Contribution à la segmentation de textes manuscrits anciens

Abderrazak Zahour – Bruno Taconet – Saïd Ramdane

Equipe GED - Université du Havre
IUT du Havre
Place Robert Schuman
F-76 610 Le Havre

email{abderrazak.zahour,bruno.taconet,ramdane}@univ-lehavre.fr

Résumé : *Dans cet article, nous présentons une méthode de segmentation en lignes de textes manuscrits arabes. Les documents traités sont issus dans leur grande majorité de textes manuscrits anciens numérisés et stockés dans des bases documentaires. La diversité des styles des écritures utilisés, ainsi que les différentes présentations des textes montrent que la retro-conversion de ces documents reste un défi majeur pour la prochaine décennie. Nous pensons qu'il faut plusieurs stratégies de segmentation pour arriver à extraire toutes les lignes de texte de tels documents. La méthode proposée (en cours de développement) s'adresse à des textes manuscrits anciens comme ceux de la figure 1. L'approche utilisée segmente souvent une page de document en trois types de blocs de texte : les petits blocs représentent généralement les symboles diacritiques ; les blocs moyens correspondent au corps du texte et les grands blocs reflètent le chevauchement entre mots des lignes voisines. La segmentation des grands blocs ainsi que l'appariement entre tous les blocs générés permet de trouver les lignes de texte. Les premiers résultats obtenus dans le stade actuel de la méthode sur une dizaine de textes manuscrits sont encourageants.*

Mots-clés *: Ecriture manuscrite; Segmentation en ligne; classification automatique; K-means*

1 Introduction

Une des premières étapes dans la conception d'un système de reconnaissance de l'écriture manuscrite est l'identification et l'extraction des lignes de texte. L'objectif de cette étape est d'assigner chaque composant du texte à une ligne appropriée ; ce qui permet de préparer les données pour les traitements ultérieurs tel que la normalisation, la segmentation en mots et l'extraction des caractéristiques. Ces étapes nécessitent que les données correspondent à une ligne d'écriture.

La segmentation de textes manuscrits non contraints est rendue difficile par la variation de la distance interligne et l'ondulation des lignes de base générant des orientations différentes du texte. Les caractères de deux lignes de texte peuvent se toucher ou se chevaucher. Ce qui complique considérablement la segmentation en ligne. Dans l'écriture arabe, ces situations existent fréquemment à cause de la présence de caractères ascendants et descendants La présence massive de symboles diacritiques génère souvent des fausses lignes. Une recherche active a été menée pour les textes manuscrits latins hors ligne ([BRU], [LIK], etc)ou en ligne ([RAT]…). Des méthodes concernant l'écriture orientale ont été proposées. Citons [TSE] pour l'écriture chinoise, et [PAL] pour l'écriture Bangla. Très peu de méthodes à notre connaissance ont été proposées pour les textes manuscrits arabes. Celles recensées dans la littérature [AMI] sont surtout basées sur la technique de la projection horizontale ou récemment sur la technique du RLSA [HAD] et ne peuvent être appliquées pour les textes manuscrits.

Nous avions développé par le passé [BEN], [ZAH] une méthode de segmentation de textes manuscrits basée sur la technique de la projection partielle et du suivi partiel du contour inférieur des mots. La méthode, testée sur une centaine de textes manuscrits modernes (lettres, notes manuscrites, …) donne un taux de bonne extraction de l'ordre de 82%. Si la méthode résout assez bien le cas des chevauchements et des ondulations du texte, elle échoue dans le cas des caractères collés appartenant à des lignes voisines. Les symboles diacritiques fournissent souvent de fausses lignes.

Une analyse d'un échantillon de documents manuscrits anciens montre que les textes sont dense. L'interligne est très faible. On rencontre souvent des situations où les caractères de lignes voisines sont collés (figure 1). Nous avons modifié la méthode pour qu'elle s'adapte à ces nouvelles configurations. L'idée consiste à extraire - pour un document analysé - les paramètres de segmentations (seuils) de manière automatique et sans apprentissage préalable. L'approche est alors globale. Elle est basée sur la détection et la classification de l'ensemble des blocs de l'image. Pour améliorer les résultats de segmentation, il est nécessaire d'intégrer l'information locale (relations de voisinage entre les blocs). Nous présentons dans les sections suivantes les étapes de l'approche adoptée.

2 Segmentation en ligne

2.1 Décomposition de l'image en blocs rectangulaires

Le document numérisé sous format JPEG est recueilli de la bibliothèque nationale tunisienne [TUN]. Le cadre de l'image est supprimé manuellement (actuellement la

méthode ne permet pas de détecter et de supprimer de manière automatique le cadre). La page de document est ensuite binarisée puis découpée en 8 colonnes (figure 1). Pour un document avec le format A4, 8 colonnes semblent donner de bons résultats. Nous déterminons ensuite l'histogramme des projections horizontales pour chaque colonne. Les minima de ces histogrammes nous fournissent les blocs de textes (figure 1). Les caractéristiques de l'image réellement traitée dont une partie est montrée (figure 1) sont résumés ci-dessous :

Nombre de colonnes : 8 nbr_bloc[i] : nombre de blocs de la colonne i

Toutes les données qui suivent sont exprimées en pixels.

largeur de l'image : 530 ; Hauteur de l'image : 884 ; Nombre de pixels par colonne : 66; reste : 2 ; Nombre de pixels de la colonne 7 : 68:

nbr_bloc[0] = 20.000000 nbr_bloc[1] = 14.000000
nbr_bloc[2] = 22.000000 nbr_bloc[3] = 16.000000
nbr_bloc[4] = 20.000000 nbr_bloc[5] = 20.000000
nbr_bloc[6] = 21.000000 nbr_bloc[7] = 20.000000

nombre total des blocs : 153 ; Hauteur du plus grand bloc : 217 ;

2.2 Classification des blocs

Une page de document texte peut contenir trois type de blocs selon leur hauteur: les petits blocs représentent les symboles diacritiques mais aussi les composantes générées par le découpage de l'image en colonnes. Les blocs moyens correspondent au tracé principal des mots. Enfin les grands blocs résultent du chevauchement des caractères ascendants et descendants mais aussi des caractères collés des lignes voisines. Chaque type de bloc représente alors une classe. L'histogramme des hauteurs des blocs de l'image réellement traitée (taille 23 cm*14 cm) est donné à la figure 2.

La détection des types de blocs est réalisée à l'aide de l'algorithme de classification automatique des k-means avec en sortie 3 classes (les trois types de blocs). L'algorithme converge au bout de 5 itérations.

Nous donnons ci-dessous une description sommaire de l'algorithme :

On fournit en entrée le tableau des hauteurs des blocs et le nombre total *n* des blocs : $(h_i)_{1\le i\le n}$.

Initialisation :

Les centres de masse m1, m2, m3 des trois classes C1, C2, C3 sont initialisés comme suit : $m_1 = 0;\quad m_2 = \frac{1}{n}\sum_{i=1}^{n} h_i;\quad m3 = h_{\max}$; m2 est la moyenne statistique; n_i est le nombre de blocs de hauteur h_i ; h_{max} est la hauteur du plus grand bloc.

Pour chaque bloc, trois distances euclidiennes sont évaluées: $(d_j)_{1\le j\le 3} = \left\| h_i - m_j \right\|$. On affecte le bloc i à la classe j si la distance dj est minimale.

Trois nouveaux tableaux sont générés : $(h_{i,j})_{1\le i\le n_j, 1\le j\le 3}$; n_j est le nombre de blocs de la classe j.

Itérations :

A chaque itération nous calculons les nouveaux centres de masse $(m_j)_{1\le j\le 3}$ $(m_j)_{1\le j\le 3} = \frac{1}{n_j}\sum_{i=1}^{n_j} h_{i,j}$ et reprenons le calcul précédent des distances $(d_j)_{1\le j\le 3}$

Arrêt :

Les trois classes sont définitives quand les centres de masse ne varient plus.

Les résultats obtenus pour l'image réellement traitée sont résumés ci-dessous:

Toutes les grandeurs sont données en pixels.

Paramètres de la classe C1 :

m1 = 8.482759 ; n1 = 29 ; hmin1 = 1.000000 ; hmax1 = 17.000000

Paramètres de la classe C2 :

m2 = 25.799999 ; n2 = 80 ; hmin2 = 18.000000 ; hmax2 = 38.000000

Paramètres de la classe C3 :

m3 = 58.571430 ; n3 = 44 ; hmin3 = 40.000000 ; hmax3 = 217.000000

Avec: ***mi*** : centre de masse de la classe **i** ; ***ni***: nombre de blocs de la classe **i** ; ***hmini*** (resp. ***hmaxi***) la hauteur mini (resp. max) des blocs de la classe **i**.

Le seuil S1 de séparation entre la classe C1 et la classe C2 est S1 = (m1+m2)/2 = 17.141379

Le seuil S2 de séparation entre la classe C2 et la classe C3 est S2 = (m2+m3)/2 = 42.1857145

La figure 3 montre les trois types de blocs générés : En blanc, les grands blocs. En gris foncé, les blocs moyens (tracé principal des mots). Enfin en gris clair les petits blocs. On peut remarquer que certains blocs frontière peuvent migrer de leur classe d'origine à une autre classe; par exemple le premier bloc de la colonne 0.

2.3 Segmentation des grands blocs

Cette phase consiste à segmenter les grands blocs. Nous avons besoin de connaître avec la plus grande précision possible les distances interlignes. Le document peut contenir des lignes incomplètes qui peuvent générer des espaces blancs importants entre lignes voisines dans certaines régions. Une classification automatique des espaces blancs entre lignes s'impose. Nous utilisons

l'algorithme des k-means présenté précédemment avec deux classes CB1 et CB2. Les espaces blancs appartenant à la classe CB1 constituent pratiquement les espaces interlignes.

On fournit comme entrée à l'algorithme le tableau des distances des espaces blancs entre deux blocs voisins. De telles distances sont calculées simplement en évaluant la différence entre l'ordonnée inférieure d'un bloc de texte et l'ordonnée supérieure du bloc situé immédiatement en dessous et appartenant à la même colonne. Les paramètres des deux classes de l'image de la figure 1 sont résumés ci dessous :

Paramètres de la classe CB1 :

m1 = 2.000000 ; n1 = 88 ; dmin1 = 0.000000 ; dmax1 = 5.000000

Paramètres de la classe CB2 :

m2 = 9.824561 ; n2 = 57 ; dmin2 = 6.000000 ; dmax2 = 44.000000

Avec ***dmini*** (resp. ***dmaxi***) la distance minimale (resp. maximale) entre deux blocs blancs voisins de la même colonne appartenant à la classe **i**

Le seuil de séparation entre les classes CB1 et CB2 est calculé comme suit : S = (m1+m2)/2 ; S = 5.9122805

Les blocs blancs dont la taille est inférieure à S (appartenant dans leur grande majorité à la classe CB1) sont les seuls utilisés pour la segmentation des grands blocs et sont rangés dans un tableau qui regroupe alors les distances interlignes estimées. Nous l'appelons CB.

Un grand bloc de texte est généralement issu de la fusion entre mots qui contiennent des caractères ascendants et descendants. Ces mots possèdent des hauteurs relativement importantes et appartiennent naturellement aux blocs de texte de la classe C2. Nous résumons ci-dessous la procédure de segmentation des grands blocs (blocs de textes de la classe C3) :

– Déterminer la hauteur hm du bloc moyen de la classe C2 et retenir uniquement les blocs j dont la hauteur h_j est supérieure à hm ,

soit hg_i la hauteur du bloc i (à segmenter) de la classe C3. Soit hb_k la hauteur du bloc blanc d'indice k du tableau CB,

- Calculer le nombre réel $Nls_{j,k} = \dfrac{hg_i - h_j}{h_j + hb_k}$ pour tout bloc j de la classe C2 tel que $h_j \geq hm$ et pour tout bloc k de hauteur hb_k du tableau CB. $Nls_{j,k}$ indique le Nombre de Lignes de Segmentation du bloc i, pour un couple (j,k) donné.

-déduire la fraction de pixels résiduelle $\varepsilon_{j,k} = Nls_{j,k} - (\text{int})Nls_{j,k}$,avec (int) la partie entière de…

Le couple (j,k) qui minimise $\varepsilon_{j,k}$ est utilisé pour reconstituer de manière dynamique les blocs et la distance interligne d'origine.

Les résultats obtenus sont montrés (figure 4). On peut voir par exemple que le premier bloc de la colonne 0 admet un $Nls_{j,k}$ égal à zéro. Il réintègre sa classe d'origine C2. Autre exemple :le quatrième bloc de la colonne 1 a été segmenté par cette méthode en deux blocs. La première séparatrice en pointillés (gris foncé) passe par l'ordonnée inférieure du premier bloc généré. La seconde séparatrice en pointillés (gris clair) immédiatement en dessous passe par l'ordonnée supérieure du deuxième bloc généré. La très faible hauteur entre les deux séparatrices correspond à la distance interligne estimée.

2.4 Recherche des lignes de texte

Une fois la segmentation des grands blocs réalisée, il faut chercher maintenant les lignes du texte. Un appariement entre les blocs des différentes colonnes est effectué. Nous utilisons pour cela les distances euclidiennes entre les ordonnées inférieurs des blocs. Nous comparons les blocs de la colonnes i avec ceux des colonnes i-1 et i+1, sauf pour les colonnes extrêmes 0 et 7 pour lesquels la comparaison se fait respectivement avec la colonne 1 et la colonne 6. Les blocs dont la distance entre leurs ordonnées inférieures est minimale sont appariés ensemble. Les séparatrices de lignes correspondent aux ordonnées inférieures de ces blocs. comme le montre la figure 5.

3 Discussion

Nous avons utilisé pour cette étude l'algorithme de classification automatique des k-means. La procédure développée ne demande pas d'apprentissage et utilise des informations de nature globale. Des situations de sous ou sur - segmentation des blocs peuvent avoir lieu. Les petits blocs peuvent générer des doublons dans les séparatrices de lignes. Nous pensons intégrer par la suite l'information locale (relations de voisinage entre les blocs des colonnes voisines) pour pallier à ces inconvénients et valider le module de segmentation en lignes.

Nous avons testé la méthode sur des textes imprimés et manuscrits latins ainsi que des textes imprimés arabes. Quand l'image du document présente les trois types de blocs, l'algorithme fournit des résultats semblables à ceux obtenus pour le texte de la figure 1. A titre d'exemple, l'image de la figure 6 correspond au fichier d'imprimante de la première page de cet article (format PDF) sorti en format non PostScript (par erreur) et constitue un cas d'étude. Les résultats sont donnés aux figures 7,8 et 9.

Pour les textes imprimés (ou manuscrits) avec interligne suffisant et peu inclinés il y a disparition des grands blocs. Deux types de blocs existent : les petits et les moyens blocs et donc deux classes au lieu de trois. Nous

avons modifié la méthode pour qu'elle s'adapte à tout type de document. L'implémentation actuelle tient compte de ces modifications. Nous résumons ci-dessous la démarche adoptée.

- Utiliser l'algorithme des k-means avec trois classes.
- Choisir une mesure de plausibilité de la classification. Nous avons utilisé pour cela l'indice proposée par [WU] appelé "Clustering Validity Index (CVI)" et basé sur le calcul de la densité intra et inter-classe.
- Refaire les étapes a) et b) avec 2 classes.
- Le CVI tend vers un maximum quand le nombre de classe est optimum.

Quand le CVI est grand pour deux classes, nous faisons abstraction sur le module de segmentation des grands blocs. Le calcul du CVI sur les images des figures 1 et 6 sont résumés ci-dessous :

Image traitée (une partie est montrée à la figure 1): k-means avec 3 classes

Centre des trois classes: m1 = 8.482759 m2 = 25.799999 m3 = 58.571430

Nombres d'éléments de chaque classe n1 = 29 n2 = 80 n3 = 44

Ecart_type de la classe i (ETi), Ecart_type moyen (ETM)

ET1=9.090569 *ET2*=13.480455 ET3=39.408474 *ETM*=24.612930

dist. intra_classe (INTRA)=4021.666748 *dist. inter_classe (INTER)*=272.000000

coef. de sep. des classes (SEP)=0.719793 *(CVI=SEP*INTRA)*=2894.767573

Image de la figure 1 : k-means avec 2 classes

Centre des deux classes: m1=21.955751 m2=73.849998

Nombre d'éléments de chaque classe: n1=113 n2=40

Ecart_type de la classe i (ETI), Ecart_type moyen(ETM) :ET1=9.791932 *ET2*=32.657158 *ETM*=24.107800

dist. intra_classe (INTRA)=6260.500000 *dist. inter_classe (INTER)*=138.000000

coef. de sep. des classes (SEP)= 0.373340 *(CVI =SEP*INTRA)*= 2337.29507

Image de la figure 6 : k-means avec 3 classes

Centre des trois classes : m1=6.000000 m2=18.051470 m3=70.266670

Nombres d'éléments de chaque classe : n1=8 n2=136 n3=38

Ecart_type de la classe i(ETi), Ecart_type moyen(ETM)

ET1=4.461433 *ET2*=4.440205 *ET3*=23.719994 *ETM*=14.168720

dist. intra_classe (INTRA)=6682.000000 *dist. inter_classe (INTER)*=319.000000

coef. de sep. des classes (SEP)= 0.444071 *(CVI =SEP*INTRA)*= 2967.283782

Image de la figure 6 : k-means avec 2 classes

Centre des deux classes : m1=20.215279 m2=66.000000

Nombre d'éléments de chaque classe : n1=144 n2=38

Ecart_type de la classe i (ETi), Ecart_type moyen(ETM) :

ET1=4.461433 *ET2*=25.538525 *ETM*=18.331947

dist. intra_classe (INTRA)=10341.000 *dist. inter_classe (INTER)*=166.000

coef. de sep. des classes (SEP)= 0.274160 *(CVI =SEP*INTRA)*= 2835.088536

Nous remarquons pour les deux images, que le CVI obtenu pour trois classes est le plus grand. D'où le choix de trois types de blocs.

4 Conclusion

Nous avons présenté une méthode (en cours d'élaboration) de segmentation en lignes de textes manuscrits arabe anciens. L'expérimentation, menée sur une dizaine de textes montre des premiers résultats encourageant, mais loin d'être satisfaisant. Quelques réglages restent à faire pour la segmentation des grands blocs, où les séparatrices traversent quelques fois les mots. L'intégration de notre algorithme de suivi de contour partiel [ZAH], appliqué aux grands blocs permettrait de résoudre ce type de problème. Des fausses lignes peuvent être générées par les petits blocs de texte (par exemple le troisième petit bloc de la colonne 6, entre autres). Une analyse locale permettrait de s'affranchir de ces situations. Nous y travaillons activement.

5 Bibliographie

[AMI 98] AMIN. A, « Off_line Arabic characters Recognition: The State Of the Art », *Pattern Recognition*, vol 31, n° 5, 1998, p 517-530.

[BEN.00] BENNASRI. A, ZAHOUR. A, TACONET. B, « Arabic script processing and application to postal addresses », *Actes de ACIDCA'2000, Vision and Pattern Recognition*, Mounastir, 22-24 mars 2000, Tunis, p. 74-79.

[BRU 99] BRUZZONE. E, COFFETTI. M.C, « An algorithm for Extracting Cursive Text Lines », *Actes de ICDAR'99*, Bangalore, 20-22 septembre 1999, India, p. 749-752.

[HAD 03] HADJAR. K, INGOLD. R, « Arabic Newspaper Page Segmentation », *Actes de ICDAR'2003*, Edinburgh, 3-6 Août 2003, Scotland, p. 895-899.

[LIK 94] LIKFORMAN-SULEM. L , FAURE. C, « Une méthode de résolution des conflits d'alignements pour la segmentation des documents manuscrits », *Actes de. la conférence nationale sur l'écrit et le document CNED'94*, Rouen,6-8 juillet 1994, France, p. 265-272.

[PAL 03] Pal. U, Datta. S, « Segmentation of Bangla Unconstrained Handwritten Text », *Actes de ICDAR'2003*, Edinburgh,Aout 3-6 2003, Scotland.

[RAT 00] RATZLAFF. E, « Inter-Line Distance and Text Line Extraction for Unconstrained On Line Handwriting », *Actes de IWFHR-7*, Amsterdam, 11-13 septembre 2000, p. 33-42,

[TSE 99] TSENG.Y.H, LEE H.J, « Recognition-based handwritten Chinese character segmentation using a probabilistic Viterbi algorithm », *Pattern Recognition Letters*, vol. 13, 1999, p. 791-806,

[TUN 03] http://www.bibliotheque.nat.tn/

[ZAH 01] ZAHOUR. A, TACONET. B, MERCY. P, RAMDANE. S,« Arabic hand-written text-line extraction », *Actes de ICDAR'2001*, Seattle, 8-11 septembre 2001,p. 281-285.

[WU 04] WU. S, CHOW. T.W.S,« Clustering of the self-organizing map using a clustering validity index based on inter and intra-cluster density », *Pattern Recognition*, vol 37, n°°2, février 2004, p 175-188.

6 Figures

(les images présentées ci-dessous sont une partie des images réellement traitées)

FIG. 1 - Segmentation en blocs

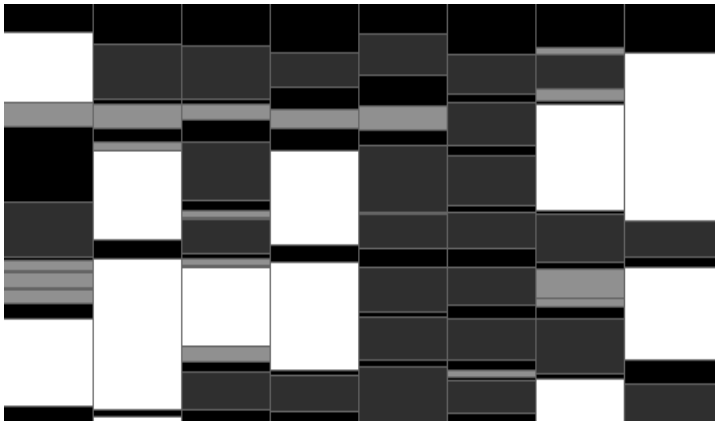


FIG. 3 - Classification des blocs

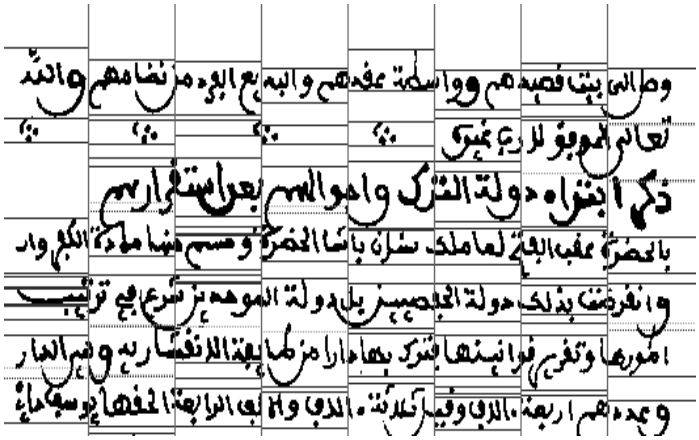


FIG. 4 - Segmentation des grands blocs

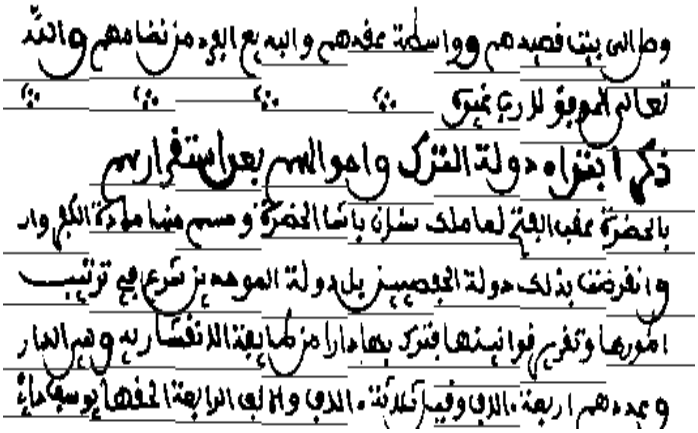
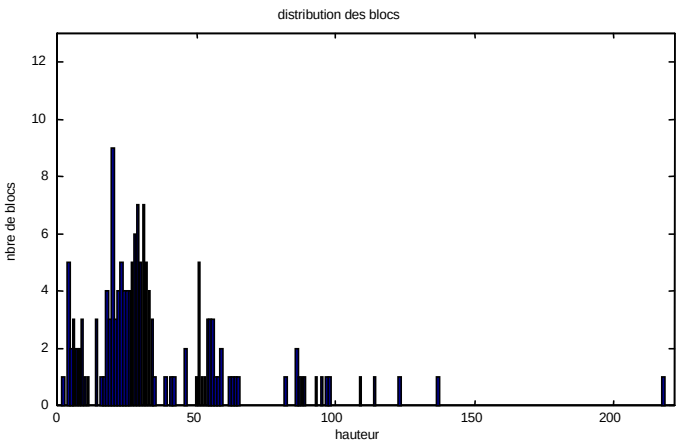


FIG. 5 - Lignes de texte



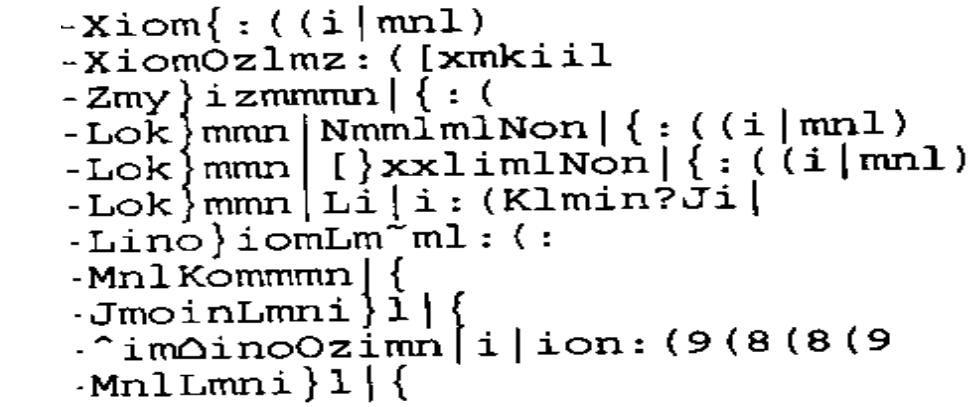


FIG. 9 - Lignes de texte

FIG. 6 - Image brute binarisée

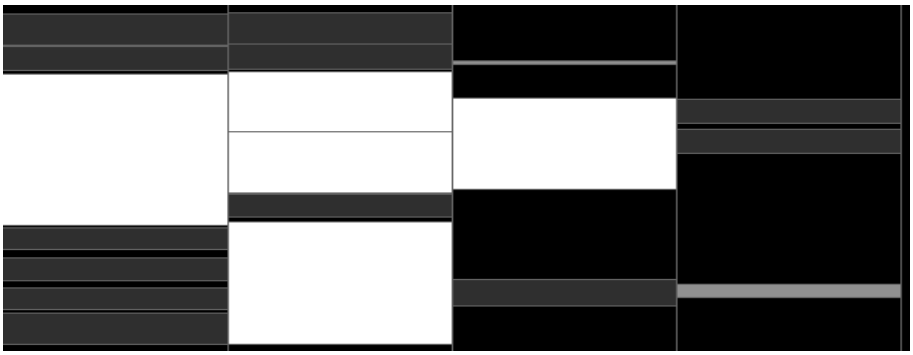


FIG. 7 - Classification des blocs

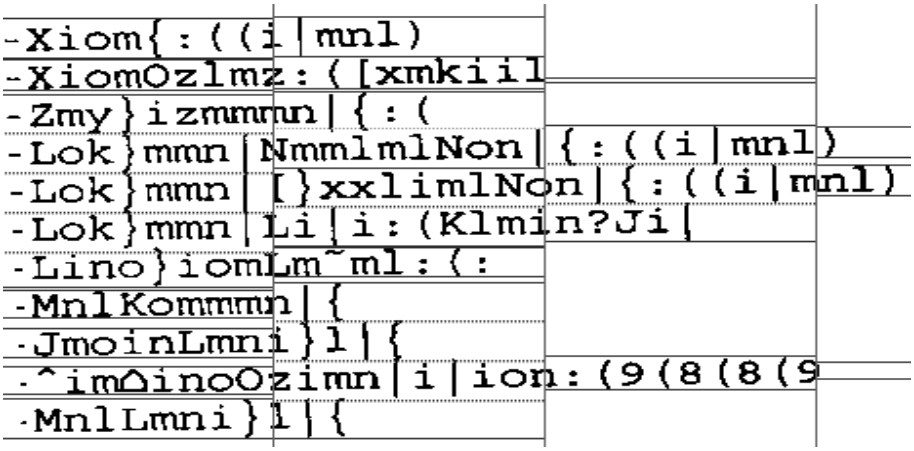


FIG. 8 - Segmentation des grands blocs