

## ***Reconnaissance hybride des caractères Arabes imprimés***

***Ouarda Hachour***

Laboratoire Communication Parlée

Centre de Recherche Scientifique et Technique pour le Développement de la langue Arabe (C.R.S.T.D.L.A) Université d'Alger – Annexe de Bouzaréah - 1, rue Djamel Eddine al-Afghani – Bouzaréah tél/fax: (213) (021) 94-12-38

B.P. 225 Alger-Rostomia

16011 Alger

*E-mail:* [hachour\\_ouarda@hotmail.com](mailto:hachour_ouarda@hotmail.com)

### **Résumé**

Nous proposons un système de reconnaissance de caractères Arabes dédié à la lecture automatique des caractères en quelque soit leurs formes présentés. Dans ce présent travail, nous présentons une partie du projet développée qui a pour rôle reconnaître et faire comprendre les caractères (écrire et faire parler le contenu écrit), nous présentons une nouvelle technique de l'imagerie qui permet la caractérisation des caractères. Le système développé s'articule autour de quatre modules distincts. Un module de prétraitements, un module de segmentation, un module de reconnaissance et un module de détection de symboles de classification. Ce dernier est un classifieur flou combiné avec la technique des systèmes experts chargé d'extraire l'information topologique et contextuelle pour reconnaître les caractères. Sa sortie sera combinée avec celle du module de reconnaissance pour la reconstitution du caractère. Les résultats obtenus sur les bases de données RCA sont prometteurs.

### **Mots Clés**

Reconnaissance, Morphologique, Statistique, Logique floue, Système experts, Classification, Règles, Apprentissage, Inférence.

### **1 Introduction**

Les recherches sur la reconnaissance des caractères Arabes exposent un domaine qui s'étend rapidement et indéfiniment évoquées par une place aussi importante dans les deux dernières décennies (Kharm, 1999). C'est ainsi que la reconnaissance des caractères Arabes constitue aujourd'hui une préoccupation dont la pertinence est incontestée par la communauté de chercheurs qui ont dévoués leurs efforts à réduire les contraintes et à élargir le royaume de la reconnaissance des caractères Arabes. Ecrire pour communiquer a été de tous les temps une préoccupation première de l'Homme. L'écrit a été, et restera, l'un des grands fondements des civilisations et le mode par excellence de conservation et de transmission du savoir. Malgré les avancées d'autres moyens de communication tel que l'audio visuel, nombreuses sont les applications dont l'existence commence sur le papier, plus particulièrement dans la bureautique, en publication assistée par ordinateur (pour faciliter la composition à partir d'une sélection de plusieurs documents), dans la poste (lecture des adresses et tri automatique), dans les banques (traitement des chèques, des factures). Cependant malgré les progrès

technologiques, le clavier reste encore un moyen obligé de communication avec l'ordinateur (Cheriet, 1993), (HO, 1995), (Hassanein,1997). Nous présentons dans ce travail une partie d'un système global de reconnaissance des caractères Arabes en cours de développement dans notre laboratoire, dont la tâche est la reconnaissance des caractères Arabes imprimés en quelques soit leurs formes, police de caractère, la taille et qui comprend toutes les caractéristiques de la langue Arabe. Dans ce présent travail, nous présentons une nouvelle technique d'imagerie pour reconnaître les caractères Arabes Imprimés RCI, cette technique ne tient pas en considération la taille, la fonte et la surface. L'objectif essentiel est de reconnaître la forme d'un caractère Arabe en quelque soit la forme qui' il est présenté, si on veut reconnaître la lettre « Ha » par exemple petit, grand, ou bien gras le système doit le reconnaître du moment qu'il s'agit de la lettre « Ha » et n'on pas d'un autre caractère. (voir la figure 1).

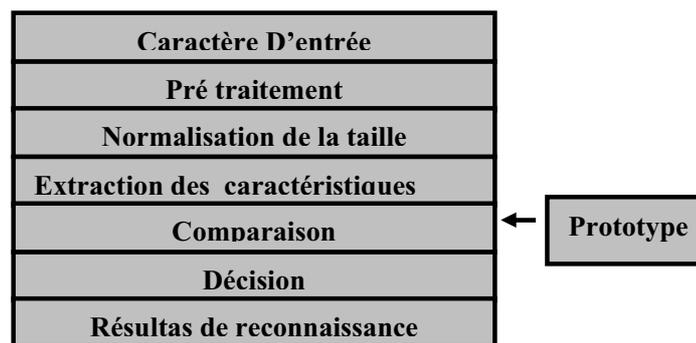


Figure 1 : le système de reconnaissance d'un caractère Arabe

## 2 Prétraitements et Segmentation

L'image extraite subit divers traitements avant son découpage et sa reconnaissance. Ainsi, pour éviter tout un problème de reconnaissance survenu durant les phases d'acquisition et de binarisation (BUNKE et al.,1995), (COCQUERESZ<sub>1</sub> et al.,1995), (COCQUERESZ<sub>2</sub> et al.,1995), (MARION, 1987), (POSTAIRE, 1987), nous procédons au processus du contour externe du tracé. Ceci permet d'éviter la fermeture d'éventuelles boucles de taille petite en gardant intactes des propriétés morphologiques importantes dans l'image. Nous passons par la suite à l'étape de segmentation. Cette dernière est une opération très importante pour le problème de reconnaissance des caractères (CHEN, 1995), elle consiste de générer un découpage de l'image en des éléments susceptibles d'être reconnus, en décortiquant par exemple le caractère en sous caractères simples à identifier (BUNKE , 1995). Dans notre cas pour la segmentation, nous avons testé la couleur noire de chaque pixel d'un caractère donné, le principe consiste de trouver un moyen de séparation entre le dernier pixel noir renvoyé d'un caractère et celui juste présenté devant lui d'un autre caractère dans un même mot en phase de reconnaissance.

## 3 Reconnaissance et classification

### 3.1 Extraction des caractéristiques

L'étude morphologique de l'arabe au niveau de l'écriture montre qu'il est difficile d'opérer la reconnaissance au niveau du caractère et ceci pour les raisons suivantes :

- Le caractère arabe peut avoir jusqu'à 4 formes différentes selon sa position dans le mot.

- Plusieurs groupes de caractères possèdent le même corps mais un nombre et/ou un emplacement de points diacritiques différents. Ces points se situent au dessus ou au dessous de la ligne de base à des endroits différents dans l'image en fonction du caractère et dans certains cas de sa position dans le mot.
- Certains caractères arabes incluent une boucle qui peut avoir différentes formes. La boucle est très souvent bouchée ou ouverte.
- Les ligatures horizontales qui se traduisent par un allongement de la ligature de base ce qui complique le processus de segmentation.
- Les ligatures verticales, bien qu'elles soient rares, posent d'importants problèmes lorsqu'elles existent. Le chevauchement vertical des caractères modifie souvent la morphologie de certains d'entre eux et la ligne de base n'est plus horizontale.

Pour pallier ces problèmes, notre système de reconnaissance de l'écriture nécessite deux étapes : une étape d'extraction de primitives et une étape de classification dans laquelle des règles de décision pour séparer les classes sont définies. Ces primitives sont généralement classées en deux familles : les primitives morphologiques et les primitives statistiques qui dérivent des mesures de distribution spatiale des pixels. Les caractéristiques morphologiques et statistiques sont complémentaires dans la mesure où deux sortes de propriétés sont mises en relief.

### **3.1.1 Primitives morphologiques et statistiques**

L'aspect morphologique est très important dans la perception humaine, et son étendue spatiale est tout aussi importante surtout dans la mesure où l'information locale n'est pas suffisante pour pouvoir représenter un motif donné. A cet effet, nous avons utilisé les caractéristiques basées sur les régions de la segmentation. Lors de l'extraction des primitives, l'image est divisée en une grille de 32 zones égales, car nous avons 8 régions et un vecteur de 4 valeurs pour chaque zone de la grille (voir la figure 2 et 4). Le nombre de pixels du contour dans chacune des zones, appartenant à chacune des 8 régions est déterminé. Les primitives statistiques utilisées sont basées sur le code en chaîne du contour des caractères. Ce dernier est calculé par un suivi en run-length de l'image. Cet ensemble de primitives est un vecteur de 12 valeurs représentant les statistiques sur les directions et les courbures des pixels dans chacune des 32 zones définies par la grille qui englobe l'image. Donc nous avons un vecteur de  $12 * 32 = 384$  valeurs. Les valeurs de courbure calculées sont quantifiées en quatre valeurs. L'ensemble des primitives est composé donc de  $(8+4+4) * 16 = 256$  valeurs, le premier nombre 4 : est le nombre de pixels de chaque côté (voir la figure 3), le deuxième nombre 4 est le nombre des valeurs de courbure calculées, le nombre 8 est le nombre de régions de zones à analyser et le nombre seize (16) est le nombre des valeurs décrites dans la figure 4. Afin de normaliser ces primitives, nous avons converti ces nombres en pourcentage par rapport aux nombres de pixels du contour dans chaque zone (voir la figure 4). Lors de la classification, nous avons utilisé une approche hybride : la logique floue et les systèmes experts pour pouvoir répertorier les caractères Arabes avec une décision logique, pertinente, et intelligente en parallèle. La base de données des caractères Arabes RCA utilisée pour la reconnaissance a été extraite à partir d'images. Elle comporte un ensemble de 116 images utilisées pour le test de résolution 300DPI.

La logique floue s'avère un outil robuste pour résoudre tout un problème imprécis ou d'autres. Les trois étapes connues comme : la fuzzification, les règles d'inférence et la défuzzification sont les étapes clés dans tout un système flou (Zadeh, 1992),(Zadeh, 1996),(Zadeh,1965), (Pedrycz, 1990), (Pedrycz,1988),(Shih-Sen,1985). Le modèle flou utilisé

lors de la conception est illustré dans la figure 5, l'algorithme à suivre pendant la phase de reconnaissance est décrit dans la figure 9. Les entrées de ce modèle sont les primitives décrites dans la figure 4. Le degré de vérité d'une entrée correspond à la fonction d'appartenance est montré dans la figure 6. Les termes linguistiques  $P_i$  flous utilisés sont : **P** : caractère petit, **M** : caractère moyen, **G** : caractère grand. L'apprentissage du classifieur se fait en deux étapes. Une étape de génération de règles et une autre d'adaptation des fonctions d'appartenance. La génération de règles s'opère selon la distribution de l'ensemble d'apprentissage en termes linguistiques flous (**P**, **M**, **G**). Les règles d'inférences utilisées lors de la conception sont illustrées dans la figure 7. La sortie de notre classifieur flou est  $N$  qui prendra les termes linguistiques flous : **PC** (petit caractère), **CM** (caractère moyen), **GC** (grand caractère). Les fonctions d'appartenance de la sortie du modèle flou sont illustrées dans la figure 7. La décision finale (defuzzification) est accomplie pour convertir les entrées en des sorties d'après les règles d'inférence. La defuzzification est calculée par la formule suivante (centre de gravité) :

$$règle. \mu^G = \frac{\sum (\mu_i * g_i)}{\sum \mu_i} \quad 1 \leq i \leq m.$$

Ou  $\left\{ \begin{array}{l} m : \text{le nombre de règle d'inférence.} \\ \mu : \text{le degré d'appartenance pour chaque} \\ g : \text{le centre de gravité de chaque règle.} \end{array} \right.$

Cette tâche intelligente utilise les termes linguistiques flous et calcule pour chacune le degrés d'appartenance sous forme d'un système expert. Un système expert est un programme soft de l'ordinateur, qui peut négocier avec des connaissances spécialisées, généralement possédées par les experts humains. Un système expert qui représente une bonne partie des activités de l'Intelligence Artificielle, est capable de dessiner des conclusions sans avoir d'autant d'informations à l'avance, et capable de diriger l'acquisition de la nouvelle information dans une manière effective (Medsker, 1995), (Mitra, 1995), (HATON, 1990). Dans notre cas, le principe de la technique consiste à vérifier pour chaque forme de caractère inconnue un ensemble de règles, ou chaque règle est de la forme : *Si <cond> alors <nom de la tâche>*, où <cond> est une combinaison de prédicats traduisant les relations spatiales entre les primitives de la forme inconnue (si la logique utilisée par le système expert est celle des prédicats).

Nombre de pixels dans la partie supérieure d'un caractère (N1)	Nombre de pixels dans la partie inférieure d'un caractère (N2)	Nombre de pixels dans la partie gauche d'un caractère (N3)	Nombre de pixels dans la partie droite d'un caractère (N4)
--	--	--	--

Figure 2 : Vecteur de primitives morphologique de zone

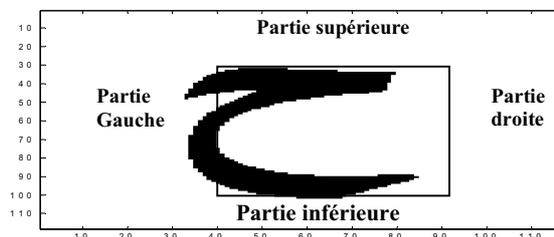


Figure 3 : exemple d'extraction des caractéristiques du caractère « ha »

### 3.1.2 Résultats

Pour l'apprentissage nous avons utilisé 116 images issues de la base de données de RCA. Nous avons utilisé une technique hybride pour pouvoir répertorier les caractères. Cette

technique est basée sur la logique floue et les systèmes experts. La combinaison de telles approches permet d'atteindre les performances d'experts humains en prenant une décision logique, intelligente et pertinente pour une classification d'un caractère Arabe donnée. Nous avons obtenus un taux d'erreurs de 1,03%, cet échec est du pour deux raisons : la première c'est que il y a toujours des similarités dans les composantes du vecteur des caractéristiques, le problème qui est souvent rencontré en phase de reconnaissance. Donc il faut trouver un moyen pour augmenter le nombre des composantes du vecteurs des caractéristiques en utilisant d'autres concepts comme les réseaux de neurones pour éviter le problème de similarité entre les caractères (apprentissage couche par couche). La deuxième raison est du aussi que si la résolution augmente plus de 300 DPI, le système marque un échec, car l'algorithme de programmation est basé sur les techniques de traitement d'images. Le taux de reconnaissance a une moyenne de 98.97%. Le tableau 1 représente les résultats de simulation pour les taux de reconnaissance de quelques caractères.

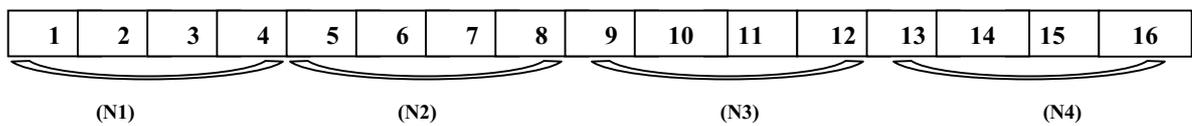


Figure 4: Format du vecteur de primitives

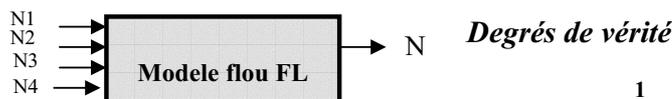


Figure 5 : Le modèle flou de la Conception

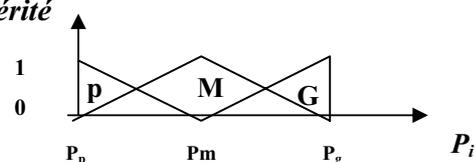


Figure 6: Repartition des fonctions d'appartenance d'une entrée du classifieur.

IF (N1 is P)	Then (N is PC)
IF (N1 is M)	Then (N is CM)
IF (N1 is G)	Then (N is GC)
IF (N2 is P)	Then (N is PC)
IF (N2 is M)	Then (N is CM)
IF (N2 is G)	Then (N is GC)
IF (N3 is P)	Then (N is PC)
IF (N3 is M)	Then (N is CM)
IF (N3 is G)	Then (N is GC)
IF (N4 is P)	Then (N is PC)
IF (N4 is M)	Then (N is CM)
IF (N4 is G)	Then (N is GC)

Figure 7: Les règles d'inférence utilisées.

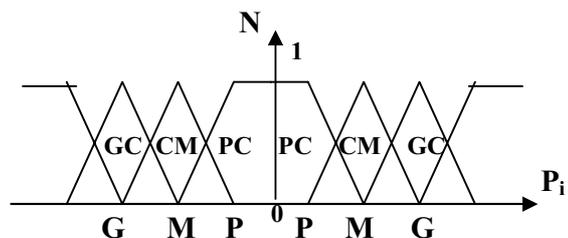


Figure 8 : les fonctions d'appartenance de la sortie du modèle flou N.

Caractère	Taux de Reconnaissance
« Alif »	98,32%
« Ba »	99,33%
« Del »	98,16%
« Ha »	98,75%
“seen”	98,83%
“cheen”	99,23%

Tableau 1 : Taux de reconnaissance .

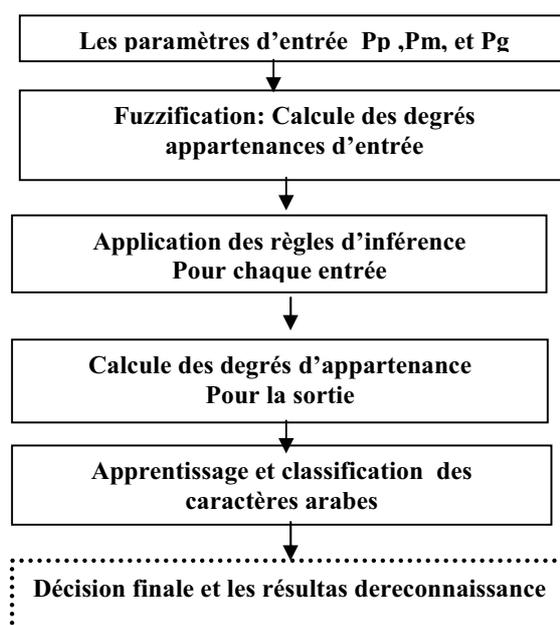


Figure 9 : L'algorithme de décision schématique

## 6 Conclusion

Nous avons proposé dans ce travail un prototype d'un système de lecture automatique de caractères Arabes imprimés. Cet article s'articulait essentiellement autour de deux parties. Une partie d'apprentissage et une partie de reconnaissance. Nous avons présenté dans la première un nouveau jeu de primitives basé sur les caractéristiques morphologiques et locales, la deuxième utilise le principe flou pour le classification et s'identification. Les résultats sont très satisfaisants au vu de la taille de la base de donnée utilisée. Pour ce faire, nous avons considéré un ensemble de primitives géométriques et topologiques propres aux chaînes de caractère et dont la pertinence a été montrée par les résultats exposés. Nous avons utilisé une nouvelle technique basée sur les techniques d'Intelligence Artificielle « *Système expert* » et sur un raisonnement flou, cette procédure de travail répond parfaitement aux problèmes de reconnaissance. La nouvelle méthodologie de conception mesure une opportunité importante dans la mesure que le système n'a marqué aucun rejet d'un caractère donné.

## Références

- CHERIET.M , SUEN.C.Y (1993), Extraction of key letters for cursive script recognition, *Pattern Recognition Letters* 14 , pp. 1009-1017.
- BUNKE.H, ROTH.M, SCHUKAT-TALAMAZZINI.E.G , Off - line cursive handwriting recognition using hidden Markov models.*Pattern Recognition*, Vol. 28, N° 9, pp 1399-1413, 1995.

## *Reconnaissance hybride des caractères Arabes imprimés*

NAWWAF N.KHARMA & RABAB K.WARD (1999) , Qu'est-ce qu'un système de reconnaissance de caractères?, *IEEE Canadian Review - Summer / Été* .

CHEN.M.Y., KUNDU.A, S. SRIHARI (1995)Variable duration hidden Markov model and morphological segmentation for handwritten word recognition, *IEEE Transactions on Image Processing*, Vol. 4, N° 12, Dec.

HO.W, LEE.S.W, KIM .J.H (1995), Modeling and recognition of cursive words with hidden Markov models.*Pattern Recognition*, Vol. 28, N° 12, pp 1945-1953.

HATON.J.P (1990), Intelligence artificielle et reconnaissance de formes,Techniques de l'Ingénieurs, Vol H2 *informatique*, pp H1900-1 - H1900-5 .

HASSANEIN.J, WESOLKOUSKI .*SCHARACTER* (1997), Extraction from Grayscale Images of Financial Documents Based on Relative Figures of Importance,*1997 International Conference on Image Processing (ICIP '97)* 3-Volume Set-Volume 3 mOctober 26 - 29, Washington, DC p. 224

BUNKE.H, ROTH.M, E.G. SCHUKAT-TALAMAZZINI (1995), *Off* - line cursive handwriting recognition using hidden Markov models.*Pattern Recognition*, Vol. 28, N° 9, pp 1399-1413.

COCQUEREZ J-P, PHILPP.S (1995), *Analyse d'image: filtrage et segmentatio*, Edition Masson.

COCQUEREZ J-P, PHILPP.S (1995), *Analyse d'image : filtrage et segmentation*, Edition Masson.

MARION.A 1987 *Introduction aux techniques de traitement d'image*, Édition Eyrolles.

POSTAIRE.J.G ( 1987 ) *Analyse des images numériques et théorie de la décision*, édition DUNOD INFORMATIQUE, PARIS.

ZADEH L.A. (1992) : The calculus of Fuzzy if/ then rules, *AI expert*, March,pp.23-27.

ZADEH L.A. (1996) : Fuzzy logic = computing with words, *IEEE Trans on Fuzzy Systems*, Vol.4, .2, May, pp.103-111.

ZADEH L.A. (1965 ) : Fuzzy sets, *Inform & Control*, ,pp.338-353.

MEDSKER L.R. (1995): hybrid intelligent systems, *Kluwer Academic Publish*.

MITRA. S.AND PAL, S.K. 1995: Fuzzy multi layer perceptron, Inferencing and rule generation. *IEEE Trans on Neural Networks*, Vol.6, N.1, , pp.51-63.

PEDRYCZ W. (1990) : Relevancy of fuzzy models, *Information Sciences* 52(, 285-302.

PEDRYCZ W. ( 1988) Fuzzy control and Fuzzy systems & sons, *Jhon Wiley*.

SHIH-SEN C. (1985) : Fixed point theorems for fuzzy mapping, *Fuzzy sets and systems*17, 181-187.