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