Reconnaissance de l'écriture arabe multifonte à très basse résolution

Oussama Zayene*, — Fouad Slimane***

**Groupe DIVA, Dptmt d'Informatique*
Université de Fribourg [1]
Bd de Pérolles 90, 1700 Fribourg,
Suisse

***UR : SAGE, ENISo*
Université de Sousse [2]
BP 264 Sousse Erriadh, 4023 Sousse,
Tunisie

{oussama.zayene, fouad.slimane}@unifr.ch

RÉSUMÉ. *Nous proposons dans ce travail une approche de reconnaissance de textes arabes imprimés hors-ligne à vocabulaire ouvert et à très basse résolution (72 dpi). La méthode est basée sur les Modèles de Markov Cachés en utilisant la boite à outils HTK. La nouveauté de notre travail est dans l'analyse de ce type de systèmes sur trois fontes de calligraphie complexe et présentant de fortes ligatures : DiwaniLetter, DecoTypeNaskh et DecoTypeThuluth. Nous proposons une extraction de caractéristiques basée sur l'usage de primitives statistiques et structurelles permettant une description robuste des différentes variabilités morphologiques des fontes considérées. Le système proposé est évalué sur la base APTI (Arabic Printed Text Image).*

ABSTRACT. *We propose in this work an approach for automatic recognition of printed Arabic text in open vocabulary mode and ultra low resolution (72 dpi). This system is based on Hidden Markov Models using the HTK toolkit. The novelty of our work is in the analysis of three complex fonts presenting strong ligatures: DiwaniLetter, DecoTypeNaskh and DecoTypeThuluth. We propose a feature extraction based on statistical and structural primitives allowing a robust description of the different morphological variability of considered fonts. The validation of the proposed approach was operated on the APTI database (Arabic Printed Text Image).*

MOTS-CLÉS: *OCR arabe, reconnaissance, MMC, très basse résolution, multi-fontes.*

KEYWORDS: *Arabic OCR, recognition, HMM, Ultra low resolution, multi-fonts.*

[1]Rolf Ingold et Jean Hennebert
[2]Najoua Essoukri Ben Amara

JJC 2012, pp. 443–448, Bordeaux, 21-23 mars 2012

1. Introduction

Les textes intégrés dans les pages web sous forme d'image sont généralement compressés et générés à très basse résolution (72 - 100 dpi) afin d'accélérer leur téléchargement. Les images obtenues en faisant des captures d'écrans sont également à très basse résolution. Malgré leur efficacité relative dans la reconnaissance des textes à haute résolution (>150 dpi), les AOCR (Arabic Optical Character Recognition) classiques présentent de piètres performances sur les textes à basse résolution. Ce phénomène a été mis en évidence lors de la compétition *Arabic Recognition Compétition* organisée à ICDAR'11 (Slimane et *al.* 2011).

Plusieurs systèmes basés sur les MMCs (Modèles de Markov Cachés) ont été développés pour la reconnaissance de texte arabe (Ben Amara et *al.* 2000, El-Hajj et *al.* 2005, Ben Amor et Ben Amara 2006, Al-Muhtaseb et *al.* 2008, Slimane et *al.* 2008). Le succès de ces modèles revient essentiellement à leur bonne capacité d'intégration du contexte et de modélisation des variabilités. Ces modèles présentent également l'avantage de fournir une segmentation implicite en caractère lors du décodage de la séquence d'observations.

Dans ce travail, nous nous intéressons plus spécifiquement à la reconnaissance de texte arabe hors-ligne à vocabulaire ouvert et à très basse résolution en tenant compte des effets d'anti-aliasing. Nous basons la reconnaissance sur les MMCs en utilisant la boite à outils HTK (Young et *al.* 2001). La nouveauté de notre travail est dans l'analyse de ce type de systèmes pour trois fontes complexes présentant de fortes ligatures et des chevauchements entre caractères : DiwaniLetter, DecoTypeNaskh et DecoTypeThuluth.

Ce papier est organisé de la façon suivante : la section 2 présente les particularités des fontes arabes. Dans la section 3, nous décrivons l'approche proposée. L'évaluation de notre système avec la base APTI est décrite dans la 4ème section. Avant de conclure, nous discutons les résultats obtenus dans la section 5.

2. Particularités des fontes arabes

L'écriture arabe varie selon les milieux et les régions, d'une extrême simplicité formelle (par exemple, la fonte *Simplified Arabic* qui ne présente pas des ligatures) à la complexité exhaustive de l'arabesque (les fontes fortement ligaturées comme le DiwaniLetter). Il existe plus de 450 fontes dont seulement quelques-unes sont couramment utilisées dans le monde arabo-musulman (Ben Amara 1999). Nous nous intéressons dans ce travail, aux trois fontes les plus complexes : DiwaniLetter, DecoTypeThuluth et DecoTypeNaskh.

La figure 1 présente un exemple de mot arabe généré en 4 fontes différentes. L'image de mot présentée dans la fonte *Simplified Arabic* ne présente ni de ligature ni de chevauchement entre les caractères alors que dans les 3 autres fontes, elles présentent différents types de ligatures et de chevauchements.

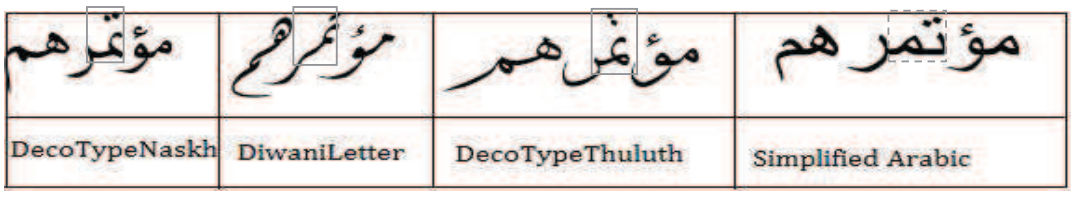

Figure 1. *Exemple d'image de mot arabe généré en 4 fontes différentes*

3. Approche proposée

3.1. *Présentation du système*

Le système proposé est basé sur les MMCs. Il a été développé en utilisant la boite à outils HTK. Comme illustré à la figure 2, il fonctionne en deux phases : l'apprentissage et la reconnaissance. Pour les deux phases, nous effectuons le même prétraitement et nous extrayons les mêmes caractéristiques.

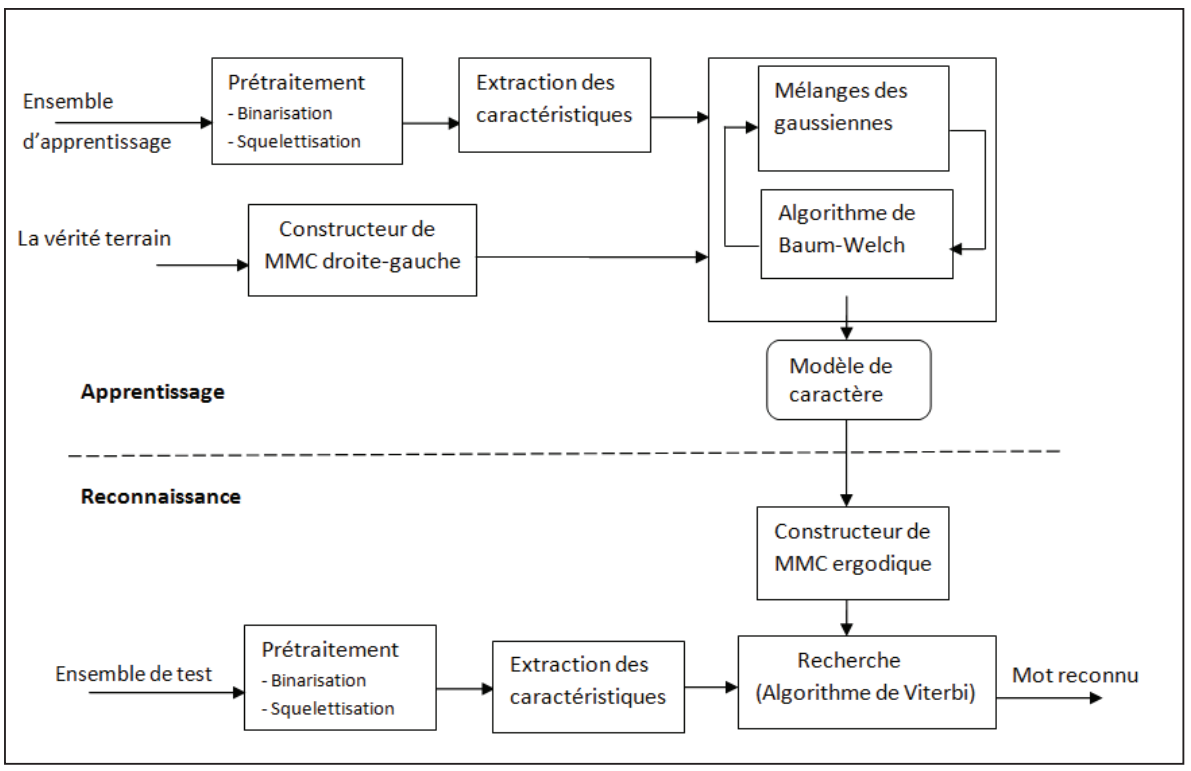

Figure 2. *Mode de fonctionnement du système de reconnaissance, basé sur le HTK*

3.2. *Prétraitement*

L'objectif de cette étape est de préparer les images de mots pour l'étape d'extraction des caractéristiques. Notre objectif est d'utiliser plusieurs caractéristiques, soit basées directement sur les images en niveau de gris, soit basées sur le contour ou le squelette de l'image comme les descripteurs de Fourier et les directions de Freeman. Les deux techniques de prétraitement que nous avons utilisé pour une partie des caractéristiques, sont la binarisation et la squelettisation.

3.3. *Extraction des caractéristiques*

Dans notre travail, nous avons utilisé le principe de la fenêtre glissante utilisé par des nombreux chercheurs (Slimane et *al.* 2009, Ben Amara et *al.* 2000,) pour la phase d'extraction des caractéristiques. L'utilisation de la fenêtre glissante se diffère d'un chercheur à un autre par le nombre de caractéristiques, leur type et la taille de la fenêtre. Dans notre cas, la largeur de la fenêtre est 6 pixels, la hauteur est égale à celle de l'image et le chevauchement entre deux fenêtres successives est égal à 1

pixel. Chaque fenêtre est divisée verticalement en N cellules, N dépendant de la fonte (pour le DiwaniLetter, par exemple, N=7).

Un ensemble des caractéristiques est extrait dans chaque fenêtre. Chaque mot est transformé donc en une matrice dont le nombre de lignes correspond à celui des fenêtres et le nombre des colonnes à celui des caractéristiques. Nous avons effectué plusieurs tests afin de trouver les caractéristiques structurelles et statistiques adéquates pour chaque fonte.

3.4. *Apprentissage et reconnaissance*

L'idée est de représenter chaque mot par un modèle de Markov caché composé par des modèles de Markov cachés associés aux caractères composant le mot. Par exemple dans la figure 3 nous avons le mot غدا (traduction : demain), à 3 caractères, chaque caractère est représenté par un MMC à 5 états en plus des deux états : S (début) et E (fin).

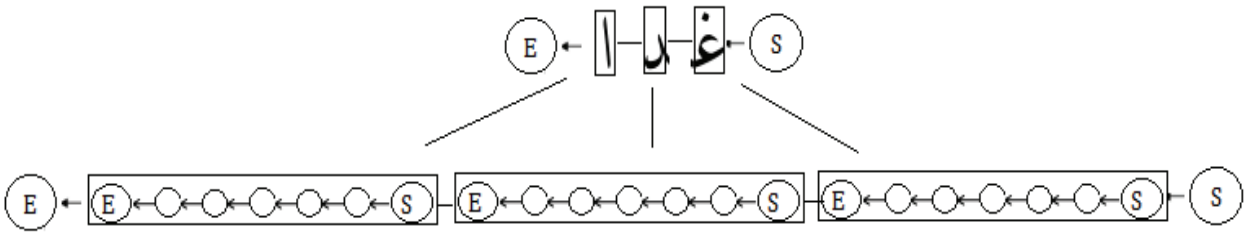

Figure 3. *Modèle de Markov Caché associé aux caractères d'un mot*

- Dans la phase d'apprentissage : nous préparons pour chaque image de mot le modèle de Markov droite-gauche qui la représente. Nous préparons aussi les fichiers de caractéristiques. Une fois les paramètres initialisés, un entraînement itératif des paramètres des modèles est effectué avec l'algorithme de Baum-Welch.

- La reconnaissance se fait en utilisant l'algorithme de Viterbi qui cherche la séquence de caractères la plus vraisemblable identifiant le mot à reconnaitre en utilisant un modèle ergodique défini par l'ensemble de modèles des caractères. Dans le modèle ergodique, toutes les transitions entre sous-modèles sont possibles, ce qui lui permet de reconnaître n'importe quel mot arabe.

4. Evaluation du système

4.1. *La base APTI- Arabic Printed Text Image*

Pour évaluer la performance de la méthode proposée, des expériences ont été réalisées sur un sous ensemble de la base APTI. A notre connaissance, APTI est la première base « publique » de données à vocabulaire large et à très basse résolution (72 dpi) pour la reconnaissance de texte arabe multi-fontes, multi-tailles et multi-style Cette base a été développée en 2009 dans le cadre d'une collaboration entre le groupe DIVA (Document, Image and Voice Analysis) de l'université de Fribourg-Suisse et le groupe REGIM (Research Group in Intelligent Machines) de l'université de Sfax-Tunisie (Slimane et *al.*, 2009). La base de données est synthétique, générée en utilisant un lexique de 113'284 mots arabes dans 10 fontes, 10 tailles et 4 styles.

4.2. *Tests et résultats expérimentaux*

4.2.1. *Système de base*

Notre système s'inspire des travaux de Slimane (Slimane et *al.* 2010). Les résultats de ce dernier pour les trois fontes choisis sont présentés dans le tableau 1. On remarque bien que les résultats obtenus sont très faibles.

	% mot	**% caractère**
DiwaniLetter	72.5 %	94.9 %
DecoTypeNaskh	64.1 %	96.0 %
DecoTypeThuluth	70,7 %	94,6 %

Tableau 1. *Résultats du système de base*

Ces taux assez bas sont dus à la morphologie relativement complexe des fontes utilisées : nos 3 fontes présentent plusieurs ligatures, des chevauchements et collisions entre les caractères et les marques diacritiques d'un seul mot, en plus des allongements horizontaux dans le cas de DecoTypeNaskh.

Pour remédier à ces problèmes, nous avons entamé une étude de pertinence des caractéristiques pour chaque fonte selon sa morphologie et ces spécificités.

4.2.2. *Sélection des caractéristiques*

Nous avons testé, avec les MMCs, un ensemble des caractéristiques telles que l'histogramme des directions de Freeman, les moments de Zernike, les descripteurs de Fourier, les 5 configurations locales, la position et le nombre de marques diacritiques, etc. Le choix des bonnes caractéristiques pour chaque fonte a été effectué en se basant sur le principe d'élimination, c'est-à-dire on ne garde que les primitives ayant un taux de reconnaissance acceptable (Zayene 2012).

	% mot	**% caractère**
DiwaniLetter	93.7%	99.1%
DecoTypeThuluth	90.3%	99.0%
DecoTypeNaskh	82.2%	99.2%

Tableau 2. *Résultats de système après le choix des bonnes caractéristiques*

Comme illustré dans le tableau 2, une amélioration de 4% du taux de reconnaissance de caractère a été enregistrée dans le cas de la fonte DiwaniLetter. Les performances sur les mots se sont améliorées de manière significative pour les trois fontes.

4.2.3. *Modèles de caractères ligaturés*

L'analyse de la matrice de confusion montre que la majorité des erreurs de suppression et d'insertion proviennent des ligatures entre les caractères. Nous avons créé pour cela des nouveaux modèles représentant les caractères ligaturés. Comme illustré dans le tableau 3, ces nouveaux modèles permettent d'éliminer une bonne partie des erreurs.

	% mot	% caractère
DiwaniLetter	98.6%	99.8%
DecoTypeThuluth	97.6%	99.7%
DecoTypeNaskh	92.4%	99.8%

Tableau 3. *Résultats de système après la création des nouveaux modèles*

5. Conclusion

Dans ce travail, nous avons présenté un système pour la reconnaissance hors-ligne de l'écriture arabe imprimée multifonte à très faible résolution. Il est basé sur les MMCs continus de type droite-gauche, en utilisant le HTK. Grâce à une sélection efficace des caractéristiques adéquates pour chacune des trois fontes et une gestion détaillée des erreurs de sortie, nous avons obtenu des résultats intéressants reflétant un système de reconnaissance robuste et performant ; ce qui représente une contribution intéressante au domaine de reconnaissance de l'écriture arabe multifonte à très basse résolution, et plus précisément au screen-based OCR.

6. Bibliographie

Al-Muhtaseb H. A., Mahmoud Sabri A., Qahwaji R. S., «Recognition of off-line printed Arabic text using Hidden Markov Models», Signal Processing, vol. 88, 2008, p. 2902-2912.

Ben Amara N., Utilisation des Modèles de Markov Cachés Planaires en reconnaissance de l'Ecriture Arabe imprimée, Thèse de doctorat, Université de Tunis, 1999.

Ben Amara N., Belaid A., Ellouze N., « Utilisation des modèles markoviens en reconnaissance de l'écriture arabe, état de l'art. », CIFED'00, Lyon, France, 2000.

Ben Amor N., Ben Amara Essoukri N., «Combining a hybrid Approach for Features Selection and Hidden Markov Models in Multifont Arabic Characters Recognition», DIAL'06, 2006, p. 103-107.

El-Hajj R., Likforman-Sulem L., Mokbel C., «Arabic Handwriting Recognition Using Baseline Dependant Features and Hidden Markov Modeling», ICDAR'05, Korea, 2005.

Slimane F., Ingold R., Kanoun S., Hennebert J., Alimi A. M., « Modèles de Markov Cachés et Modèle de Longueur pour la Reconnaissance de l'écriture Arabe à basse résolution », MajecSTIC'09, Avignon, France, 2009.

Slimane F., Ingold R., Kanoun S., Alimi A. M. Hennebert J., «A New Arabic Printed Text Image Database and Evaluation Protocols», ICDAR'09, Barcelone, Espagne, 2009.

Slimane F., Kanoun S., Alimi A. M., Hennebert J., Ingold R., «Comparison of Global and Cascading Recognition Systems Applied to Multi-font Arabic Text», DocEng '10, Actes des 10 e ACM symposium on Document engineering, New York USA, 2010.

Slimane F., Kanoun S., El Abed H., Alimi A., Ingold R., M. Hennebert J., «Arabic Recognition Competition: Multi-font Multi-size Digitally Represented Text», ICDAR'11, Pékin, Chine, 2011.

Young S., Evermann G., Kershaw D., Moore, D., Odell, J., Ollason D., valtchev V., Woodland P., The HTK Book, Cambridge University Engineering Dept, 2001.

Slimane F., Ingold R., Alimi A. M., Hennebert J., «Duration Models for Arabic Text Recognition using Hidden Markov Models», CIMCA'08, Austria, 2008, pp. 838-843.

Zayene O., Contribution à la reconnaissance de l'écriture Arabe à très basse résolution, Thèse de master, Ecole nationale d'Ingénieurs de Sousse, Université de Sousse, 2012.