems
Braunschweig, Germany
{elabed,v.maergner}@tu-bs.de

Résumé : *Cet article présente un aperçu sur l'évolution du domaine de recherche de la reconnaissance de l'écriture manuscrite arabe. La première partie de cet article traite l'importance et la diversité de bases de données pour les systèmes de reconnaissance de l'écriture. La deuxième partie présente les bases de données existantes de l'écriture manuscrite arabe. Dans la troisième partie nous présentons un aperçu sur les systèmes de reconnaissance de l'écriture manuscrite arabe, les deux dernières compétitions d'évaluation de ces systèmes et la planification de la prochaine compétition ICDAR 2009.*

Mots-clés : Reconnaissance de l'écriture, base de donnée IfN/ENIT, Compétition, Évaluation de systèmes de reconnaissance.

1 Introduction

Les premiers testes dans le domaine de la reconnaissance automatique des caractères étaient développés dans les années 30, ces systèmes utilisaient de simple méthodes de comparaison de formes sur des caractères de texte bien choisis. Pour ces systèmes on a essayé de développer des fonts bien simples à lire et à classifier et on a utilisé des formulaires bancaires pour les tester. Ces travaux ont permis de développer de puissantes méthodes de reconnaissance de textes dans les chèques bancaires [PAQ 93].

Les systèmes de tri de post étaient début des années 60 les premiers systèmes à lire automatiquement les adresses imprimées sur les enveloppes. Le véritable "boom" était provoqué début des années 90 par le développement et la publication de grandes bases de données. Ces bases de données étaient accessibles aux chercheurs dans le domaine de la reconnaissance de l'écriture. Une autre raison pour cette évolution, était les compétitions annuelles des systèmes de reconnaissance de l'écriture, comme les ateliers organisés par l'institut de recherche des sciences de l'informations (ISRI), à l'université de Nevada (par exemple [RIC 96]). La troisième raison était le développement de méthodes standardisés d'évaluation et de comparaison de la qualité des systèmes de reconnaissance (National Institute of Standards and Technology (NIST) [NIS] et ISRI [RIC 96]).

L'évolution des systèmes de reconnaissance était soutenue par les recherches effectuées dans les domaines de l'analyse de documents et l'analyse des mécanismes de la production de l'écriture [BEL 78, BEL 08]. D'autre travaux au niveau de méthodes de fusion et de combinaison de différents méthodes de classification, ont permis de diminuer les taux d'erreur des systèmes de reconnaissance de l'écriture [HEU 04]. Ces techniques ont été souvent appliquées sur les écritures à base de caractères latins et surtout sur les textes français (on dénombre une quantité très importante de travaux français effectués dans ce domaine [CIF]). Pour les langues comme l'arabe, on a commencé, depuis quelques années à adopter ces méthodes.

La première section de cet article présente une description générale des bases de données pour la reconnaissance de textes, les caractéristiques des base de données synthétiques et les exigences pour une base de données pour de textes arabes. Les différentes bases de données existantes dans ce domaine sont présentées dans la deuxième section. Les compétitions de la reconnaissance de l'écriture manuscrite arabe sont exposées dans la troisième section. Dans la dernière section, nous discutons les différents points restants encore à développer dans ce domaine de recherche.

2 Base de données pour la reconnaissance de texte

Les bases de données dans le domaine de la reconnaissance de formes sont souvent classées en deux catégories. Cette classification est basée sur la nature et l'origine des données. La première catégorie représente les bases de données dite réelles, dans lesquels les données sont directement recuites de l'application réelles. Une caractéristique très importante pour ces bases données est le fait qu'on a un a faire avec un acteur "ignorant", par exemple pour la reconnaissance de l'écriture manuscrite le scripteur n'est pas au courant de la procédure de la reconnaissance au moment où il a rédiger son texte. Cette caractéristique à aussi une facette négative, pour des raisons morales et juridiques la publication de ces données, parfois personnelles, est très restreinte. Ces restrictions sont l'une des raisons pour lesquels on constate un manque de présence de tels bases de données. La deuxième catégorie est celle des données synthétiques ou

artificielles, qui sont collecter par des formulaires spécifiques et des limitations de vocabulaires ou de styles de l'écriture.

2.1 Bases de données synthétiques et artificielles

Les coûts élevés des bases de données commerciales et des phases de collections de données ont motivé plusieurs groupes de recherche à développer des méthodes de création de base de données synthétiques [KAN 99]. Ces données ont eu pour buts de faciliter les taches de reconnaissance et d'évaluation des systèmes de reconnaissance, et surtout les systèmes basés sur les méthodes statistiques, qui ont besoin d'une grande quantité de données pour atteindre de bonnes performances.

Märgner et Pechwitz [MÄR 01] ont présenté un système de génération de données synthétiques a partir d'un texte arabe imprimé. Ce système est basé sur la génération automatique de données en différents formats, ainsi que les images et les annotations correspondantes (utilisant LaTeX). Une variété de méthodes de dégradation d'images est aussi présentée.

2.2 Les caractéristiques générales des bases de données de textes arabes

Dans cette partie, nous décrivons les caractéristiques générales et les exigences des bases de données de manuscrits arabes. Ces caractéristiques doivent être souvent prises en compte dans le processus de développement des bases de données. On peut classifier ces caractéristiques en trois catégories :

- Une première caractéristique concerne la classification des données. Il peut s'agir de données liées directement à des applications industrielles ou de données synthétiques. Dans ce dernier cas, les données sont généralement collectées en utilisant des formulaires bien spécifiques, qui permettent d'accélérer et de faciliter les étapes ultérieures de la collection des données.
- Une des caractéristiques importantes des données est la présence d'annotations (Anglais : label ou Ground Truth). Ces informations sont toujours présentes dans une base de données (sinon on ne pourrait pas effectuer les phases d'apprentissage et d'évaluation d'un système de reconnaissance). Elles peuvent être classifier en plusieurs types :
  - Dans le cas d'annotation de données contenant des lignes de texte, il est important d'avoir l'information sur la position du mot dans le texte.
  - Dans le cas des bases de données composées de mots, c'est important de connaître la séquence des caractères dans le mot. Ce point est important surtout pour les données de l'écriture arabe, car les caractères prennent des formes différentes en fonction de leurs positions dans un mot.
- D'autres caractéristiques des données sont optionnelles, et peuvent enrichir une base de données et élargir ces champs d'applications. Parmi ces caractéristiques on peut lister :
  - des informations sur le scripteur, par exemple âge, profession, genre.
  - des informations sur les sources des images et des données.
  - des informations sur l'organisation générale des données, par exemple l'information sur l'appartenance d'une image à un set d'apprentissage ou un set d'évaluation.
  - des informations de qualité de l'écriture, de l'image, etc.
  - des informations sur la position des lignes de base, des ligatures, etc.

3 Bases de données de texte arabe

On présente dans cette partie des bases de données de manuscrits arabes utilisées pour des systèmes de reconnaissance de l'écriture arabe. Malheureusement, La plupart de ces bases de données ne sont plus accessibles, elles étaient développées pour un travail de recherche bien défini.

3.1 ERIM

La base de données ERIM était développée début 1995 [SCH 95] par l'institut de la recherche environnementale du Michigan. Cette base de données est composée de plus de 750 pages de textes arabes imprimés. Cette base n'est plus disponible.

3.2 Al-Isra

Les données de la base de données Al-Isra [KHA 99] étaient collectées dans l'université de Amman. Cette base de données contient une variation d'images de mots arabes (37000), de chiffres (10000 arabes et indiens), de signatures (2500) et de textes (500 paragraphes).

3.3 CENPARMI

Cette base de données est publiée par Al-Ohali et al. [ALO 00] en 2000. Elle est composée de 7000 images de chèques saoudiens. Ces chèques sont reparties en plusieurs sous-ensembles : un premier composé de 1547 montants littéraux, un deuxième contient 1547 montants numériques, un troisième contient 23325 pseudo-mots et un quatrième composé de 9865 chiffres indiens isolés.

3.4 Farsi-City

Dehghani a présenté en 2001 [DEH 01b] une base de données composée de 17000 images de 198 noms de villes iraniennes.

3.5 AHDB

Une centaine de scripteurs étaient invités à écrire des mots provenant du vocabulaire des montants numériques, et quelques lignes de textes libres. Ces données sont composées d'un lexique de 47 mots. Cette base de données était présentée par Al-Ma'adeed [ALM 02] en 2002.

3.6 IfN/ENIT

Cette base de données est développée par l'institut des technologies de communications (IfN) en coopération avec l'école nationale d'ingénieurs de Tunis (ENIT) en 2002 [PEC 02]. Cette base de données est composée de 5 sous-ensembles, contenant en totalité (en version v2.0p1e) 32492 images de noms de villes/villages tunisiennes (plus de détails

set	images	caractères	pseudo-mots
a	6537	51984	28298
b	6710	53862	29220
c	6477	52155	28391
d	6735	54166	29511
e	6033	45169	22640
Total (v2.0p1e)	32492	257336	138060
f	8671	64781	32918
s	1573	11922	6109

TAB. 1 – Les statistiques de la base de donnée IFN/ENIT (v2.0p1e) avec les sets **f** et **s** utilisés dans la compétition ICDAR 2007.

sont présentés dans le tableau 1). Ces noms sont collectés auprès de plus de 1000 scripteurs, de différents âges et professions. Ces scripteurs ont été priés de remplir des formulaires avec les noms de villes/villages tunisiennes et les codes postaux correspondants. L'annotation des données est effectuée automatiquement et la vérification est faite manuellement. Les fichiers d'annotations contiennent des informations sur l'image, le nom de ville/village en code ASCII, la séquence détaillée des caractères, la position de la ligne de base (pour le set a), le nombre de mots, caractères et pseudo-mots, et enfin des informations sur la qualité de la base de ligne et de l'écriture (un exemple de cet annotation est présente dans le tableau 2).

A cause de la variété des informations contenues dans la base de données, plusieurs équipes de recherches utilisent la base de données IfN/ENIT pour de différentes applications, par exemple la reconnaissance de mots manuscrits, la vérification des étapes de pré-traitement, comme la segmentation ou la position de la ligne de base, l'identification de scripteur, etc. Cette base de données est utilisée par plus de 54 groupes de recherches dans plus de 27 pays au monde. Cette base de données est accessible aux groupes de recherches sous www.ifnenit.com.

Image	
Annotation :	
code postal	3070
nom en ASCII	قرقنة
séquence des caractères	قB|ر|E|قB|ن|M|ت|E|
ligne de base y1,y2	77,83
qualité de la ligne de base	B1
nombre de mots	1
nombre de pseudo-mots	2
nombre de caractères	5
qualité de l'écriture	W1

TAB. 2 – Un exemple d'image de la base de données IfN/ENIT. Les symboles M, B, A, E présentent la position des caractères dans le mots (au milieu, au début, seul, ou à la fin d'un mot).

3.7 Amin

Une base de données composée de 4800 caractères arabes isolés est présentée par Adnan Amin en 2003 [AMI 03].

3.8 CEDARAB

C'est une base de données composée d'environ 20000 mots (10 scripteurs ont écris 10 pages de textes, chacune comprend entre 150 et 200 mots) publiée par Srihari et al. en 2006 [SRI 08].

3.9 Arabic-Handwriting-1.0

Fin 2007 la base de données commerciale Arabic-Handwriting-1.0 a été publiée de la part de la société Applied Media Analysis. Cette base de données comprend 5000 images, dans lesquels on trouve 200 documents arabes manuscrits, des notes, des diagrammes, des poésies, des formulaires et des chiffres arabes et indiens.

3.10 Les bases de données présentées en 2008

Fin 2008, quatre nouvelles bases de données reliées à l'écriture arabe seront présentées. Mozaffari et al. [MOZ 08] présenteront une base de données intitulée IfN/Farsi et Bidgoli et al. [BID 08] présenteront une base de données. Chacune de ces deux dernières bases de données contient des noms de villes/villages iraniennes. Kherallah et al. [KHE 08] présentent une base de données qui peut être utilisée pour la reconnaissance en-line et hors-line. Enfin Alamri et al. [ALA 08] présenteront une base de données pour le texte manuscrit arabe.

4 Compétitions

L'intérêt à la reconnaissance de l'écriture arabe et aux méthodes d'évaluation de ces systèmes de reconnaissances a évolué massivement les dernières années. Dans cette partie de l'article nous présentons un aperçu sur les systèmes de reconnaissance de l'écriture manuscrites arabe, les deux dernières compétitions pour évaluer ces systèmes et la planification de la prochaine compétition ICDAR 2009.

4.1 Les systèmes de reconnaissance de mots manuscrits arabes

L'un des premiers systèmes de reconnaissance de l'écriture manuscrite arabe était développé par Amin et al. [AMI 80] début des années 80. Les 20 années suivantes ont vu le développement de plusieurs systèmes de reconnaissance de l'écriture arabe (imprimée et manuscrite) [LOR 06]. Trois techniques de reconnaissances ont été dominantes dans ce domaine de recherche [CHE 07]. Le premier groupe de méthodes est basé sur les techniques du logique floue, utilisé pour les systèmes de reconnaissance en-ligne [ALI 97]. Le deuxième groupe est celui qui contient les variations des réseaux de neurones (NN). Le dernier groupe est présenté par les méthodes statistiques, et en particulier, les méthodes de reconnaissance utilisant les modèles de markov cachés (MMC, ou aussi Hidden Markov Models HMM).

Le point commun entre ces systèmes est qu'ils ont pris pour les tests des bases de données limitées ou "privées". Il n'y avait pas la possibilité de comparer facilement un système développé à un autre système. Un autre aspect observé

Auteur (s)	Description	Base de données	Résultats
Abuhaiba et al. [ABU 94]	Modelés floue (FCCGM)	1410 caractères	$\approx$ 99.4%
Amin [AMI 96]	NN	3000 caractères	92%
Alimi [ALI 97]	neuro-floue	100 mots	89%
Dehghani et al. [DEH 01a]	Multiple HMM	Farsi-City	71.82%
Maddouri [Sno 02]	TD-NN	70 mots, 2070 images	97%
Khorsheed [KHO 03]	Universal HMM	Documents ancients	87%
Alma'adeed et al. [ALM 04]	Multiple HMM's	AHDB	45%
Haraty [HAR 04]	NN	2132 caractères,	73%
Souici-Meslati [Sou 04]	NN	55 mots,	92%
Farah et al. [FAR 04]	ANNK-NN, fuzzy K-NN	48 mots (100 scripteurs)	96%
Safabakhsh et Adibi [SAF 05]	CD-VD-HMM	50 mots	91%
Les systèmes utilisant la base de données IfN/ENIT (* participant à la compétition ICDAR2005 (** participant à celle de 2007))			
Pechwitz et al. (ARAB-IfN)* [PEC 03, El 07]	SC-1D-HMM	apprentissage : a-c ; test : d apprentissage : a-d ; test : e	2003 : 89% 2005 : 74.69%
Jin et al. (TH-OCR)* [JIN 05]	méthodes statistiques	apprentissage : a-d ; test : e	29.62%
Touj et al.(REAM)* [TOU 05]	Planar HMMs	apprentissage : a-d ; test : e	
Hines et al. (MITRE)** [KUN 07]	VDHMM	apprentissage : a-e ; test : f	61.70%
Ball et al. (CEDAR)** [BAL 07]	HMM	apprentissage : a-e ; test : f	59.01%
Kimura et al. (MIE)** [PAL 06]		apprentissage : a-e ; test : f	83.34%
Schambach et al. (SIEMENS)** [SCH 03]	HMM	apprentissage : a-e ; test : f	87.22%
Al-Hajj et al. (UOB-ENST)*,** [ALH 06]	HMM	apprentissage : a-d ; test : e apprentissage : a-e ; test : f	2005 : 75.93% 2007 : 81.93%
Abdulkadr (ICRA)*,** [ABD 06]	NN (Two-Tier Approche)	apprentissage : a-d ; test : e apprentissage : a-e ; test : f	2005 : 65.74% 2007 : 81.47%
Menasri et al. (Paris V)** [MEN 07]	hybride HMM/NN	apprentissage : a-e ; test : f	80.18%
Benouareth et al. [BEN 08]	HMM	apprentissage : a-c ; test : d	89.08%
Zavorin et al. (CACI)** [ZAV 08]	HMM	apprentissage : abce ; test : d	52%
Dreuw et al. [DRE 08]	HMM	apprentissage : a-d ; test : e	80.95%

TAB. 3 – Un aperçu sur les systèmes de reconnaissance de l'écriture manuscrite arabe.

sur ces systèmes est qu'ils, des taux de reconnaissances en voisinage de 100% (la partie supérieure du tableau 3).

4.2 Compétition ICDAR 2005

La première compétition pour les systèmes de reconnaissance du manuscrit arabe a été basée sur les images de la base de données IfN/ENIT, et les résultats de cette compétition ont était présentes à la conférence internationale de l'analyse et reconnaissance du document (ICDAR) 2005 [MÄR 05]. Les systèmes qui ont participé, ont été développés en utilisant la base de données IfN/ENIT et les cinq participants ont envoyé leurs systèmes pour être testés au sein du laboratoire IfN.

Dans le tableau 3, les systèmes participants à la première compétition sont présentés, avec les méthodes de reconnaissance utilisées et les taux de reconnaissance pour le set **e** (un set de test inconnu pour tous les participants, actuellement intégré dans la version v2.0p2e de la base de donnée IfN/ENIT). Cette première compétition a était une motivation pour nous et pour les participants pour travaille à améliorer les systèmes de reconnaissances de manuscrits arabes.

4.3 Compétition ICDAR 2007

La deuxième compétition a été organisée avec le même principe que celui de la première. Deux changements ont été faits, le premier concerne les sets d'apprentissage et de test et le deuxième concerne les méthodes d'évaluation des systèmes. Dans cette compétition, on a demandé aux participants d'utiliser les sets **a**,**b**,**c**,**d** et **e** pour l'apprentissage et on a utilisé les sets **f** et **s** pour les tests (ces deux sets ont été crées en 2007 et ils ne sont pas encore publiés). Ces tests sont été composés du test principal effectué sur set f, du test de robustesse effectué sur set s, et des tests de vitesse effectués sur des sous-ensembles des différents sets d'apprentissage.

Les résultats ont été présentés à la conférence internationale de l'analyse et reconnaissance du document (ICDAR) 2007 [MÄR 07]. Cette compétition a comparé 14 systèmes soumis par 9 groupes (quelques groupes ont soumis plus qu'un système). Pour la première fois, on a constaté l'intérêt de groupes de recherches industrieux, présentant 43% des participants. Cette concurrence a montré une amélioration considérable des taux de reconnaissance, qui ont atteint les 87% (les détails des résultats sont présentés dans le tableau 3), comparé avec les taux obtenus lors de la première compétition, qui était de 76%.

4.4 Compétition ICDAR 2009

Les deux premières compétitions avaient motivé plusieurs groupes de recherches travaillant dans le domaine de reconnaissance de formes en général. Des groupes qui ont développé des systèmes de reconnaissance de son ou de reconnaissance de formes, ont adapté leurs systèmes à la reconnaissance de l'écriture manuscrite arabe pour participer à la compétition et montrer que leurs techniques de reconnaissance sont universelles et surtout très performantes. Pour ces raisons nous envisageons d'organiser une compétition pour l'évaluation des systèmes de reconnaissance de l'écri-

ture manuscrite arabe lors de l'ICDAR 2009.

5 Conclusion et perspectives

Dans cet article, nous avons proposé un aperçu sur les bases de données de mots arabes et les systèmes de reconnaissance de l'écriture arabe. Ces bases de données présentent un outils indispensable dans le processus de développement de tels systèmes de reconnaissance. C'est surtout dans les phases d'apprentissage et de test qu'on a besoin de travailler avec des bases de données standardisées . L'organisation de compétitions régulièrement permet aux chercheurs d'évaluer leurs systèmes et les motive à améliorer continuellement leurs techniques de reconnaissance.

Bien d'autres aspects liés directement aux systèmes de reconnaissance de l'écriture n'ont pas été abordés faute de place. C'est le cas des différentes phases de l'acquisition d'image, de pré-traitement, des techniques de segmentation, de définition et d'extraction de caractéristiques pour la classification et des méthodes de post-traitement qui permettent souvent d'améliorer les résultats de reconnaissance.

Le domaine de la reconnaissance de l'écriture en générale et de l'écriture manuscrite arabe en particulier, est un domaine de recherche dynamique qui s'étend rapidement à de nouvelles disciplines. Mais aussi un domaine où il reste encore plusieurs points à développer concernant :

- Le développement de systèmes de reconnaissance de l'écriture manuscrite arabe sans l'utilisation explicite de lexique. Cette étape va élargir énormément le domaine d'application des tels systèmes et va ouvrir de nouveaux horizons dans le domaine de la recherche de la reconnaissance de l'écriture.
- L'intégration de modules du post-traitement, par exemple l'utilisation des connaissances linguistiques ou des informations spécifiques à l'écriture manuscrite arabe (par exemple la présence de certaine types de ligatures dans l'écriture arabe orientale).
- Le développement de bases de données réelles.
- Le défi du développement de systèmes de reconnaissance de l'écriture manuscrite arabe puissants (avec lesquels on peut participer et gagner à la compétition ICDAR 2009).

Références

[ABD 06] ABDULKADR A., Two-Tier Approach for Arabic Offline Handwriting Recognition, *10th International Workshop on Frontiers in Handwriting Recognition (IWFHR)*, pp. 161–166, 2006.

[ABU 94] ABUHAIBA I., MAHMOUD S., GREEN R., Recognition of handwritten cursive Arabic characters, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 16, n° 6, pp. 664–672, 1994.

[ALA 08] ALAMRI H., SADRI J., SUEN C., NOBILE N., A Novel Comprehensive Database for Arabic Off-Line Handwriting Recognition, *11th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, 2008.

[ALH 06] AL-HAJJ R., MOKBEL C., LIKFORMAN-SULEM L., Reconnaissance de l'écriture Arabe cursive : Combinaison de classifieurs MMCs á fenêtres orientées, *Colloque International Francophone sur l'Ecrit et le Document (CIFED)*, pp. 271-276, 2006.

[ALI 97] ALIMI A. M., An evolutionary neuro-fuzzy approach to recognize on-line Arabic handwriting, *4th Intern.l Conference on Document Analysis and Recognition*, pp. 382–386, 1997.

[ALM 02] AL-MA'ADEED S., ELLIMAN D., HIGGINS C., A data base for Arabic handwritten text recognition research, *8th International Workshop on Frontiers in Handwriting Recognition (IWFHR)*, pp. 485–489, 2002.

[ALM 04] AL-MA'ADEED S., HIGGINS C., ELLIMAN D., Offline recognition of handwritten Arabic words using multiple hidden Markov models, *Knowledge-Based Systems*, vol. 17, n° 2–4, pp. 75–79, 2004.

[ALO 00] AL-OHALI Y., CHERIET M., Databases For Recognition Of Handwritten Arabic Cheques, *7th International Workshop on Frontiers in Handwriting Recognition (IWFHR)*, pp. 601–606, 2000.

[AMI 80] AMIN A., KACED A., .HATON J. P., .MOHR R., Handwritten Arabic Character Recognition by the IRAC system, *5th International Conference on Pattern Recognition (ICPR)*, pp. 729–731, 1980.

[AMI 96] AMIN A., AL-SADOUN H., FISCHER S., Hand-Printed Arabic Character Recognition System Using an Artificial Network, *Pattern Recognition*, vol. 29, n° 4, pp. 663–675, 1996.

[AMI 03] AMIN A., Recognition of hand-printed characters based on structural description and inductive logic programming, *Pattern Recognition Letters*, vol. 24, n° 16, pp. 3187–3196, 2003.

[BAL 07] BALL G. R., Arabic Handwriting Recognition using Machine Lerning Approaches, PhD thesis, The Faculty of Graduate School of State University of New York at Buffalo, 2007.

[BEL 78] BELAÏD A., Segmentation de tracés en vue de leur analyse. Application à la reconnaissance de caractères manuscrits, *congrès francophone AFRIF-AFIA, Reconnaissance des Formes et Intelligence Artificielle*, 1978.

[BEL 08] BELAÏD A., CHOISY C., *Arabic and Chinese Handwriting Recognition*, Chapitre Human reading based strategies for off-line Arabic word recognition, pp. 36–56, Lecture Notes in Computer Science, 2008.

[BEN 08] BENOUARETH A., ENNAJI A., SELLAMI M., Arabic Handwritten Word Recognition Using HMMs with Explicit State Duration, *Journal on Advances in Signal Processing*, , 2008.

[BID 08] BIDGOLI A. M., SARHADI M., Handwritten City Names, a very large database of handwritten Persian word, *11th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, 2008.

[CHE 07] CHERIET M., KHARMA N., LIU C.-L., SUEN C., *Character Recognition Systems : A Guide for Students and Practitioners*, Wiley-Interscience, 2007.

[CIF] *Colloque International Francophone sur l'Ecrit et le Document (CIFED).*

[DEH 01a] DEHGHANI A., SHABANI F., NAVA P., Off-Line Recognition of Isolated Persian Handwritten Characters Using Multiple Hidden Markov Models, *International Conference on Information Technology : Coding and Computing (ITCC)*, 2001.

[DEH 01b] DEHGHANI M., FAEZ K., AHMADIAND M., SHRIDHAR M., Handwritten Farsi (Arabic) word recognition : a holistic approach using discrete HMM, *Pattern Recognition*, vol. 34, n° 5, pp. 1057–1065, 2001.

[DRE 08] DREUW P., JONAS S., NEY H., White-Space Models for Offline Arabic Handwriting Recognition, *19th International Conference on Pattern Recognition (ICPR)*, 2008.

[El 07] EL ABED H., MÄRGNER V., Comparison of Different Preprocessing and Feature Extraction Methods for Offline Recognition of Handwritten Arabic Words, *9th International Conference on Document Analysis and Recognition (ICDAR)*, vol. 2, pp. 974–978, 2007.

[FAR 04] FARAH N., SOUICI L., FARAH L., SELLAMI M., Arabic Words Recognition with Classifiers Combination : An Application to Literal Amounts, *Artificial Intelligence : Methodology, Systems, and Applications*, pp. 420–429, 2004.

[HAR 04] HARATY R., GHADDAR C., Arabic Text Recognition, *International Arab Journal of Information Technology (IAJIT)*, vol. 1, pp. 156–163, 2004.

[HEU 04] HEUTTE L., NOSARY A., PAQUET T., A multiple agent architecture for handwritten text recognition, *Pattern Recognition*, vol. 37, n° 665–674, page 4, 2004.

[JIN 05] JIN J., WANG H., DING X., PENG L., Printed Arabic Document Recognition System, *Proc. SPIE-IS&T Electronic Imaging*, vol. 5676, pp. 48–55, 2005.

[KAN 99] KANUNGO T., HARALICK R., An automatic closed-loop methodology for generating character groundtruth for scanned documents, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 21, n° 2, pp. 179–183, 1999.

[KHA 99] KHARMA N., AHMED M., WARD R., A new comprehensive database of handwritten Arabic words, numbers, and signatures used for OCR testing, *Canadian Conference on Electrical and Computer Engineering*, vol. 2, pp. 766–768, 1999.

[KHE 08] KHERALLAH M., ELBAATI A., EL ABED H., ALIMI A. M., The On/Off (LMCA) Dual Arabic Handwriting Database, *11th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, 2008.

[KHO 03] KHORSHEED M. S., Recognising Handwritten Arabic manuscripts using a single Hidden Markov Model, *Pattern Recognition Letters*, vol. 24, n° 14, pp. 2235–2242, 2003.

[KUN 07] KUNDU A., HINES T., PHILLIPS J., HUYCK B., VAN GUILDER L., Arabic Handwriting Recognition Using Variable Duration HMM, *9th International Conference on Document Analysis and Recognition (ICDAR)*, vol. 2, pp. 644 –648, 2007.

[LOR 06] LORIGO L., GOVINDARAJU V., Offline Arabic handwriting recognition : a survey, *Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, n° 5, pp. 712–724, 2006.

[MÄR 01] MÄRGNER V., PECHWITZ M., Synthetic data for Arabic OCR system development, *6th International Conference on Document Analysis and Recognition (ICDAR)*, pp. 1159–1163, 2001.

[MÄR 05] MÄRGNER V., PECHWITZ M., EL ABED H., ICDAR 2005 Arabic handwriting recognition competition, *8th International Conference on Document Analysis and Recognition (ICDAR)*, vol. 1, pp. 70–74, 2005.

[MÄR 07] MÄRGNER V., EL ABED H., ICDAR 2007 Arabic Handwriting Recognition Competition, *9th International Conference on Document Analysis and Recognition (ICDAR)*, vol. 2, pp. 1274–1278, 2007.

[MEN 07] MENASRI F., VINCENT N., AUGUSTIN E., CHERIET M., Shape-Based Alphabet for Off-line Arabic Handwriting Recognition, *9th International Conference on Document Analysis and Recognition (ICDAR)*, vol. 2, pp. 969–973, 2007.

[MOZ 08] MOZAFFARI S., EL ABED H., MÄRGNER V., FAEZ K., AMIRSHAHI A., IfN/Farsi-Database : A Database of Farsi Handwritten City Names, *11th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, 2008.

[NIS] NIST, NIST Special Databases and Software from the Image Group, www.itl.nist.gov/iaui/ vip/databases/defs/, 2006.

[PAL 06] PAL U., ROY K., KIMURA F., A Lexicon Driven Method for Unconstrained Bangla Handwritten Word Recognition, *10th International Workshop on Frontiers in Handwriting Recognition (IWFHR)*, pp. 601–606, 2006.

[PAQ 93] PAQUET T., LECOURTIER Y., Recognition of handwritten sentences using a restricted lexicon, *Pattern Recognition*, vol. 26, n° 3, pp. 391–407, 1993.

[PEC 02] PECHWITZ M., MADDOURI S. S., MÄRGNER V., ELLOUZE N., AMIRI H., IFN/ENIT-Database of Handwritten Arabic Words, *Colloque International Francophone sur l'Ecrit et le Document (CIFED)*, pp. 127-136, 2002.

[PEC 03] PECHWITZ M., MÄRGNER V., HMM based approach for handwritten Arabic word recognition using the IFN/ENIT - database, *7th International Conference on Document Analysis and Recognition (ICDAR)*, pp. 890–894, 2003.

[RIC 96] RICE S. V., JENKINS F. R., NARTKER T. A., The Fifth Annual Test of OCR Accuracy, rapport, April 1996, Information Science Research Institute, University of Nevada, Las Vegas.

[SAF 05] SAFABAKHSH R., ADIBI P., Nastaaligh Handwritten Word Recognition Using a Continuous-Density Variable-Duration HMM, *Arabian Journal for Science & Engineering*, vol. 30, n° 1B, pp. 95–118, 2005.

[SCH 95] SCHLOSSER S., ERIM Arabic Document Database, http ://documents.cfar.umd.edu/resources/database/, 1995, Environmental Research Institue of Michigan (ERIM).

[SCH 03] SCHAMBACH M.-P., Model length adaptation of an HMM based cursive word recognition system, *7th International Conference on Document Analysis and Recognition (ICDAR)*, pp. 109–113vol.1, 3-6 Aug. 2003.

[Sno 02] SNOUSSI MADDOURI S., BELAÏD A., CHOISY C., AMIRI H., Modéle perceptif neuronal á vision globale-locale pour la reconnaissance de mots manuscrits arabes, *Colloque International Francophone sur l'Ecrit et le Document*, pp. 11-20, 2002.

[Sou 04] SOUICI-MESLATI L., SELLAMI M., A Hybrid Approach for Arabic Literal Amounts Recognition, *Arabian Journal for Science and Engineering (AJSE)*, vol. 29, n° 2B, pp. 174–194, 2004.

[SRI 08] SRIHARI S. N., BALL G. R., SRINIVASAN1 H., *Arabic and Chinese Handwriting Recognition*, Chapitre Versatile Search of Scanned Arabic Handwriting, pp. 57–69, Lecture Notes in Computer Science, 2008.

[TOU 05] TOUJ S., BEN AMARA N., AMIRI H., Arabic Handwritten Words Recognition Based on a Planar Hidden Markov Model, *International Arab Journal of Information Technology*, vol. 2, n° 4, 2005.

[ZAV 08] ZAVORIN I., BOROVIKOV E., DAVIS E., BOROVIKOV A., SUMMERS K., Combining different classification approaches to improve off-line Arabic handwritten word recognition, *Proceedings of SPIE - The International Society for Optical Engineering*, vol. 6815, 2008.