UNE METHODE RAPIDE DE SEGMENTATION ET DE RECONNAISSANCE DE CARACTERES MANUSCRITS ARABES

K. ROMEO-PAKKER , A. AMEUR

La3i - Université de Rouen - UFR des Sciences
76821 MONT - SAINT - AIGNAN Cédex
Téléphone : 35 14 65 88, Télécopie : 35 14 63 49

RÉSUMÉ

Nous décrivons une méthode structurelle de reconnaissance de l'écriture manuscrite arabe. Le problème majeur de la lecture automatique de caractères manuscrits cursifs étant la segmentation d'un tracé en éléments constituants, nous segmentons le texte d'abord en lignes, puis en mots et ensuite en caractères. Cette segmentation est fondée sur les propriétés contextuelles propres à l'écriture arabe. L'étape de la reconnaissance est étroitement liée à celle de la segmentation car elle utilise les données déjà calculées pour une pré-classification. Nous affinons la description des caractères dans un groupe plus petit par l'analyse hétérarchique qui permet le retour sur la forme pour enlever certaines ambiguïtés.

ABSTRACT

A structural method of arabic handwritten characters is proposed. The major problem in cursive text recognition is segmentation into representative strokes, our segmentation is therefore done successively into lines, then words and finally into characters. This segmentation is based on the contextual properties of arabic grammar. The recognition phase is quite dependent on the segmentation phase, the information already calculated being used for a first classification. We are following a heterarchical analysis approach which exercises goal directed feedback control on the pattern for a better discrimination of characters in the recognition process.

1. Introduction

La segmentation et la reconnaissance de textes manuscrits sont deux axes de recherche importants; il existe des produits industrialisés dont les performances doivent être améliorées. Dans ce domaine, si on se penche sur le problème de l'écriture cursive, on rencontre les difficultés de la segmentation des caractères qui s'attachent en se chevauchant, des caractères incomplètement fermés et aussi des caractères dont le tracé est séparé en plusieurs segments.

Les caractères arabes manuscrits sont un exemple typique de l'écriture cursive. Non seulement les caractères sont liés, mais ils obéissent aussi à des règles particulières: Certains caractères par exemple, s'attachent seulement à gauche et d'autres, seulement à droite... La segmentation effectuée par Almuallim [1], repose sur les traits d'écriture dans chaque mot, les noeuds et branches sont analysés pour proposer pour chaque caractère un code spécifique à comparer à un modèle. Amin, dans son dernier travail [2], segmente les mots à chaque noeud, après un prétraitement avec un algorithme d'amicissement. Le problème de l'emplacement des points est soulevé, mais reste non résolu à chaque fois. El-Dabi [3], essaie de reconnaitre les caractères en calculant différents moments, puis de les segmenter, en prenant en compte les caractères qui empiètent sur leurs voisins.

Partant de connaissances a priori sur les caractères arabes, nous avons élaboré une méthode de segmentation rapide qui ne nécessite ni squelettisation ni la recherche de primitives et de points caractéristiques comme croisements, les noeuds, etc... Les caractères (à part deux qu'on a choisi de sursegmenter) gardent leurs propriétés initiales. Notre méthode de reconnaissance diffère des travaux effectués jusqu'à maintenant par la méthode utilisée et aussi parce que nous avons choisi d'analyser les caractères par rapport à leur place dans le mot, et de prendre en compte les points et leur situation d'appartenance à certains caractères.

Dans le travail qu'on va introduire, le texte arabe manuscrit est saisi à l'aide d'une caméra CCD noir et blanc, reliée à une carte de vision permettant la visualisation et le traitement d'une image de 512 x 512 pixels. La figure 1 donne un exemple d'image utilisée.

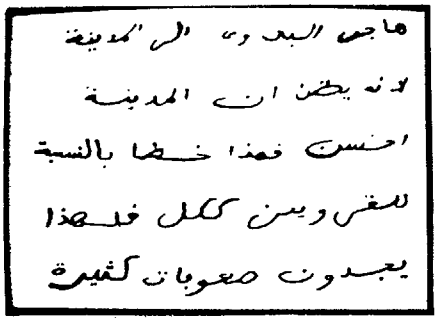


Figure 1. Image d'un texte arabe.

. 2. Caractéristiques de l'écriture arabe

L'écriture arabe s'écrit de droite à gauche. Elle est cursive, c'est-à-dire que les lettres sont liées généralement entre elles. Chaque caractère peut prendre quatre formes différentes, suivant sa position dans le mot (Fig 2). Un ensemble de pixels noirs adjacents les uns aux autres est appelé une composante connexe. Cette dernière, dans l'écriture arabe, ne représente pas forcément un mot entier, elle peut être seulement une partie du mot, car certains caractères ne doivent pas être attachés à leur successeur à gauche dans le mot. Par ailleurs, il existe des lettres différentes qui ont la même forme, mais qui se distinguent par la position et le nombre de points qui leur appartiennent. Les voyelles "a", "i" et "ou" ne sont pas utilisées systématiquement dans l'écriture arabe; des signes qui correspondent à des voyelles sont employés pour éviter des erreurs de prononciation. On peut distinguer deux types de textes : les textes avec ou sans les signes de voyelles. Quelques textes arabes (Le Coran et les livres d'apprentissage de la lecture et de l'écriture pour les enfants) contiennent des signes de voyelles. Les autres, c'est-à-dire les livres, les journaux, les publications sont des textes sans ces signes.

Nom du caractère	isolé	finale	médiane	initiale	Nom du caractère	isolé	finale	médiane	initiale
alif	ا	ا	ا	ا	dhad	ض	ض	ضـ	ضـ
ba	ب	ب	بـ	بـ	zha	ظ	ظ	ظـ	ظـ
ta	ة ت	ة ت	تـ	تـ	ain	ع	ع	عـ	عـ
tha	ث	ث	ثـ	ثـ	gain	غ	غ	غـ	غـ
jeam	ج	ج	جـ	جـ	fa	ف	ف	فـ	فـ
hha	ح	ح	حـ	حـ	qaf	ق	ق	قـ	قـ
kha	خ	خ	خـ	خـ	kaf	ك ك	ك	كـ	كـ
dal	د	د	د	د	lam	ل	ل	لـ	لـ
thal	ذ	ذ	ذ	ذ	meem	م	م	مـ	مـ
ra	ر	ر	ر	ر	noon	ن	ن	نـ	نـ
za	ز	ز	ز	ز	ha	ه	ه	هـ	هـ
seen	س	س	سـ	سـ	waw	و	و	و	و
sheen	ش	ش	شـ	شـ	hamza	ء	ء	ء	ء
tta	ط	ط	ط	ط	ya	ي	ي	يـ	يـ
sad	ص	ص	صـ	صـ					

Fig. 2 Les caractères arabes et leur position dans le mot.

3. Segmentation d'un texte arabe en caractères

La réussite de la reconnaissance de caractères dépend fortement de la bonne segmentation des lettres dans le mot. Nous pouvons donc dire que c'est une des étapes critiques pour la suite des traitements prévus pour les textes. Nous décomposons la segmentation de l'image en plusieurs étapes: D'abord la segmentation d'un texte en lignes au moyen des projections horizontales, ensuite la segmentation d'une ligne en mots en considérant les espaces entre les lettres et les mots; et enfin la segmentation des mots en caractères, beaucoup plus délicate que les deux autres. Nous avons appliqué une méthode fondée sur le trait de liaison (ou la barre de jonction) entre les caractères. Le trait de liaison représente le trait qui lie deux caractères. Son épaisseur est faible (pratiquement celle de la pointe du crayon). Nous déterminons le trait de liaison par son épaisseur (en nombre de pixels), et sa position sur la ligne de base du mot (la ligne d'écriture).

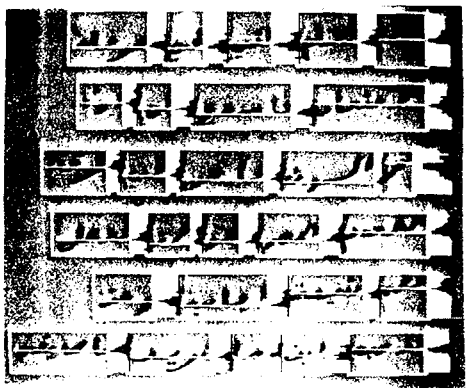


Fig. 3 Les projections (visibles à droite de chaque mot) déterminent la ligne de base.

Dans un environnement multiscripteur, les dimensions variables des tracés sont tolérées grâce à la prise en compte de valeurs relatives. Le seuil de segmentation n'est en aucun cas fixe. Notre but étant d'arriver à segmenter un mot en caractères, nous voulons détecter ces traits de liaison, plus ou moins longs suivant les écritures. Mais il est presque impossible de segmenter de façon idéale, -vu le nombre de scripteurs rencontrés et des types d'écriture utilisés- en se servant seulement de l'information fournie par des projections. Les caractères présentent également un problème de chevauchement. Deux caractères se chevauchent lorsque l'on ne peut pas encadrer un caractère dans une fenêtre correspondant à sa dimension sans croiser le caractère voisin, c'est-à-dire lorsqu'on ne peut pas faire passer une sonde verticale entre deux caractères voisins. Pour résoudre ce problème de chevauchement nous utilisons la méthode de suivi de contour en même temps qu'un étiquetage des contours détectés. Pour cela, nous avons choisi une méthode analysant les huit directions de Freeman. On effectue un balayage de haut en bas et de droite à gauche car c'est le sens de l'écriture arabe. Le suivi de contour se déplace sur les pixels blancs qui entourent les composantes connexes du mot, pour les étiqueter. On détermine ainsi le nombre de contours internes et externes du mot, les étiquettes et les dimensions des contours. L'analyse de chaque composante sera indépendante des autres, ainsi les chevauchements des caractères ne peuvent en aucun cas provoquer de confusion. Certains caractères ne peuvent se différencier que par le nombre et la position des points qui leurs sont associés. De ce fait, avant

d'identifier un caractère, il est nécessaire de déterminer ce nombre de points et leur position par rapport au tracé principal. Un point est une forme, qui n'a de liaison ni à droite ni à gauche et tel que, ni sa hauteur ni sa largeur ne dépasse trois fois l'épaisseur du trait de crayon.

Nous avons choisi de commencer la segmentation par le premier caractère à gauche du tracé, car c'est là que la grammaire arabe situe les caractères à appendices finaux généralement plus longs et plus stylisés que les autres. Lorsqu'une partie de ce caractère est au-dessous de cette ligne, nous cherchons le sens de l'ouverture de cette partie. Si l'ouverture est orientée vers l'ouest ou vers le nord, la position du début de ce caractère sera sur la colonne passant par l'extrémité droite de l'ouverture. Ainsi, le dernier caractère est séparé du reste du tracé. Dans le cas où l'ouverture est orientée vers l'est, cette partie du caractère peut englober d'autres lettres qui se trouvent à la droite du dernier caractère (Fig. 4). La branche inférieure de ce caractère va être suivie jusqu'à la fin et on autorise la segmentation du dernier caractère au-dessus de cette branche uniquement.

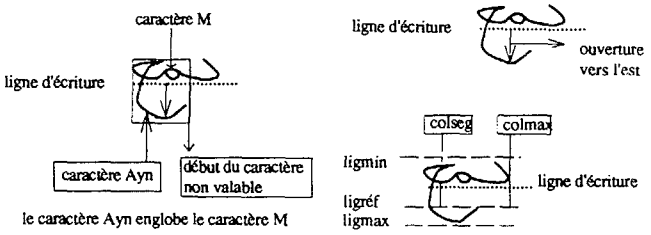
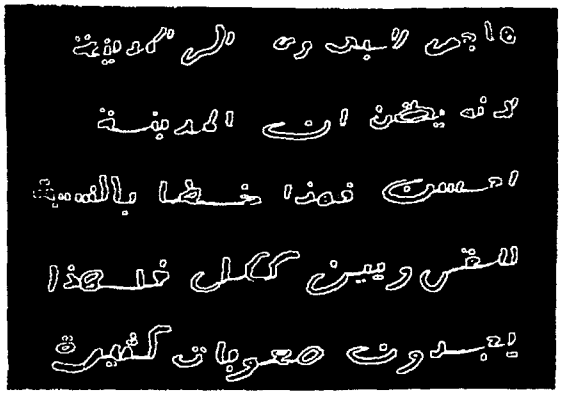


Fig. 4 Segmentation autorisée horizontalement entre colseg et colmax, et verticalement entre ligmin et ligref.

La segmentation des caractères qui précèdent le dernier caractère d'un tracé, se fait par marquage de la plus basse branche dont l'épaisseur est inférieure ou égale au seuil de segmentation défini par l'épaisseur du trait de liaison. Une correction est apportée pour des lettres comme "dal", "gain", etc... où la branche inférieure du caractère est une partie significative de la lettre.

Après la segmentation du mot en caractères, nous effectuons un réétiquetage en gardant le niveau de gris du caractère (noir) et celui du trait de liaison qui restent toujours une information pour la reconnaissance. Nous appliquons ensuite de nouveau un suivi de contour de tous les caractères du mot segmenté (Fig. 5).

Nous avons choisi de sursegmenter le caractère 'sin' en une succession de trois petits segments verticaux sans points qui seront reconnus comme tels. Nous n'avons pas essayé de joindre ces segments pour ne pas lier des couples de caractères de mêmes dimensions mais avec des points comme "ba", "ta", etc...



a) Suivi de contour appliqué sur les caractères segmentés. Les traits de liaison sont colorés en blanc.

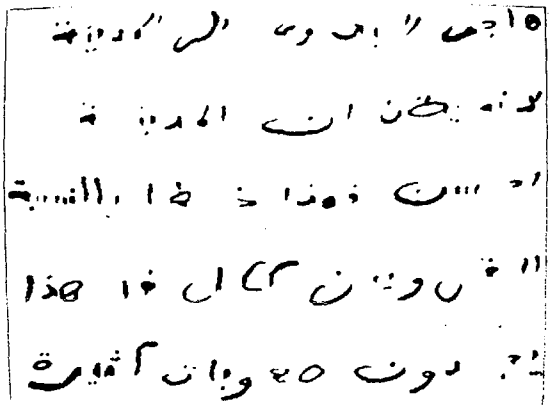


Fig. 5.b) Les caractères segmentés.

4. La Reconnaissance des caractères

La classification d'un caractère, correspond à une décision d'appartenance à une classe. Cette décision est prise après une analyse de propriétés de cette classe. On considère trois types d'analyse dans l'approche structurelle : La première, l'analyse hiérarchique descendante, essaie d'extraire les informations, en effectuant des transformations successives, jusqu'à ce qu'on arrive à une représentation finale. La deuxième, l'analyse hiérarchique ascendante, part de la description pour arriver à la vérification de l'hypothèse en appliquant les transformations en sens inverse par rapport à l'analyse descendante, des données brutes en entrée vers la représentation utilisée dans l'interprétation finale. La troisième, l'analyse hétérarchique, amène une solution plus souple au problème de la multiplicité des tests à effectuer, en contrôlant l'extraction de données (Fig. 6). L'analyse hétérarchique attache plus ou moins d'importance à un détail suivant son caractère discriminant.[4][5] Ceci nous permet de ne pas nous trouver "noyé sous un flot" d'information, qui dans beaucoup de cas ne sont pas utiles. Le contrôle des

données se fait avec des retours sur la forme initiale des caractères.

Les informations utiles pour la vérification des hypothèses émises sont ainsi recherchées, quand le besoin se présente pendant la phase d'interprétation. La stratégie utilisée pour l'analyse hétérarchique ressemble à celle que met en oeuvre le système de perception visuelle humain qui se sert des rétroactions, des "aller-retours", entre les données extraites de l'image et les données mémorisées lors des apprentissages.

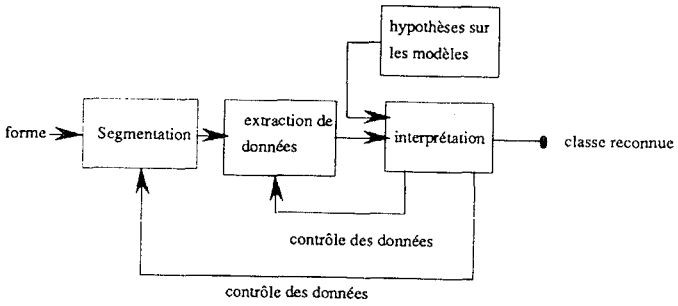


Fig. 6 L'analyse hétérarchique.

L'étape de reconnaissance est étroitement liée à celle de la segmentation car elle utilise les données déjà calculées pendant la segmentation. En nous appuyant seulement sur l'existence ou non des occlusions et sur le type de liaison avec les caractères voisins, nous pouvons séparer déjà en dix familles un ensemble représentatif de caractères arabes.

La grande difficulté dans la méthode analytique utilisée pour la reconnaissance de l'écriture manuscrite est l'augmentation du nombre de classes à reconnaître pour prendre en compte les variations du graphisme. Avec notre première classification grossière, le problème du nombre de classes change de dimension. En effet, nous pouvons analyser un nombre réduit de classes de caractères dans chaque famille et affiner ainsi la recherche sur les détails pertinents. Comme pour une même famille les caractères ont des formes ressemblantes, nous pouvons prévoir les différentes classes possibles.

Quand les positions des points sont erronées, une correction est apportée avec la connaissance a priori du nombre de points possible pour chaque caractère. Les caractères voisins sont analysés pour une meilleure répartition des points.

En appliquant la stratégie de prédiction-vérification [6] et suivant les hypothèses émises sur les caractères se trouvant dans chaque famille, des arbres de décisions sont élaborés. Le parcours en est très rapide et les cas non prévus peuvent être rajoutés dans les familles de caractères concernés.

7. Résultats Expérimentaux

Nous avons testé notre méthode sur 100 textes manuscrits écrits par 16 différents scripteurs. 8443 caractères ont été analysés. Nous obtenons 98.9 % de bonne segmentation. Certains caractères comme "ya", "ain", "gain", "sin" et "noun" ont les appendices en fin de mot séparé du caractère. Par ailleurs, "lamalif" est le caractère sursegmenté. Pour la reconnaissance, le taux obtenu est de 83 % de bonne reconnaissance, 9.7 % de rejet et 7.3 % de confusion avec d'autres caractères ressemblants. Le rejet de certains caractères est dû soit à des caractères mal écrits, ou à des formes jamais rencontrées; soit à une distribution des points sur le tracé faite d'une manière désordonnée, soit à des dimensions de points trop grandes... Nous avons également des confusions liées au fait que certains caractères ronds manuscrits semblent avoir des angles, ou bien les angles ne sont pas assez accentués par rapport à la ligne de texte. Le recours à une analyse contextuelle au niveau du mot permettra de lever la plupart de ces ambiguïtés.

8. Conclusion

Nous avons initié donc, par ce travail une méthodologie nouvelle, différente de celles des travaux effectués jusqu'à maintenant dans la reconnaissance des caractères arabes. Nous espérons que cette étude pourra contribuer à l'avancement de la recherche dans ce domaine, et que les utilisateurs de l'écriture arabe pourront bénéficier dans un proche avenir de logiciels de reconnaissance de caractères efficaces et performants.

REFERENCES BIBLIOGRAPHIQUES

[1] H. Almuallim et S. Yamaguchi.
A méthode of recognition of arabic cursive handwriting. IEEE Trans, PAMI No 5, p715-721, 1987.

[2] A.Amin et H.B. Al-Sadoun.
A new segmentation technique of arabic text. Proc. IEEE, p 441-445, 1992

[3] S. El-Dabi, R. Ramsis, A. Kamel.
Arabic Character recognition system: A statistical approach for recognizing cursive typewritten text.
Pattern Recognition, p 485-495, 1990

[4] K. Pakker, A. Chehikian.
Reconnaissance de caractères alphanumériques multipolice par analyse structurelle hétérarchique.
Congrès AFCET-INRIA RFIA, Paris, 1985

[5] H. Bollon.
Contrôle hétérarchique en vision par ordinateur. Thèse DI, INP de Grenoble, 1986

[6] G. Ménier, G. Lorette.
Segmentation et reconnaissance en ligne d'écriture cursive à l'aide de plusieurs niveaux d'information contextuelle. CNED'92, p 318-324, Juillet 1992