EXTRACTION DES LIGNES D'UN TEXTE MANUSCRIT ARABE

A.BENNASRI, A.ZAHOUR, B. TACONET
Laboratoire D'informatique du Havre
Place Robert Schuman, 76610, Le havre, France.
E_mail: bennasri@hotmail.com, zahour}@iut.univ-lehavre.fr

Abstract

One of the first stages in the conception of a writing recognition system is the segmentation of the text into lines. This operation is relatively easy if the text is not inclined and if lines do not overlap. These conditions can be satisfied for printed writing.

In the case of the handwriting without constraint, the writing fluctuates and can present an important slant with regard to the horizontal ; two adjacent lines can overlap, giving back the delicate separation.

In this paper, we propose an original method to extract lines of an Arabic handwritten text without any constraint for the writer. After having detected start points of all lines, by a partial projection, we then proceed to a partial contour following of every line ; first in the direction of the writing, then in the opposite direction. At the output of this operation, the adjacent lines are perfectly separated. The diacritical points, on which stay a doubt, are marked such and their definitive affectation will be validated at the time of the recognition. This method has been tested on about one-hundred Arabic texts written by different writers.

Résumé :

Une des premières étapes dans la conception d'un système de reconnaissance de l'écriture est la segmentation du texte en lignes. Cette opération est relativement facile si le texte n'est pas incliné, et si les lignes ne se chevauchent pas. Ces conditions peuvent être satisfaites pour l'imprimé. Dans le cas du manuscrit sans contrainte, l'écriture fluctue et peut présenter une inclinaison importante par rapport à l'horizontale: deux lignes adjacentes peuvent se chevaucher, rendant la séparation délicate.

Dans cet article, nous proposons une méthode originale pour extraire les lignes d'un texte manuscrit arabe sans aucune contrainte pour le scripteur. Après avoir détecté les points de départ de toutes les lignes, par une projection partielle, nous procédons à un suivi de contour partiel de chaque ligne : d'abord dans le sens de l'écriture, puis dans le sens opposé. A l'issue de cette opération, les lignes adjacentes sont alors parfaitement séparées. Les points diacritiques, sur lesquels demeure un doute sont marqués tel et leur affectation définitive sera validée lors de la reconnaissance. La méthode a été testée sur une centaine de textes arabes écrits par différents scripteurs.

Mots Clés :

Ecriture manuscrite arabe, segmentation, projections partielles, suivi partiel de contour.

I) Introduction :

La segmentation d'un texte manuscrit en lignes d'écriture est une étape nécessaire dans le développement d'un système de reconnaissance automatique de l'écriture. Cette opération est rendue délicate, dans le cas de l'écriture manuscrite, par la présence des espacements irréguliers entre lignes et des fluctuations de la ligne directrice de l'écriture par rapport à l'horizontale. Les lignes de textes peuvent être imbriquées ou collées lorsque hampes et jambages appartenant à deux lignes consécutives sont proches ou se touchent. Différentes directions de lignes peuvent coexister sur une même page.

Pour l'écriture arabe, la présence massive des points diacritiques complique en plus cette tâche comme le montrent les textes choisis dans les figures de cet article.

La plupart des études sur la segmentation d'une page en ligne s'appuient sur une décomposition de l'image en composantes connexes.

Dans [1], les composantes connexes sont extraites puis regroupées en alignement. Des situations de conflit peuvent apparaître pendant le processus de regroupement à cause de l'interpénétration des lignes ou de leur chevauchement. Dans ce cas, une analyse locale du conflit est réalisée. Si celle-ci suffit, les composantes connexes ambiguës sont affectées à un alignement unique. Dans le cas contraire l'analyse locale est suivie d'une analyse globale qui affine la détection des alignements.

Dans [2] on propose une méthode basée sur la segmentation ascendante par fusion des composantes connexes. La page est d'abord segmentée en blocs. On

calcule la hauteur d'un caractère sans hampe ni jambage, tel que le «x» ; cette estimation locale au bloc, fournit une valeur relativement précise. Le calcul est basé sur l'étude de la fonction de distribution de la hauteur des composantes connexes du bloc. La méthode est appliquée aux documents imprimés présentant une petite inclinaison.

Dans [3] la segmentation en lignes et en mots est réalisée en plusieurs étapes. Les composantes connexes sont d'abord détectées et étiquetées. Les zones de chevauchement entre lignes adjacentes sont localisées. Les composantes connexes sont ensuite regroupées en lignes et en mots.

La méthode proposée dans [4] utilise trois segmentations successives et indépendantes, de même structure de base. Chacune d'entre elle fonctionne selon le schéma suivant :

1) découpage de l'image en colonnes,
2) construction de l'histogramme de la projection horizontale pour chaque colonne,
3) définition des chemins identifiant les lignes de textes par l'utilisation des densités maximales des histogrammes,
4) segmentation en lignes,
5) affinement de la segmentation par un suivi de contour des caractères.

Cette méthode permet de détecter la plupart des erreurs de segmentation. Par contre les auteurs ne proposent aucune procédure de correction. Le principe de la projection partielle a été aussi utilisé dans [5].

Dans une étude récente consacrée à l'écriture arabe[6], les techniques utilisées pour la segmentation en lignes sont basées en grande partie sur la projection horizontale.

La méthode que nous présentons utilise la technique du suivi de contour partiel. Le point de départ de chaque ligne est déterminé à l'aide d'une projection partielle.

Cette approche impose une contrainte faible au scripteur à savoir que les lignes adjacentes ne doivent pas être collées.

Cet article est organisé comme suit :

Le paragraphe II décrit la méthode de la projection partielle appliquée à l'écriture manuscrite arabe.

Le troisième paragraphe présente les étapes de notre algorithme de segmentation en lignes.

Nous terminons cette communication par une présentation détaillée des résultats prouvant les performances de notre approche.

II) Projection partielle :

La méthode de projection horizontale totale consiste a faire la somme de tous les pixels noirs sur chaque ligne et de construire l'histogramme correspondant. Cette technique présente trois inconvénients majeurs :

a) Dès que le texte devient incliné l'histogramme obtenu n'est plus exploitable .

b) Elle est sensible au chevauchement : la détection des minima locaux devient très difficile.

c) La présence des points diacritiques peut donner lieu à de faux minima.

La méthode de la projection partielle est une alternative intéressante pour pallier au problème de l'inclinaison. La mise en œuvre de cette méthode nécessite les étapes suivantes:

- subdiviser le texte en colonnes. La largeur d'une colonne est celle d'un mot environ, car l'inclinaison d'un mot est moins importante que celle d'une ligne.
- Déterminer les minima des histogrammes résultant des projections horizontales pour chaque colonne. Ces minima correspondent à la zone de séparation entre deux lignes adjacentes.
- Représenter chaque minima par un trait horizontal de même longueur que la colonne. La liaison de ces traits entre eux permet d'avoir la séparatrice de deux lignes adjacentes.

Cette méthode permet de surmonter le problème d'une inclinaison quelconque comme le montre la figure1.

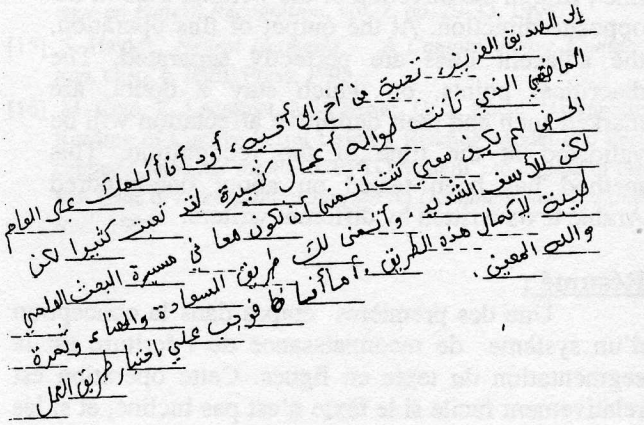
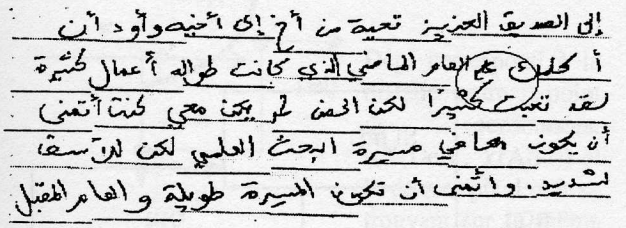


Figure 1 : application de la projection partielle sur un texte manuscrit arabe incliné

Limite de la méthode

Dans le cas où l'écriture ne serait pas alignée les traits suivent les oscillations de l'écriture. D'où la difficulté de trouver une ligne de séparation fiable. Ajoutons à cela que la présence d'un chevauchement sur une colonne fait disparaître le minima correspondant. Ceci se traduit par l'absence d'un trait séparateur comme le montre la figure 2.



Le cercle montre une zone
de chevauchement

Figure 2 : Projection partielle sur un texte manuscrit arabe présentant des oscillations.

III) Suivi de contour partiel

Cette méthode consiste à suivre le contour externe des composantes connexes en privilégiant soit le chemin vers le pixel le plus bas du mot si on se déplace de droite à gauche (phase aller), soit le chemin vers le pixel le plus haut si on se déplace de gauche à droite(phase retour).

Lors de ce suivi le déplacement d'une composante à l'autre se fait par un balayage horizontal. La mise en œuvre de l'algorithme nécessite trois étapes.

1) déterminer les points de départ du suivi de contour pour la phase aller.

2) effectuer le suivi de contour partiel pour la phase aller puis la phase retour.

3) repérer les symboles diacritiques (points, tracé secondaire d'un caractère) puis les affecter à un alignement unique.

a) Description de l'algorithme

1ére étape : On réalise une projection partielle limitée à la première colonne. Les minima locaux des histogrammes obtenus représentent les points de départ de la phase "aller". Souvent la présence de symboles diacritiques (figure 3) sur la première colonne peut générer de faux minima et donc de faux points de départ.

(a) (b)

Figure 3: Caractères présentant des ouvertures (a) kaf, (b) alif.

Pour les éliminer, on procède comme suit:

Sur l'histogramme de la projection partielle, on calcule la hauteur de chaque composante comprise entre deux minima(figure 5 (b)). On en déduit une hauteur moyenne notée (hm).

Les composantes dont la hauteur est inférieure à hm peuvent être soit des symboles diacritiques, soit

Des caractères sans hampe ni jambage tel que ceux de la figure 4.

(a) (b)

Figure 4: Caractères sans hampes ni jambage (a) mim, (b) ba

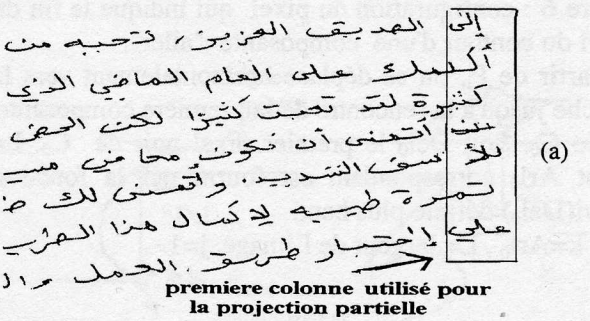
Pour les différencier on procède comme suit:

On calcule les distances entre toutes les composantes successives et on en déduit une distance moyenne (dm).

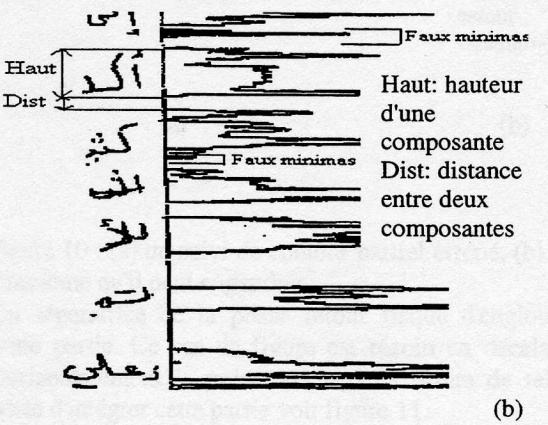
Soit Ci-1, Ci, Ci+1 respectivement la composante précédente, courante et suivante. Et soit h(Ci) la hauteur de la composante Ci.

Si d(Ci,Ci-1) ou d(Ci+1,Ci)<dm

Et si h(Ci)<hm alors la composante est un symbole diacritique. Elle est annexée à la composante la plus proche.



(a)



(b)

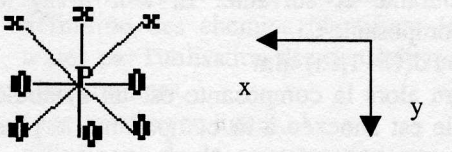
Figure 5: a)sélection de la première colonnedu texte, (b) projection horizontale corespondante.

2éme Etape :

Elle se divise en deux phases : une phase "aller" et une phase" retour".

1) Algorithme de la phase "aller"

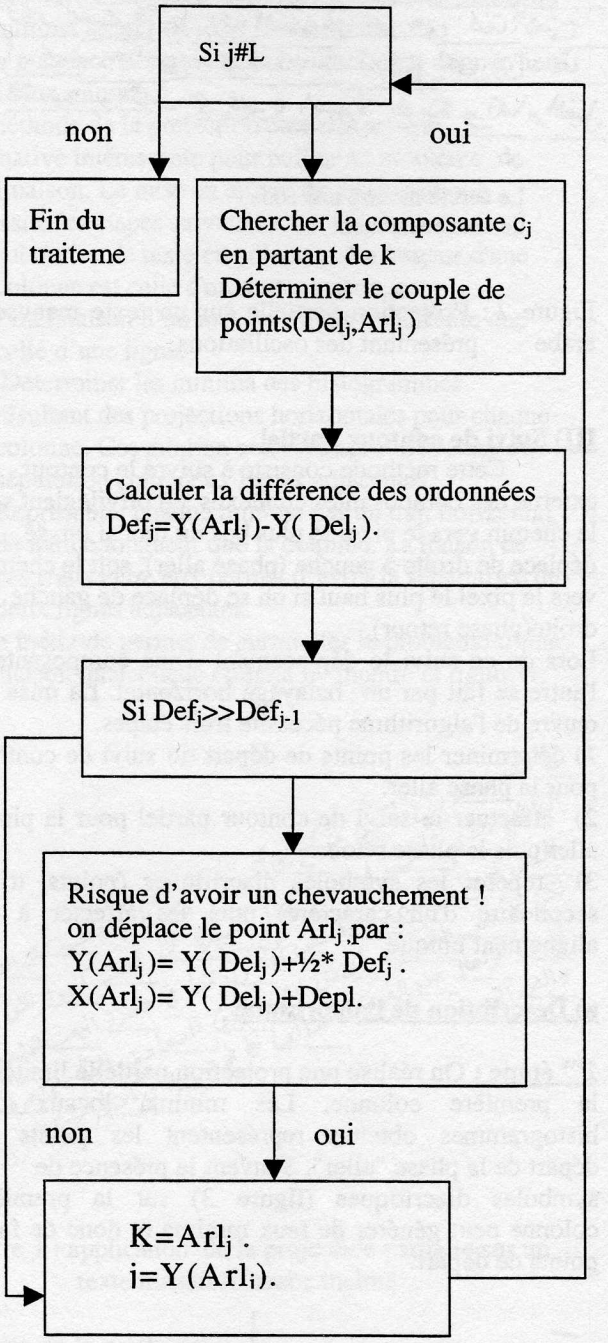
Soit R(o,x,y) le repère orthonormé utilisé. L'origine O a pour cordonnées $x_0=L-1, y_0=0$.
L étant la largeur de l'image.
Soit P_i le point de départ de la ligne L_i déterminé lors de la première étape.
On appelle "Suivi(Del_j) " la fonction qui effectue le suivi de contour partiel de la composante C_j à partir du pixel Del_j. Cette fonction retourne les cordonnées du pixel le plus bas de la composante C_j dont la configuration est donnée à la figure 6. Ce point sera noté Arl_j.



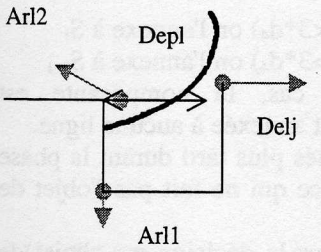
X peut prendre la valeur 0 ou 1

figure 6 : configuration du pixel qui indique la fin du suivi du contour d'une composante "aller".

A partir de P_i, on se déplace horizontalement vers la gauche jusqu'à la rencontre de la première composante notée C_0. Soit Del_0 le premier pixel noir de C_0. Le point Arl_0 correspondant est fourni par la fonction Suivi(Del_0) définie plus haut.
Soit $k=Arl_0$, L= largeur de l'image, j=1.



Depl est représenté sur la figure 7
On remonte le point Arlj , car ce choix est validé par nos tests sur plusieurs textes manuscrits arabes lisibles

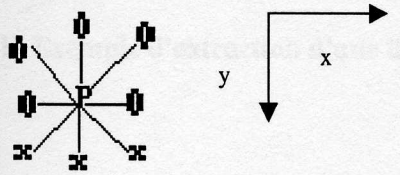


Depl correspond à la distance entre le point de cordonnées: ((X(Del) , Y(Arl2)) et le dernier pixel noir se trouvant sur la même horizontale

figure 7: Arl2 est obtenu après correction de Arl1.
Après on regroupe tous les points Arl_j et Del_j pour obtenir un tracé séparateur des lignes L_i et Li+1.

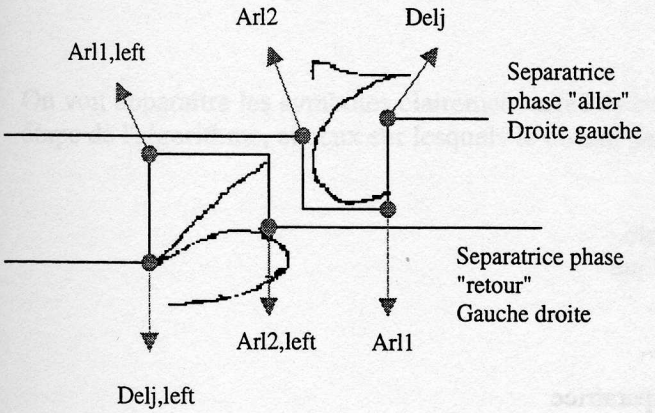
2) Algorithme de la phase "retour"

La phase retour utilise deux séparatrices S_i et S_{i+1} respectivement des lignes (L_i,L_{i+1}) et (L_{i+1},L_{i+2}) déterminées lors de la phase "aller".
Le point Arl de la dernière composante de S_{i+1} nous permet de déterminer le point de départ P_i de la phase retour. A partir de ce point on effectue le même traitement que celui de la phase "aller ". On le fait dans un déplacement gauche droite, en privilégiant le chemin vers le pixel noir le plus haut dont la configuration est celle da la figure 8.

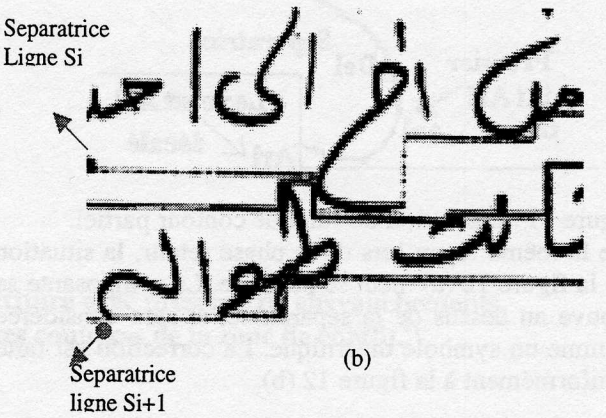


x peut prendre la valeur 0 o 1

figure 8 : configuration du pixel qui indique la fin suivi du contour d'une composante "retour".
La présence de la séparatrice S_i nous évite tout risque de chevauchement comme le montre la figure 9.



(a)



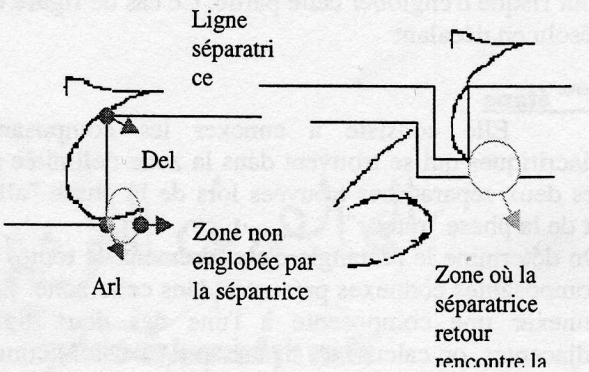
(b)

figure 9 : (a)traitement d'une zone de chevauchement. (b) illustration sur une partie d'un texte

b) Les problèmes rencontrés:

Lors de la phase "aller" on peut se trouver dans la situation de la figure 10, où une partie de la composante est en dehors de la séparatrice.

composante est en dehors de la séparatrice.



(a) (b)

figure 10 : (a) un suivi de contour partiel erroné, (b) le problème qu'il peut engendrer.
La séparatrice de la phase retour risque d'englober cette partie. Ce cas de figure est résolu en décalant horizontalement le point Arl vers la droite de telle sorte d'intégrer cette partie voir figure 11.

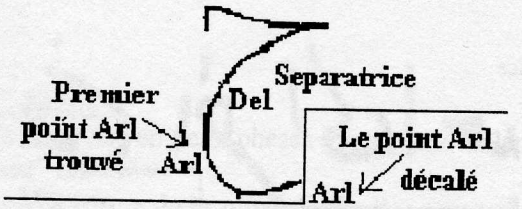
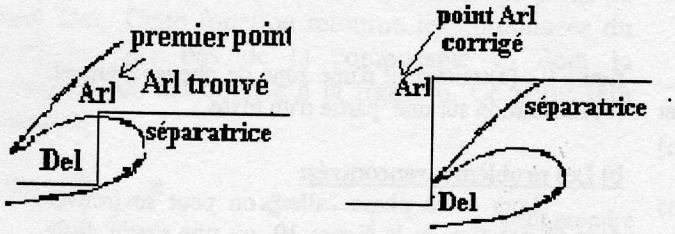


figure 11 : correction du suivi de contour partiel.

De la même façon lors de la phase retour, la situation de la figure 12 (a) :peut se produire. La composante se trouve au dessus de la séparatrice, et sera considérée comme un symbole diacritique. La correction est faite conformément à la figure 12 (b).



(a) (b)

figure 12 : (a) tracé erroné, (b) correction.

tour risque d'englober cette partie. Ce cas de figure est résolu en décalant

3eme étape

Elle consiste à annexer les composantes diacritiques qui se trouvent dans la zone délimitée par les deux séparatrices trouvées lors de la phase "aller" et de la phase "retour".

On détermine le rectangle d'encadrement de toutes les composantes connexes présentes dans cette zone. Pour annexer une composante à l'une des deux lignes adjacentes, on calcule les distances d_i(i=1 à 4) comme indiqué sur la figure 13.

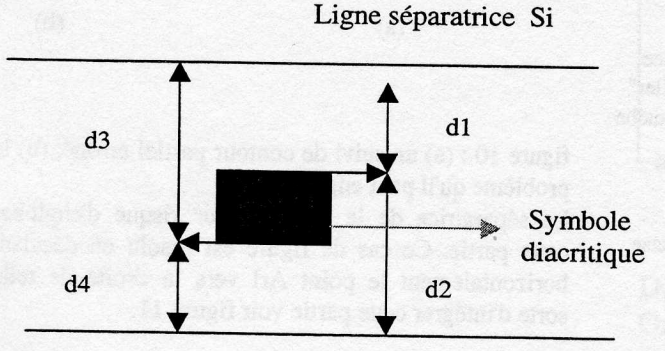


figure 13 : traitement d'un symbole diacritique

Si (d_2>3*d_1) et (d_3<3*d_4) on l'annexe à S_i

Si (d_2<3*d_1) et (d_3>3*d_4) on l'annexe à S_{i+1}

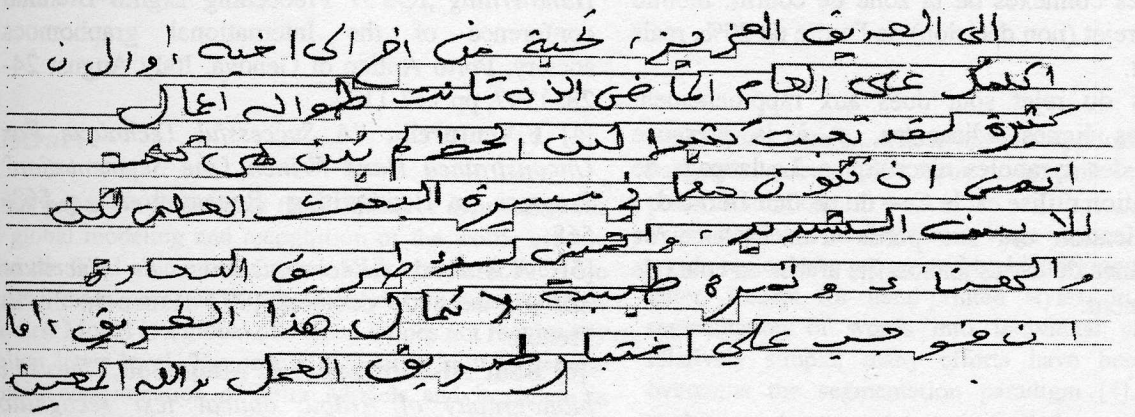
Dans les autres cas, la composante est marquée, mais n'est annexée à aucune ligne.

Ces cas seront traités plus tard durant la phase de la reconnaissance qui ne fait pas l'objet de ce travail.

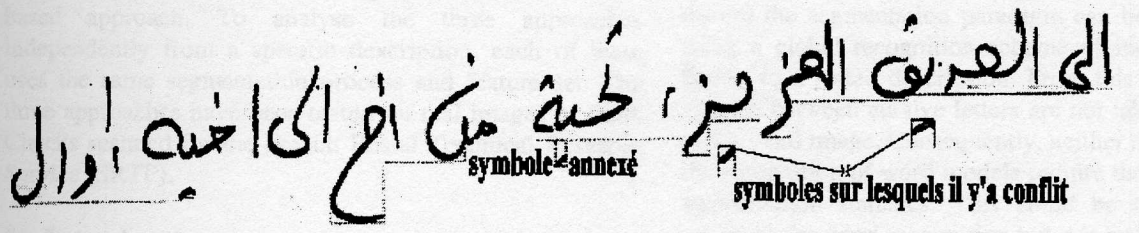
Le seuil utilisé pour la décision est choisi de manière à privilégier le rejet.

IV) Résultats:

a)Cas d'un texte présentant des oscillations de l'écriture avec présence de chevauchements. "Séparation en ligne et Affectation des composantes connexes de la zone de conflit."



b) Exemple d'extraction d'une ligne.



On voit apparaître les symboles clairement affectés à cette ligne à l'issue de la troisième étape de l'algorithme, et ceux sur lesquels le conflit persiste.

V) Conclusion :

Nous avons présenté une méthode originale de la segmentation en lignes d'un texte manuscrit arabe. Aucune contrainte n'est imposée au scripteur hormis le fait que les lignes ne doivent pas être collées. La méthode a été testée sur 100 textes manuscrits écrits par 20 scripteur, équivalant à 2000 lignes. Les résultats de segmentation à l'issue des deux premières étapes de l'algorithme donnent un taux 98%. Les causes d'erreurs sont principalement dues à la présence au début des lignes, de caractères avec symboles diacritiques qui faussent la détection des points de départ.

La troisième étape, concernant l'affectation des composantes connexes de la zone de conflit, montre un taux de rejet (non décision) de l'ordre de 20%, mais 0% d'erreur.

Les causes du rejet sont dues aux rapprochement exagéré des lignes adjacentes, et à la présence importante des symboles diacritiques. Le langage de programmation utilisé est le C++ du produit Borland.

Cette application est une partie d'un système de reconnaissance de textes manuscrits arabes en cours de développement.

Références :

[1] **Likforman-Sulem L, Faure C.** "*Une méthode de résolution des conflits d'alignements pour la segmentation des documents manuscrits.*" CNED'94, 3ème Colloque National sur l'écrit et le Document. Rouen 6,7et 8 Juillet 1994, pp 265-272, in French.

[2] **Derrien-Péden D,Morbé O.,Thépaut A.** "*Analyse et reconnaissance de documents par méthodes morphologique et neuronale.*", CNED'94, 3ème Colloque National sur l'écrit et le Document. Rouen 6,7et 8 juillet 1994, pp 419-425 , in French.

[3]**Wu ,Xiaoying et Graham ,Leedham.** "*Separating Lines and Words in Unconstrained Handwriting*",IGS'97 Proceeding Eighth Biennial conference of the International graphomocs society, Porto Antico di Genova, Italy August 24-28, 1997, pp 117-118

[4] **F.Venturelli.** "*A Successful Technique For Unconstrained Hand-Written Line Segmentation*" Progress in Handwriting Recognition" , p 563-568.

[5] **A.Ameur.** "*Reconnaissance de l'écriture arabe*", thèse de doctorat 1992 Rouen la 3i, In French.

[6] **Badr Al-Badr,Sabri A. Mahmoud**, "*Survey and bibliography of Arabic optical text recognition.*" Signal Processing 41(1995) p 49-77.