Segmentation texte /graphique : Application au manuscrits Arabes Anciens

Wafa Boussellaa, Abderrazak Zahour, Bruno Taconet, Abdellatif Benabdelhafid, Adel Alimi

Segmentation texte /graphique : Application au manuscrits Arabes Anciens

Wafa Boussellaa [1] – Abderrazak Zahour [2] – Bruno Taconet [2] – Abdellatif Benabdelhafid [2]– Adel Alimi[1]

[1] REsearch Group on Intelligent Machines (REGIM)
Université de Sfax, ENIS, DGE, BP. W-3038 - Sfax – Tunisie

[2] Université du Havre,
IUT du Havre, Place Robert Schuman, F-76 610 Le Havre

Wafa.boussellaa@gmail.com, adel.alimi@ieee.org
{abderrazak.zahour, bruno.taconet,benabdelhadid} @benuniv-lehavre.fr

Résumé *: Cet article présente une nouvelle méthode de segmentation d'images de documents couleur de type manuscrits arabes anciens. La méthode développée opère directement sur la luminance. L'analyse multi-échelle permet une séparation entre le fond et l'avant plan. Des caractéristiques statistiques ont extraites de l'avant plan obtenue et sont utilisées par l'algorithme de classification c-moyen floue pour la segmentation texte/graphique de l'avant plan. Notre méthode a été testée sur 50 images de documents manuscrits rares, à structure complexe, extraits d'une base de 2000 manuscrits de la Bibliothèque Nationale Tunisienne. Les tests menés montrent des résultats satisfaisants pour la segmentation avant/arrière plan. La segmentation de l'avant plan en texte/graphique reste à améliorer.*

Mots-clés *: Segmentation, ondelettes, c-moyen floue, fond/texte/graphique, manuscrit arabe ancien.*

1 Introduction

Les ouvrages anciens conservés dans la bibliothèque nationale de la Tunisie forme une bonne partie de son patrimoine culturel et scientifique. Le traitement automatique de ces documents en vue de leur restauration, indexation et exploitation offre un avantage certain. Cependant, on est confronté à de nombreuses difficultés dues au mauvais état de conservation de ces manuscrits et à la complexité de leur contenu. Les manuscrits composés de texte et de graphique forment une collection rare. Ils sont à structure complexe et ont de nombreuses particularités qui mettent en échec les algorithmes classiques de segmentation. La figure 1 illustre une variété de documents anciens à structure complexe.

FIG. 1- Images de documents manuscrits arabes anciens

A notre connaissance, peu de travaux concernant la segmentation texte/graphique des manuscrits anciens sont référencés dans la littérature. Le système DEBORA[1] a proposé une méthode pour la séparation texte/graphique des images de documents anciens du XVI^{ème} siècle en niveaux de gris. Cette approche est basée sur la morphologie mathématique [MUG 00] pour l'extraction des zones graphiques et la détection des zones texte. Des opérations d'érosion et de dilatation permettent de séparer ou de fusionner des formes particulières plus ou moins éloignées. Li et al [LI 00] et Menoti et al [MEN 03] ont proposé un algorithme de

[1] DEBORA: Digital AccEss to BOoks of the RenAissance.

segmentation d'images de documents en quatre classes : fond, texte, image, et graphique. Les caractéristiques utilisées pour la classification sont basées sur les modèles de distribution des coefficients d'ondelettes dans les bandes à haute fréquence. Chuai-Aree et al [ARE 01] ont développée une méthode de segmentation de l'image de document en trois classes : fond, texte et image. Cette technique se base sur l'algorithme de classification des C-moyen flous et utilise des caractéristiques statistiques. La méthode proposée par Hamza et al [SMI 05] est basée sur une classification par la carte organisatrice de Kohonen. L'algorithme des K-moyennes est utilisé pour le regroupement des classes. Quatre classes sont considérées: fond, lettrine, écriture avant plan, écriture arrière plan.

Cet article présente une nouvelle méthode de segmentation pour la séparation fond/texte/graphique pour les images manuscrits arabes anciens couleur. La méthode proposée opère en deux phases : utilisation de l'analyse multi-échelle pour la segmentation avant/arrière-plan, puis segmentation texte/graphique par l'algorithme des C-moyen flous. . Le reste de l'article décrit la méthode proposée et les résultats expérimentaux.

2 Description de la méthode

La méthode proposée réalise la segmentation fond/texte/graphique des images de documents anciens couleurs. L'image couleur originale en coordonnées RGB est d'abord transformée en coordonnées YIQ. La composante Y capture l'intensité lumineuse. Les deux canaux I et Q codent respectivement la teinte et la saturation. Les documents anciens présentent un contraste faible et une irrégularité de l'intensité lumineuse de l'arrière plan. Ces informations sont concentrées uniquement dans la composante Y. Notre choix pour cet espace de représentation exploite cette propriété. Une analyse multi échelle sur la base d'ondelettes de Daubechies 4 (db4) de niveau 1 permet l'extraction de l'arrière plan de l'image du document. Des caractéristiques statistiques basées sur la moyenne et de l'écart type de chaque bloc de pixels sont extraites de l'avant plan résultant. La segmentation texte/graphique est obtenue à l'aide du classifieur C-moyen flou (CMF). L'architecture générale du système est donnée dans la figure 2

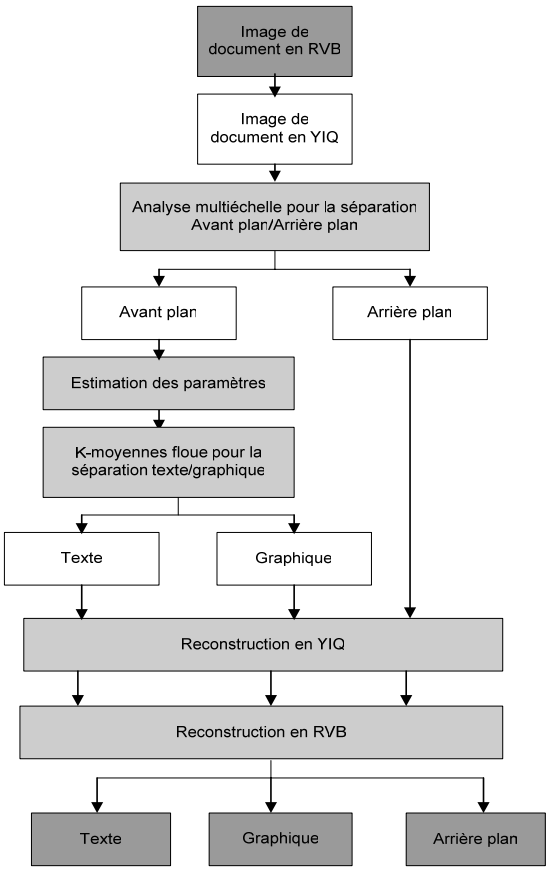


FIG. 2 – Processus général de segmentation

2.1 Extraction de l'arrière plan par analyse multi-échelle

2.1.1 Choix de l'ondelette

La transformée en ondelette est caractérisée par l'utilisation de fonctions bien localisées, à la fois dans le plan géométrique et dans le plan fréquentiel. Ces ondelettes sont générées par translation et dilatation de l'ondelette mère.

La transformée en ondelette correspond à une analyse multi-échelle de l'image. A chaque niveau d'analyse, les dimensions géométriques de l'image sont réduites dans un rapport de deux par un ensemble de filtres orthogonaux dont les caractéristiques sont imposées par la famille d'ondelette utilisée. Le résultat de l'analyse est composé de 4 imagettes : l'une représente la réduction de l'image source (image "approximation"), les 3 autres contiennent les informations hautes fréquences spatiales perdues lors de la réduction (images "détail"). La décomposition en ondelettes consiste à appliquer un filtre passe-bas (BB) permettant d'obtenir l'image d'approximation. Les détails diagonaux sont obtenus à l'aide d'un filtre passe haut (HH). La combinaison d'un filtre passe bas et d'un filtre passe haut (HB, BH) fournissent les détails horizontaux et verticaux (figure 3).

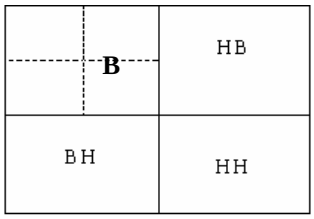


FIG. 3- Décomposition multi-échelle niveau 2 d'une image

La transformée d'ondelette est toujours implémentée sous forme d'un arbre binaire de filtres. Le choix de l'ondelette adaptée n'est pas aisé. Il convient de bien cerner le problème et d'identifier le type de transformée à utiliser (continue ou discrète). En analyse d'image, il est souvent utile d'avoir une certaine redondance pour avoir plus d'informations. L'utilisation de la transformée par ondelette continue est alors conseillée. Pour une analyse multi-résolution, on préfère une base d'ondelettes orthonormale. Parmi les familles d'ondelettes que nous avons testées, celle de Daubechies 4 (db4) d'ordre un répond parfaitement à notre étude. Elles sont à supports compacts et dissymétriques.

2.1.2. L'algorithme proposé

Dans le but de détecter la texture de l'arrière plan de l'image du document, une première approche consiste à appliquer l'ondelette de Daubechies récursivement sur la composante Y et analyser les imagettes obtenues à chaque niveau de décomposition. Deux difficultés sont inhérentes à cette approche : d'une part l'exploitation de l'ensemble des imagettes obtenues est une opération fastidieuse, et d'autre part, la profondeur du niveau de décomposition n'est pas connue. La démarche que nous avons adoptée consiste à décomposer la composante Y des intensités des pixels en blocs de taille 32x32, sur lesquels est appliquée directement la transformée d'ondelette de Daubechies 4 (db4) au niveau 1. Seules les détails horizontaux (HB) et verticaux (BH) sont analysés. Les détails diagonaux (HH) sont bruités. L'algorithme de segmentation par analyse multi-échelle opère en deux phases décrites dans la figure 4 :

– Premier niveau de segmentation par analyse des blocs 32x32 ;
– Affinement du résultat par décomposition de l'image Y de l'avant plan obtenu en bloc 4x4 et correction des blocs mal classés ;

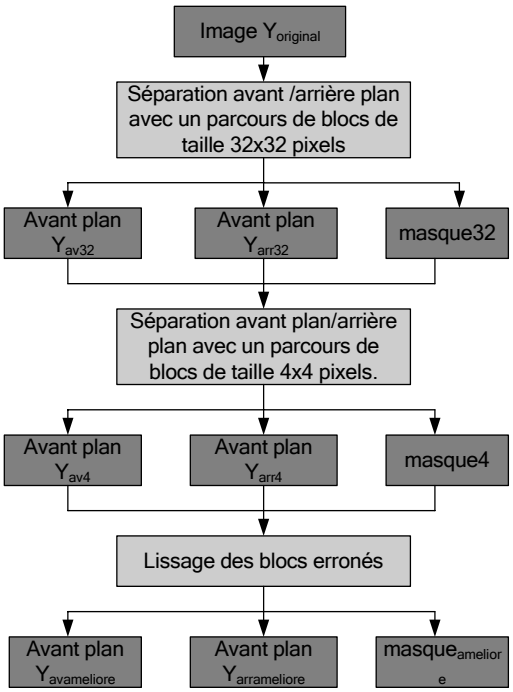


FIG. 4 - Analyse multiéchelle de l'image du manuscri

Phase 1 : Premier niveau de segmentation avant/arrière plan

Après décomposition de la matrice Y en blocs de taille 32x32 pixels, l'application de la transformation par ondelettes de Daubechies, limitée au niveau 1 à chaque bloc, permet d'obtenir les détails (BH) et (HB), qui recèlent l'information utile, comme le montre la figure 5.

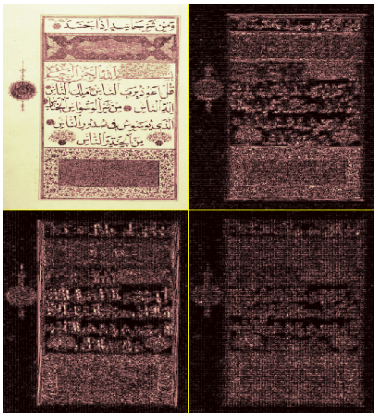


FIG. 5- Décomposition en ondelettes de Daubechies 4 niveau 1

Les coefficients d'ondelette des détails horizontaux et verticaux de chaque bloc sont fusionnés en un seul vecteur V. Une analyse minutieuse des histogrammes des coefficients de V sur plusieurs documents manuscrits a montré que ces histogrammes suivent une distribution normale centrée autour de zéro. En plus, nous avons constaté que la variance de cette distribution est très différente selon qu'il s'agit d'un bloc de texte, de graphique ou d'un bloc appartenant à l'arrière plan. L'idée intéressante était de collecter un échantillon

d'apprentissage significatif de blocs 32x32 pour le texte, un autre pour les blocs graphiques et un troisième échantillon pour les blocs appartenant à l'arrière plan. Nous avons utilisé pour cette étude 1000 blocs de taille 32*32 par échantillon issus de 50 manuscrits anciens à structure complexe. Les figures 6, 7 et 8 montrent quelques blocs issus de notre échantillon d'apprentissage.

FIG. 6 - Echantillon de la base de bloc fond

FIG.7 - Echantillon de la base bloc texte

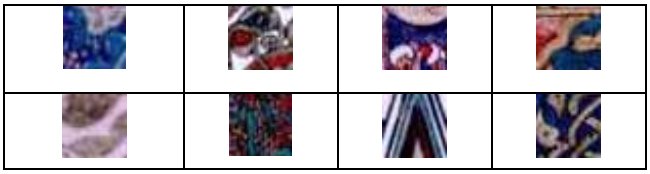
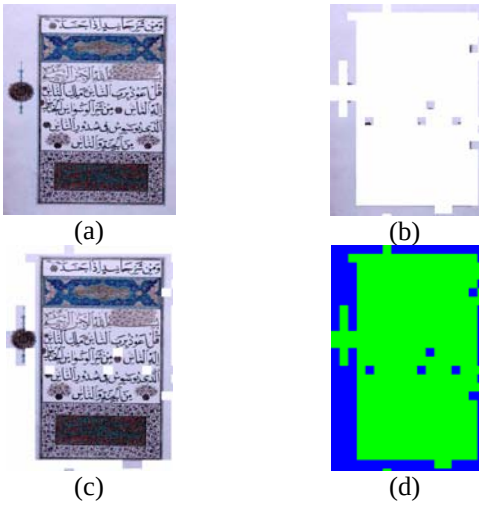


FIG. 8 - Echantillon de la base bloc graphique

Le calcul de la variance sur le vecteur V pour les trois échantillons montre que celle-ci varie entre 10^{-6} et 8.10^{-4}.pour les blocs de fond. La variance des blocs de textes varie entre 6.10^{-3} et 0,04. et celle des blocs graphiques se situe entre 4.10^{-3} et 0,2. On peut remarquer que les intervalles de variation de la variance se chevauchent pour les blocs texte/graphique. Par conséquent, il est très tôt pour opérer à une quelconque segmentation texte/graphique. Par contre la séparation est nette entre les blocs de l'arrière plan et les blocs texte/graphique. Est considéré arrière plan tout bloc dans la variance est inférieure ou égale à 8.10^{-4}. La figure 9 montre le résultat de cette étape sur un manuscrit ancien.



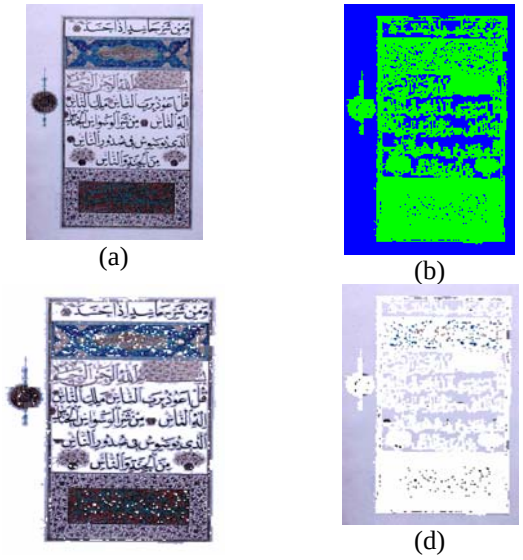
(a) (b) (c) (d)

FIG. 9 - Segmentation avant/arrière plan, block 32×32 : (a)Original , (b) Arrière plan, (c) Avant plan, (d) masque

A travers cet exemple, on remarque que certains blocs qui appartiennent normalement au fond sont mal classés (étiqueté avant plan). Ceci est dû en grande partie à la taille des blocs analysés (32x32) qui peuvent contenir un mélange de texture arrière/avant plan. D'après nos tests sur cent manuscrits anciens, les erreurs d'étiquetage concernent exclusivement les blocs du fond, qui se trouvent alors étiquetés avant plan. Dans le but d'améliorer les résultats de segmentation, nous construisons une image "masque" à partir de l'avant plan et de l'arrière plan obtenus à l'issue de cette première étape. Dans cette image masque, un bloc étiqueté avant plan est affecté d'une couleur verte et un bloc de l'arrière plan est coloré en bleu (figure 10). Ce masque est exploité puis mis à jour dans l'étape suivante de l'algorithme.

Phase 2 : Affinement de la segmentation avant/arrière plan

Dans l'étape précédente, plusieurs blocs de l'arrière plan sont mal classés. Dans le but de détecter ces blocs, la matrice de luminance Y de l'avant plan obtenue par la première phase du traitement est exploitée. Cette matrice est décomposée en blocs de taille 4x4 pixels. L'algorithme de la phase 1 est appliqué à nouveau, à l'identique, sur ces blocs. Le paramètre précédemment calculé sur nos échantillons d'apprentissage, de valeur 8.10^{-4}, est utilisé, exactement comme dans la phase 1, pour réaffecter les blocs mal étiquetés à leur classe d'origine (arrière plan). Le masque précédemment construit est mis à jour en modifiant la couleur de ces blocs qui devient bleue. La figure 10 montre les résultats obtenus. L'image RGB de l'arrière plan et de l'avant plan est construite à partir de ce masque.

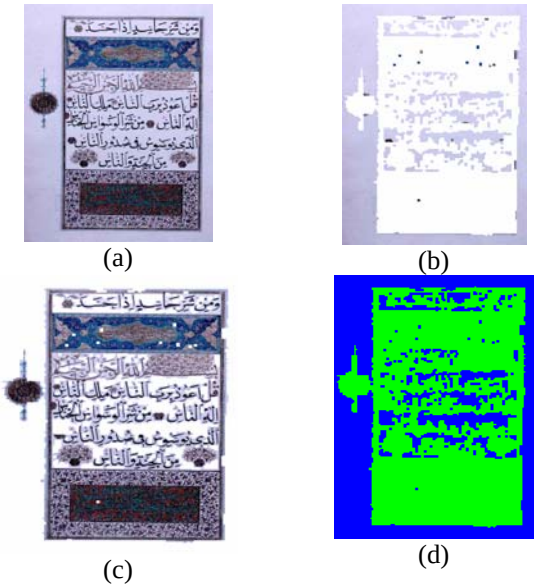


(a) (b) (c) (d)

FIG. 10 - Segmentation avant/arrière plan, block 4×4 : (a)Original , (b) Arrière plan, (c) Avant plan, (d) masque

Nous terminons cette phase par une analyse du 8-voisinage de la couleur des pixels dans le masque résultant. Si un pixel a une couleur bleue (respectivement verte) et si ses 8 voisins ont une couleur verte (respectivement bleue), alors le pixel est classé

comme appartenant à l'avant plan (respectivement arrière plan). Cette procédure est répétée jusqu'à stabilisation du masque. Les résultats de segmentation après correction par analyse du 8-voisinage sont montrés dans la figure 11.



(a) (b) (c) (d)

FIG. 11 - Correction segmentation avant/arrière plan : (a)Original, (b) Arrière plan, (c) Avant plan, (d) masque

2.2. Séparation texte/ graphique par le classifieur C-moyen flou (CMF).

L'algorithme C-moyen flou (fuzzy c-means) est un algorithme de classification floue fondé sur l'optimisation d'un critère quadratique de classification où chaque classe est représentée par son centre de gravité. L'algorithme nécessite de connaître le nombre de classes au préalable et génère les classes de telle sorte que les sommes des écarts quadratiques inter-classes et intra-classes soient respectivement maximales et minimales. La classification floue autorise le chevauchement des régions. Une segmentation floue peut être obtenue par affectation de chaque pixel à la classe pour laquelle son degré d'appartenance est maximal. Cette propriété est exploitée en traitement d'images et plus précisément en classification où les classes, appelées aussi régions, sont représentées par des ensembles flous. Cela est fort utile lorsque les régions ne peuvent pas être définies de manière nette et précise. Leur manipulation, en gardant le caractère flou, permet de traiter des données imprécises, incertaines et/ou redondantes d'une manière plus flexible.

Dans notre application, la matrice Y de l'avant plan obtenue après extraction de l'arrière plan est décomposée en n fenêtres de taille 32x32 pixels. Pour chaque fenêtre est calculé un vecteur caractéristique xi dont les composantes sont la moyenne est l'écart type. L'algorithme CMF opère à partir de ces vecteurs caractéristiques. Le résultat étant la partition des l'ensemble des fenêtres en deux classes : texte et graphique. Les valeurs des degrés d'appartenance sont regroupées dans une matrice $U = [u_{ik}]$ pour: $1 \le i \le n, 1 \le k \le c$ où u_{ik} désigne le degré d'appartenance de la fenêtre i à la classe k. Pour avoir une bonne partition, on impose aux éléments de U les contraintes suivantes : $u_{ik} \in [0,1]$ et $\sum_k u_{ik} = 1$; *ceci* $\forall \iota$.

L'algorithme du CMF fait évoluer la partition (Matrice U) en minimisant la fonction objectif suivante :

$$J_m(U,C) = \sum_i \sum_k (u_{ik})^m . \|x_i - c_k\|^2 ;$$

m >1 est un paramètre contrôlant le degré de flou (généralement m =2) et c_k : le centre de la classe k. Le principe de l'algorithme CMF est décrit ci-dessous :

1. Initialiser les centres c_k , initialiser aléatoirement la matrice de partition U.
2. Faire évoluer la matrice partition et les centres suivant les deux équations :

$$u_{ik} = 1/(\sum (d_{ik}/d_{ij})^{(2/(m-1))}),$$

// mise à jours des degrés d'appartenances, où :

$$d_{ij} = \|x_i - c_j\|,$$

$$c_k = (\sum_i (u_{ik})m . x_i / (\sum_i (u_{ik})^m)) ,$$

//mise à jours des centres.

3. Test d'arret : $\left| j^{(t+1)} - j^{(t)} \right| < seuil.$

La répartition des blocs selon leurs caractéristiques est montrée dans la figure 12.

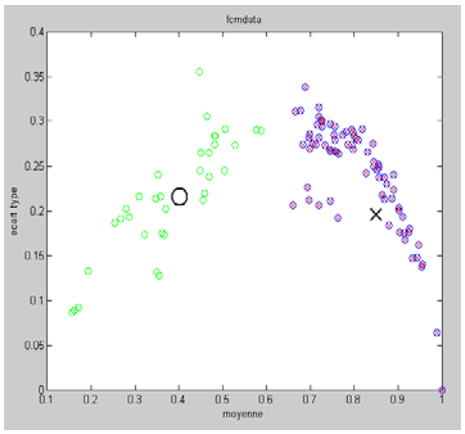


FIG. 12- Classification des blocs de l'avant plan en deux classes : texte (en vert) et graphique (en vert).

Les figures 13 et 14 présentent les résultats de séparation texte et graphique sur des images de documents anciens à structure complexe extraites de notre base.

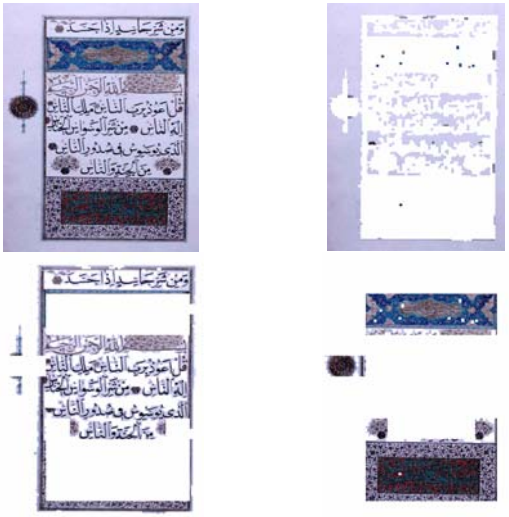


FIG. 13- Segmentation texte/graphique par CMF : (a)Original , (b) fond, (c) texte, (d) Graphique

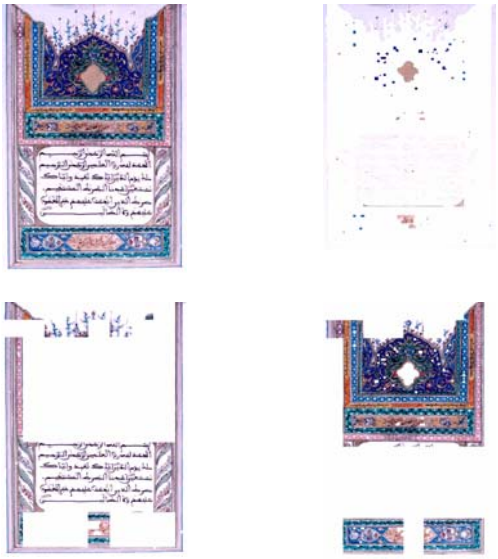


FIG. 14 - Segmentation texte/graphique par CMF : (a)Original , (b) fond, (c) texte, (d) Graphique

A travers ces exemples, nous remarquons que la séparation n'est pas parfaite : tous les blocs texte sont correctement étiquetés, mais certains blocs graphiques sont mal classés. Il reste donc à affiner le module de segmentation texte/graphique.

3 Conclusion et perspectives

Dans cet article, nous avons présenté une méthode de segmentation fond/texte/graphique en deux étapes : l'extraction du fond, puis la segmentation texte/graphique.

L'extraction du fond (arrière-plan) est réalisée en deux étapes. Une première transformation par ondelette de Debauchies db4, au niveau 1, est appliquée sur de grands blocs de la matrice Y de l'image originale. On obtient une image pour le fond et une autre pour l'avant-plan. Une phase d'affinement sur des blocs de petite taille par la même transformation effectuée sur l'avant-plan. Une correction finale par une analyse du 8-voisinage achève cette première phase de segmentation avec de très bons résultats.

La deuxième phase concerne la séparation texte/graphique. Elle commence par l'extraction d'un vecteur caractéristique (moyenne, écart-type) sur les blocs de taille 32x32 pixels de la matrice de luminance de l'avant-plan. L'algorithme c-moyen flou partitionne l'ensemble des blocs en deux classes ; tous les blocs de la classe texte sont correctement étiquetés, mais certains blocs de la classe graphique sont mal classés. Les graphiques mal classés ont un aspect visuel texturé proche de celle de la calligraphie, ce qui explique ces premiers résultats, au demeurant dans l'ensemble très encourageants.

4 Bibliographie

[MUG 00] Muge F., Granado I., M. Mengucci, Pina P., Ramos V., Sirakov N., Caldas Pinto J.R., A. Marcolino, Mário Ramalho, P. Vieira, A. Maia do Amaral, "Automatic Feature Extraction and Recognition for Digital Access of Books of the Renaissance", *Proc. of ECDL'2000 - 4th European Conference on Research and Advanced Technology for Digital Libraries*, J.Borbinha and T. Baker (Eds.), ISBN 3-540-41023-6, Lecture Notes in Computer Science, Vol. 1923, , Springer-Verlag -Heidelberg, Lisbon, Portugal, , pp. 1-13, 18-20 Sep. 2000.

[LI 00] Li J., Gray M., "Context based Multiscale classification of document wavelets coefficients distributions", *IEEE Transactions on Image Processing*, Vol. NO. 9, September 2000.

[MEN 03] Menoti D., Borges D., Facon L., J. , Britto Jr A. S., "Segmentation of Postal Envelopes for Address Block Location: an approach based on feature selection in wavelet space", *IEEE proceedings of the Seventh International Conference on Document Analysis and Recognition (ICDAR 2003)*, 0-7695-1960-1/03, 2003.

[ARE 01] AREE C., LURSINSAP C., SOPHASATHIT P., SIRIPANT S., "Fuzzy C-MEAN: a statistical feature classification of text and image segmentation method", *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, Vol. 9, No. 6, 661-671 S, 2001.

[SMI 04] Smigiel1 E., Belaid A., and Hamza H., "Self-organizing Maps and Ancient Documents", S. Marinai and A. Dengel (Eds.): DAS 2004, LNCS 3163, pp. 125–134, Springer-Verlag Berlin Heidelberg, 2004.

[MAL 89] Mallat S., "A Theory for Multiresolution Signal Decomposition: the Wavelet Representation", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, IEEE, USA, vol. 11, n.7, pp. 674-693,1989.