Reconnaissance de l'écriture arabe imprimée par transformée de Hough Généralisée

Sodien Touj, Najoua Essoukri Ben Amara, Hamid Amiri

Reconnaissance de l'Ecriture Arabe Imprimée par Transformée de Hough Généralisée

Sodien Touj [1] – Najoua Essoukri Ben Amara [2] – Hamid Amiri [3]

[1,2,3]Laboratoire des Systèmes et de Traitement du Signal-ENIT

B.P 37 Belvédère 1002, Tunisie

[1,3] Ecole Nationale d'Ingénieurs de Tunis
[2] Ecole Nationale d'Ingénieurs de Monastir

[1]Sofien.Touj@isetgb.rnu.tn – [2]Najoua.BenAmara@enim.rnu.tn

Résumé *: La reconnaissance de l'écriture Arabe imprimée reste encore un défi important à relever à cause de sa nature semi cursive et de la grande variabilité morphologique de ses différentes polices de caractères. La Transformée de Hough Standard est un algorithme classique de détection de segments de droites dans des images. Cette transformée a été généralisée pour la reconnaissance de n'importe quelle forme non paramétrique. Dans ce papier, nous proposons l'utilisation de la Transformée de Hough Généralisée pour la reconnaissance de l'écriture arabe imprimée. L'approche proposée repose sur un module de reconnaissance de caractères arabes préalablement établi permettant d'identifier et de localiser les caractères dans les pseudo-mots par une méthode de segmentation par reconnaissance.*

Mots-clés *:* Reconnaissance de l'écriture arabe imprimée, Transformation de Hough Généralisée, Segmentation par reconnaissance.

1 Introduction

L'écriture arabe est une écriture consonantique qui utilise un alphabet composé de vingt-neuf lettres. C'est une écriture plutôt semi-cursive dans le sens où le mot peut être composé d'un ou de plusieurs pseudo-mot (chaînes de caractères). Dans le pseudo-mot les lettres changent de dessin en fonction de leur position : initiale, médiane, finale ou isolée. La plupart des lettres arabes possèdent la même forme de caractère et ne se différencient que par la présence et/ou le nombre de points diacritiques se situant au-dessus ou bien au-dessous du corps du caractère ce qui réduit à 37 les formes à reconnaître (35 formes plus les caractères ﻻ et ء).

I	D	M	F	I	D	M	F	I	D	M	F
ا	-	-	ـا	ب	بـ	ـبـ	ـب	ح	حـ	ـحـ	ـح
د	-	-	ـد	ر	-	-	ـر	س	سـ	ـسـ	ـس
ص	صـ	ـصـ	ـص	ط	-	-	ـط	ع	عـ	ـعـ	ـع
ف	فـ	ـفـ	ـف	ق	قـ	ـقـ	ـق	ك	كـ	ـكـ	ـك
ل	لـ	ـلـ	ـل	م	مـ	ـمـ	ـم	ن	نـ	ـنـ	ـن
ه	هـ	ـهـ	ـه	و	-	-	ـو	ى	ىـ	ـىـ	ـى

I: Isolé, D: Début, M: Milieu, F: Fin

TAB. 1 – Différents tracés des caractères Arabes

Le problème de la reconnaissance optique de l'écriture arabe (AOCR-Arabic Optical Character Recognition) imprimée demeure à ce jour non résolu, il fût l'objet de recherches intenses depuis plus de quatre décennies [Ess 02]. La littérature montre que différentes types d'approches ont été expérimentées (structurelles, géométriques, statistiques, stochastiques…), cependant le problème reste encore ouvert [Ess 03].

La Transformée de Hough dans sa forme Standard (THS), s'est avérée souvent une technique robuste pour l'étude des alignements dans une image donnée. Jusque là, son usage s'est restreint à la détection de quelques primitives directionnelles dans l'écriture latine [Rui 00] et arabe [Fak 93]. Dans sa forme Généralisée, la THG a été aussi utilisée pour la reconnaissance des pseudo-mots arabes. Les premiers dictionnaires proposés [Zar 98] doivent inclure tous les modèles des pseudo-mots possibles ce qui contraint cette approche à des applications de reconnaissance à vocabulaire limité.

Dans ce papier, nous proposons une méthode basée sur la Transformée de Hough Généralisée (THG) pour la reconnaissance de l'écriture arabe imprimée. Elle repose sur l'identification et la localisation des caractères dans le pseudo-mot. Un module de post traitement permet de filtrer les différents caractères proposés en s'appuyant sur un ensemble de règles contextuelles.

Ce papier est organisé en cinq sections. Dans la Section 2, nous rappelons les fondements théoriques de la THG. Dans la section 3 nous proposons une approche de reconnaissance des caractères arabes basée sur la THG. Dans la section 4, nous présentons la méthode adoptée pour la reconnaissance de l'écriture arabe imprimée. Nous terminons par la conclusion.

2 Transformée de Hough Généralisée

On désigne généralement par le nom de transformées de Hough (TH) des transformations qui permettent de détecter dans des images la présence de courbes ayant une forme paramétrique (droite, conique...) à partir d'un ensemble sélectionné de points appelés points caractéristiques. La transformée de Hough utilise principalement l'information spatiale de ces points caractéristiques (leur position dans l'image) pour générer dans un espace paramétrique Ω une représentation mieux appropriée à l'utilisation. Cet espace Ω est quantifié en cellules, chacune représentant les paramètres considérés. Un compteur initialement mis à zéro est associé à chaque cellule et il est incrémenté à chaque fois que cette dernière est affectée par la TH. L'espace Ω est aussi appelé espace de Hough ou bien accumulateur de Hough ou encore tableau de vote [Mai 85].

La technique de la TH a pu être généralisée pour la détection de formes arbitraires contenues dans l'image. Les formes recherchées par la TH Généralisée (THG) ne sont pas obligatoirement définies de manière analytique mais plutôt par une silhouette particulière. Un tableau de référence nommé R-table définit donc la correspondance entre l'espace de définition de la forme recherchée (espace image) et l'espace paramétrique [Mar 01].

La THG agit donc sur des points caractéristiques de l'image, généralement le contour. Elle permet en phase d'apprentissage de décrire chaque forme par le tableau R_Table. En phase de reconnaissance, elle exploite les différents tableaux R_Tables, préalablement obtenus pendant la phase d'apprentissage, pour générer les espaces de vote qui permettent de faire la classification [Tou 03a].

Dans la section suivante, nous développons une approche de reconnaissance de caractères arabes basée sur la THG.

3 Une approche de reconnaissance de caractères arabes basée sur la THG

Dans une première phase de développement de nos travaux, nous avons exploré la THG pour la reconnaissance des caractères arabes préalablement segmentés. Ce module est à la base de l'approche de reconnaissance de l'écriture cursive imprimée arabe que nous considérons dans la section 4.

Nous décrivons dans ce qui suit les principales étapes de la méthode de reconnaissance de caractères, basée sur la THG.

3.1 Extraction des points caractéristiques

L'extraction des points caractéristiques est réalisée en appliquant un opérateur de type gradient à l'image considérée. L'orientation correspondante de la tangente θ est ensuite mémorisée. L'opérateur du gradient adopté utilise deux masques de filtrage : un masque horizontal et un masque vertical (FIG. 1).

$$\begin{bmatrix} -1 & -2 & -3 & -2 & -1 \\ 0 & 0 & 0 & 0 & 0 \\ 1 & 2 & 3 & 2 & 1 \end{bmatrix} \qquad \begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -3 & 0 & 3 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix}$$

Masque vertical Masque horizontal

FIG. 1 – Masques vertical et horizontal utilisés.

A la suite de la convolution de l'image avec ces deux masques séparément, nous obtenons deux cartes du gradient Δv et Δh. L'orientation du gradient local à chaque point (x, y) est alors donnée par l'expression suivante :

$$\theta = \arctan\left(\frac{\Delta v(x,\ y)}{\Delta h(x,\ y)}\right) \qquad [1]$$

3.2 Construction des tableaux R_Tables

Après avoir choisi le point de référence, nous construisons un tableau caractéristique appelé R_Table correspondant à chaque modèle de caractère.

Nous avons retenu comme point de référence, le centre O(xo,yo) du cadre englobant le contour du caractère. Pour chaque point caractéristique Pi(xi,yi) ayant comme orientation du gradient local θi, nous calculons l'offset Δ(xi,yi) entre ce point et le point de référence O(xo,yo) conformément à la relation [2].

$$\Delta(xi,yi) = O(xo,yo) - Pi(xi,yi) \qquad [2]$$

Le tableau R_Table sera donc formé par l'ensemble des listes des paires d'offsets ainsi obtenues pour chaque θi.

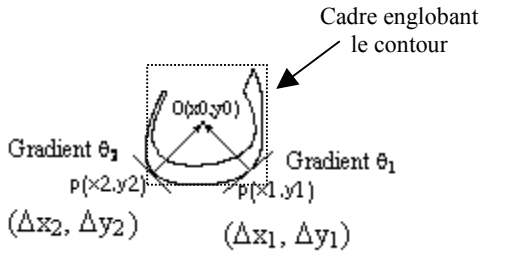


FIG. 2 – Construction du tableau R_Table.

3.3 Construction des dictionnaires

Après avoir créé l'ensemble des tableaux R_Tables relatifs aux différentes formes de base des caractères arabes imprimés, nous procédons à la construction du dictionnaire indexés et étiquetés. Ce dictionnaire

comprend, en plus des tableaux R_Tables, des informations structurelles relatives aux diacritiques (nombre et position par rapport au tracé du caractère), à la hauteur, la largeur, la densité en pixels ainsi que le décalage entre le point de référence et le niveau de la ligne d'écriture spécifiques à chaque modèle de caractère. [Tou 03b].

Modèle 0

Information structurelle + **R_Table**

boucles, diacritiques, dimensions, densité de pixels

θ en degré	Liste des paires d'offsets
0	$\Delta(x_1,y_1)$, $\Delta(x_2,y_2)$, $\Delta(x_3,y_3)$,
1	$\Delta(x_5,y_5)$, $\Delta(x_6,y_6)$
2	$\Delta(x_7,y_7)$
.	
358	null
359	$\Delta(x_n,y_n)$ …

Modèle n

Information structurelle + **R_Table**

boucles, diacritiques, dimensions, densité de pixels

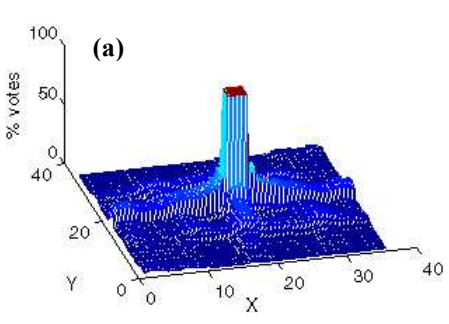
θ en degré	Liste des paires d'offsets
0	$\Delta(x_4,y_4)$, $\Delta(x_6,y_6)$,
1	null
2	$\Delta(x_1,y_1)$
.	
358	null
359	$\Delta(x_n,y_n)$ …

FIG. 3 – Structure du dictionnaire

3.4 Reconnaissance

Pour une image donnée de caractère, nous procédons d'abord à une extraction de ses points caractéristiques. Nous calculons ensuite les orientations du gradient local θi. Se servant des θi comme index dans les tableaux R_Tables, nous utilisons les offsets Δ(xi,yi) afin d'estimer la position du point de référence pour chaque modèle de caractère inclus dans le dictionnaire. Nous accumulons au fur et à mesure les résultats obtenus dans les différents tableaux de vote (ou accumulateurs) des modèles. Un processus de recherche de maximum parmi l'ensemble complet des accumulateurs permet de retenir la meilleure proposition (FIG. 4).

L'expérimentation de la méthode a été effectuée sur un jeu de 166 873 échantillons de caractères arabe dans leurs différentes formes (début, milieu, finale et isolée) dans la fonte «Arabic Transparent». Le taux moyen de reconnaissance obtenu est de 93% [Tou 03c].



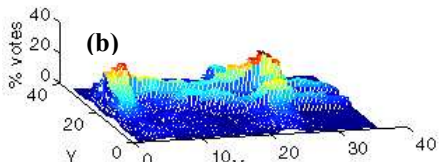


FIG. 4 – Représentation en 3D des tableaux de vote d'un échantillon du caractère « ط »obtenu à partir (a) du modèle « ط » et (b) du modèle « ن »

Par ailleurs, l'étude de la procédure de recherche du maximum de votes, a montré que la méthode est sensible aux déformations. En effet, les différentes variations se traduisent souvent par un éparpillement des votes au voisinage du point de référence. Pour remédier à ce problème, nous cumulons pour chaque point de l'image le nombre de votes qu'il a obtenu dans son voisinage. Pour cela, chaque point caractéristique ne doit voter pour une cellule et l'un de ses huit voisins qu'une seule fois. Cette mesure a conduit à l'optimisation de la procédure d'estimation du point de référence et par conséquent à une nette amélioration du taux moyen de reconnaissance qui a atteint 99% (FIG. 5)[Tou 3d].

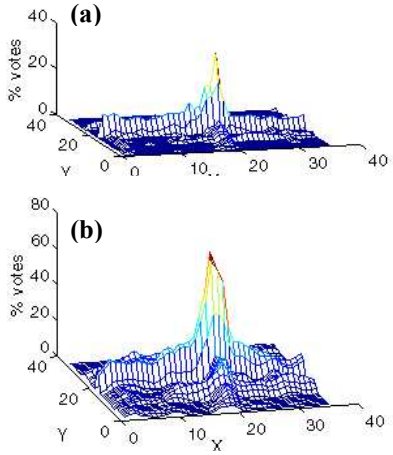


FIG. 5 – Tableaux de vote correspondant au modèle «ط» obtenu pour un échantillon du même caractère :
(a) avant et **(b)** après optimisation

4 Transformée de Hough Généralisée pour la reconnaissance de l'écriture arabe imprimée

Comme il a été expliqué précédemment, la THG peut être utilisée non seulement pour reconnaître les caractères mais aussi pour déterminer leurs positions dans l'image grâce à la localisation de leurs points de référence. Notre but est donc d'explorer cet aspect pour la reconnaissance de l'écriture arabe imprimée cursive sans avoir recours à une étape de segmentation en caractères (ou en graphèmes).

L'approche proposée repose sur le module de reconnaissance développée dans la section 3. Elle comporte trois principales étapes : une étape de prétraitements, une étape d'identification et de localisation des caractères dans le pseudo-mot; la dernière étape correspond à une phase de post traitement.

4.1 Prétraitements

A partir d'un texte arabe imprimé digitalisé à 300 dpi, nous procédons d'abord à la localisation automatique des différents pseudo-mots contenus dans le texte et l'élimination des diacritiques. Pour chaque pseudo-mot un ensemble d'informations structurelles sont extraites (telles que position dans l'image, densité en pixels, largeur, hauteur). La détermination des points caractéristiques est enfin effectuée selon la méthode expliquée dans le paragraphe 3.1.

4.2 Application de la THG sur les pseudo-mots

Pour un pseudo mot donné, nous construisons pour chaque modèle de caractère stocké dans le dictionnaire, le tableau de vote correspondant, selon la méthode expliquée au paragraphe 3.4. Chaque tableau de vote ainsi obtenu n'est autre qu'une retranscription de l'image du pseudo-mot où sont consignés les pourcentages des votes de chaque pixel. Nous définissons le pourcentage des votes Pv_{ki}, pour un pixel P_i de l'image comme suit :

$$Pv_{ki} = \frac{\text{Nombre de points caractéristiques qui ont voté pour } P_i}{\text{Nombre de points caractéristiques du modèle k considéré}}$$

Les pourcentages de votes permettent de retrouver le ou les emplacements éventuels du point de référence du modèle du caractère considéré (FIG. 6). Lorsqu'un pourcentage dépasse un seuil prédéfini, le modèle du caractère est retenu ainsi que la position de son point de référence P_{Rk} dans l'image. Par la suite, nous effectuons un ajustement des pourcentages de votes pour les points de référence enregistrés comme suit :

$$P_{PR_k} = \frac{\text{Nombre de points caractéristiques qui ont voté pour } PR_k}{\text{Nombre de points caractéristiques présents dans } F_k}$$

où F_k est une fenêtre ayant comme centre, le point de référence PR_k du caractère présumé. Sa largeur étant celle du modèle k correspondant.

A la fin de cette procédure, nous obtenons pour chaque pseudo-mot, un ensemble de propositions des caractères qui le composent, leurs taux de présence ainsi que leurs positions. La figure 6 illustre la détection des points de référence des caractères présents dans l'image : la figure 6(a) montre la présence du caractère « حـ » à la fois au début et au milieu du pseudo-mot «جحد» et la figure 6(b) montre la présence du caractère « د » à la fin du même pseudo-mot.

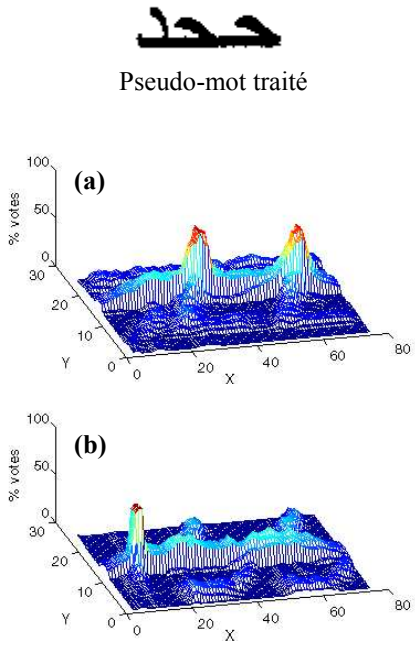


FIG. 6 – Tableaux de vote du pseudo-mot «جحد» obtenus : (a) pour le modèle «حـ» (deux pics visibles indiquant les emplacements des deux caractères) et (b) pour le modèle «د»

4.3 Post-traitement

Il s'agit de vérifier un ensemble de règles contextuelles ce qui permet d'éliminer chaque proposition qui présente une contradiction entre la forme du caractère correspondant et sa position dans le pseudo-mot. Au cas où plusieurs propositions sont candidates pour la même position, nous retenons celle qui présente le plus grand pourcentage de vote.

4.4 Expérimentations

Nous avons testé la méthode proposée sur des extraits de textes comportant environ 6400 caractères avec des pseudo-mots de longueurs variables allant jusqu'à 10 caractères successifs. Le taux de reconnaissance en caractères obtenu est de 96%.

5 Conclusion

Dans ce papier nous avons présenté une méthode de reconnaissance de l'écriture arabe imprimée par une approche de segmentation par reconnaissance basée sur la Transformée de Hough Généralisée. Le système repose sur un module de reconnaissance de caractères par THG préalablement élaboré. L'approche proposée a permis d'élargir le champ d'application de la THG à la reconnaissance de l'écriture cursive sans pour autant utiliser des dictionnaires de pseudo-mots ce qui restreindrait la méthode à des applications à vocabulaire limité. De plus, l'approche proposée permet de s'affranchir efficacement des problèmes classiques liés à la segmentation explicite en caractères. Les performances enregistrées sont très encourageantes ce qui laisse le champ ouvert à d'autres perspectives.

6 Bibliographie

[Ess 02] N. ESSOUKRI BEN AMARA: "Sur la problématique et les orientations en reconnaissance de l'écriture arabe". *In Proc. Col. Int. Francophone sur l'Ecrit et le Document CIFED02*, Hammamet, Tunisie, 2002, pp. 1-10.

[Ess 03] N. ESSOUKRI BEN AMARA, F. BOUSLAMA: "Classification of Arabic Script Using Multiple Sources of Information: State of the Art and Perspectives". *Special issue on Multiple Classifiers, Int. Journal on Document Analysis and Recognition IJDAR03*.

[Fak 93] M. FAKIR, C. SODEYAMA: "Recognition of Arabic printed scripts by dynamic programming matching method". *IEICE Trans. INF. & SYST.*, vol. E76-D, n°.2 February, 1993.

[Mai 85] H.MAITRE : "Un panorama de la transformation de Hough ". *Traitement de signal*, vol 2, 1986, pp 305-317.

[Mar 01] MARKUS ULRICH, CARSTEN STEGER, ALBERT BAUMGARTNER, AND HEINRICH EBNER, "Real-time object recognition using a modified generalized Hough transform". *In Eckhardt Seyfert, editor, Photogrammetrie --- Fernerkundung --- Geoinformation: Geodaten schaffen Verbindungen, 21. Wissenschaftlich-Technische Jahrestagung der DGPF*, Berlin, 2001, pp 571-578.

[Rui 00] J. RUIZ-PINALES, E. LECOLINET: "Cursive handwriting recognition using the Hough transform and a neural network". *ICPR'00*, pp. 231-234.

[Tou 03a] SO.TOUJ, N. ESSOUKRI BEN AMARA, H.AMIRI, "Reconnaissance de l'écriture arabe imprimée par une approche de segmentation par reconnaissance basée sur la Transformée de Hough". *Sciences, Electroniques, Technologie de l'Information et des Télécommunications SETIT'2003*, Mars, Sousse, Tunisie 2003.

[Tou 03b] SO.TOUJ, N. ESSOUKRI BEN AMARA, H.AMIRI, "Use of the Generalized Hough Transform for Printed Arabic Characters Recognition". *Signals, Systems, Decision & information technology SSD'2003*, Sousse, Tunisie 2003.

[Tou 03c] SO.TOUJ, N. ESSOUKRI BEN AMARA, H.AMIRI, "Generalized Hough Transform for Arabic Optical Character Recognition". *International Conference on Document Analysis and Recognition ICDAR'2003*, Edinburgh, Scotland 3-6 Août 2003, pp. 1242-1246.

[Tou 03d] SO.TOUJ, N. ESSOUKRI BEN AMARA, H.AMIRI, « Deux Approches basées sur la Transformée de Hough pour la reconnaissance des caractères arabes imprimées » *Sciences, Electroniques, Technologie de l'Information et des Télécommunications SETIT'2004*, Mars, Sousse, Tunisie 2004.

[Zar 98] S. ZARROUK, « Reconnaissance générale de PAW de l'arabe écrit » *DEA*, Ecole Nationale d'Ingénieurs de Tunis, Octobre 1998.