Une méthode de segmentation d'Images de Documents Composites

Mohamed Ben Jlaiel, Slim Kanoun, Adel Alimi

Une méthode de segmentation d'Images de Documents Composites

Mohamed Ben Jlaiel, Slim Kanoun, Adel M. Alimi

Groupe de **RE**cherche sur les **M**achines **I**ntelligentes (**REGIM**)

Université de Sfax, Ecole Nationale d'Ingénieurs de Sfax, BP. W-3038 - Sfax – Tunisie

benjlaiel@yahoo.fr, slim.kanoun@enis.rnu.tn, Adel.Alimi@enis.rnu.tn

Résumé : *Dans cet article, nous proposons une méthode de segmentation d'images de documents composites. Cette méthode commence par regrouper les objets connexes proches les uns par rapport aux autres dans des microstructures homogènes (texte, tableau, etc.). Elle extrait ensuite les différentes microstructures du document et sépare après entre celles qui sont textuelles des non textuelles (graphiques). La méthode proposée catégorise les microstructures graphiques (tableau, cercle, etc.) et localise leur contenu textuel. L'étude expérimentale a été effectuée sur 50 documents contenant de l'écriture Arabe (fictif, réel, simple et composite). Le taux de segmentation correcte globale était de l'ordre de 89.75 %.*

Mots-clés : Documents composites, segmentation.

1 Introduction

Face à la masse très importante d'informations échangées entre les différentes organisations, le besoin de systèmes permettant la reconnaissance, l'indexation, la recherche et la classification automatique de documents de natures imprimée et manuscrite croît continuellement. La plupart des travaux de rétro conversion d'images de documents arabes imprimés se sont limités à la reconnaissance des blocs textuels sans trop aborder les documents composites (lettres d'information, Lettre d'invitation, tout type de demandes, etc.) et qui sont composées, en général, par plusieurs blocs hétérogènes (textes, graphiques, tableaux, logos, photographies, etc.). L'analyse des structures de tels documents pose des difficultés dans le domaine de l'analyse et la reconnaissance d'image de documents.

Dans le cadre de cet article, nous focalisons notre intérêt essentiellement sur la segmentation d'images de documents composites en microstructures. Nous envisageons comme perspectives l'intégration et le développement de différents systèmes de différentiation et de reconnaissance d'écritures afin d'aboutir à un système complet de reconnaissance et d'indexation d'images de documents.

Dans la suite, nous présenterons les techniques existantes pour la segmentation d'images de documents. Nous exposerons ensuite notre démarche pour la segmentation d'images de documents composites ainsi que les résultats expérimentaux obtenus sur une base de 30 documents simples et 20 autres composites contenant de l'écriture Arabe. Ce choix de la langue Arabe est justifié par le fait que notre système est conçu pour faire partie d'une chaîne d'analyse et de reconnaissance des documents Arabes imprimés composites. Nous détaillerons après les différents systèmes qui peuvent être intégré dans le système proposé pour aboutir à la reconnaissance et l'indexation de ces documents. Nous achèverons ainsi notre article par une conclusion et des perspectives

2 Techniques existantes de segmentation d'images de documents composites

Dans la dernière décennie, plusieurs travaux ont été proposés pour la segmentation d'images de documents. Parmi les premières techniques, nous citons la technique RLSA Run Lengh Smoothing Algorithm proposée dans [WON 82] qui consiste à faire un double lissage unidirectionnel de l'image à segmenter selon deux seuils. Le premier est égal à la moyenne des espaces inter mot dans le cas du lissage horizontal et le deuxième seuil est égal à l'espace interligne dans le cas du lissage vertical. La segmentation est obtenue en appliquant l'opérateur logique "and" sur les deux images résultant respectivement d'un lissage horizontal et d'un lissage vertical. Il existe aussi d'autres versions améliorées dans [TAK 93].

Dans [AKI 87], la découpe récursive est une autre technique de segmentation procédant par découpe récursive alternant l'analyse des profils horizontaux avec celle des profils verticaux.

L'analyse des espaces est une technique de segmentation fondée sur une analyse des espaces inoccupés et qui consiste à fusionner les segments d'espaces blancs adjacents dans le but de construire des plages blanches qui serviront à estimer l'inclinaison des images ainsi qu'à segmenter ces dernières. Dans [AKI 93], l'auteur propose une technique dans laquelle les plages blanches sont supposées rectangulaires puisque les images traitées ont été préalablement redressées.

Dans [BAI 90], l'auteur propose une technique de segmentation qui consiste à déterminer dans un premier temps les blocs en se basant sur une analyse des espaces puis dans un second temps la structure des blocs textuels au moyen d'une technique de découpe

récursive. La particularité de son approche réside dans le fait qu'elle est non seulement indépendante de l'inclinaison mais aussi capable de segmenter des blocs mosaïques. L'analyse structurelle est une catégorie qui regroupe l'ensemble des techniques de segmentation guidées par des règles structurelles décrivant le but à atteindre.

Dans [KRI 93], la technique proposée consiste à subdiviser les images suivant une description des profils de projection générique associées à chaque type d'entités pouvant composer les images traitées. Les profils sont décrits au moyen d'une suite alternant les plages noires et les plages blanches chaque plage étant représentée par sa longueur.

Dans [NAG 00], l'auteur souligne que les fusions erronées de colonnes juxtaposées ont des conséquences désagréables sur la post-édition des sorties des OCR.

Certains sont spécialisées sur des types de mises en page : les journaux, les tableaux [KIE 98], les sommaires [LEB 00]. Il existe aussi des méthodes développées pour s'affranchir des défauts qui gênent la détection des colonnes comme l'inclinaison de la feuille [PAV 92].

D'autres auteurs [LEE 01] et [JAI 98], proposent comme résultats de la segmentation physique une description en termes de blocs homogènes (texte, image, ligne graphique isolée).

Dans [BEL 04] un des problèmes majeurs du rétro conversion est qu'une page de document représente une structure *physique*, alors que le but de l'analyse et de la reconnaissance est d'aboutir à une description logique du document, la plus proche possible de celle utilisée par les outils d'édition. Il faut donc résoudre deux sous problèmes :

- Trouver la structure physique, c'est-à-dire décomposer l'image de document en entités structurelles homogènes : caractères, blocs de texte, primitives graphiques, etc.

- Passer de la structure physique à la structure logique, c'est-à-dire retrouver en quelque sorte la « sémantique » du document.

L'analyse des techniques présentées ci-dessus montre plusieurs insuffisances. La première insuffisance est liée au choix arbitraire des seuils de lissage, la sensibilité aux inclinaisons et l'inadaptation à la segmentation des blocs graphiques, formules et tableaux. La deuxième s'articule autour de la segmentation et l'analyse de documents à structures complexes comme les journaux ne donnent des résultats satisfaisants [ANT 03] [HAD 03] [HAD 01] qu'après l'intervention d'un utilisateur humain.

Dans le cadre de cet article, nous proposons une méthode de segmentation des documents composites qui vient palier quelques limites des méthodes présentées précédemment et notamment RLSA et la découpe récursive. Les blocs textuels issus de cette opération seront exploités par les système de reconnaissance de l'écriture arabe [KAN 05]. Notre méthode est insensible aux inclinaisons et adaptée à la segmentation des blocs graphiques, et des tableaux.

3 Démarche proposé pour la segmentation de documents composites

La figure 1 récapitule la démarche proposée pour la segmentation de documents composites.

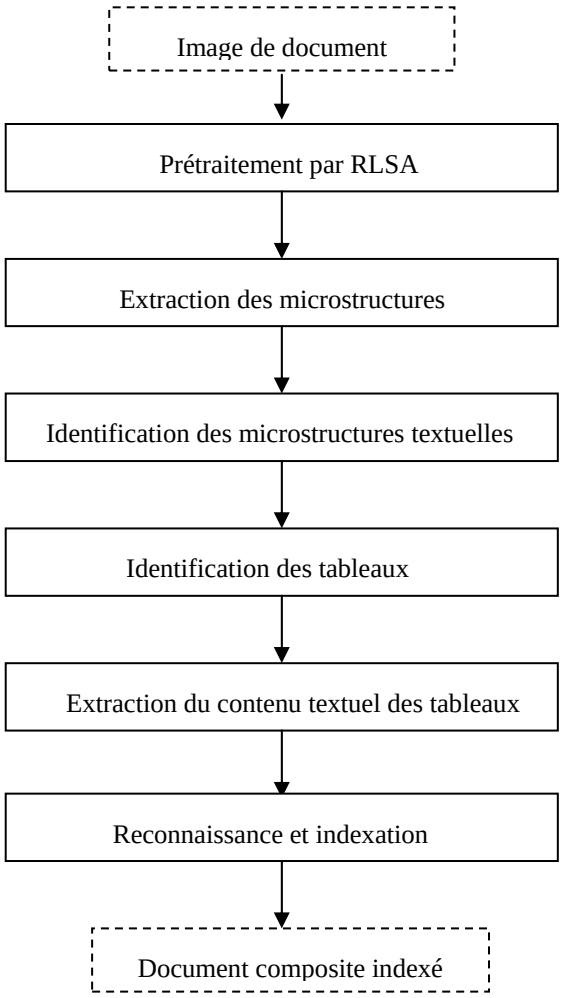


FIG.1- Démarche proposée pour la segmentation de documents composites

3.1 Prétraitement RLSA [WON 82]

En partant d'une image de document composite, nous appliquons la technique Run Lengh Smoothing Algorithm (RLSA) qui consiste à faire un double lissage unidirectionnel de l'image de documents à segmenter selon les seuils indiqués précédemment. La segmentation est obtenue en appliquant l'opérateur logique "and" sur les deux images résultant respectivement d'un lissage horizontal et d'un lissage vertical [WON 82] et [TAK 93]. La figure 2 illustre l'application de la technique RLSA sur un exemple de document et montre bien la fiabilité de cette technique au niveau du regroupement d'objets connexes proches les uns par rapport aux autres dans un bloc de texte, tableau, etc.

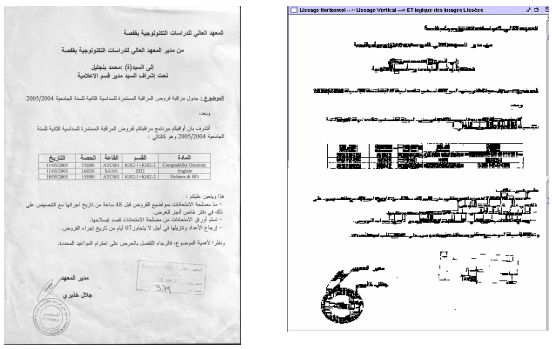


FIG. 2 - Application de la technique RLSA sur un exemple d'image de document composite

3.2 Extraction des microstructures

En partant de l'image de document composite après prétraitement par la technique RLSA, nous procédons à la segmentation de l'image de document en microstructures en se basant dans un premier temps sur l'analyse du profil de projection horizontale (figure 3). Nous appliquons ensuite dans un deuxième temps sur chaque microstructure issue de la dernière segmentation l'analyse du profil de projection verticale afin de fiabiliser davantage la segmentation de l'image de document. Comme exemple montrant l'intérêt de la dernière analyse, nous pouvons citer la présence très proche d'un texte à coté d'un tableau

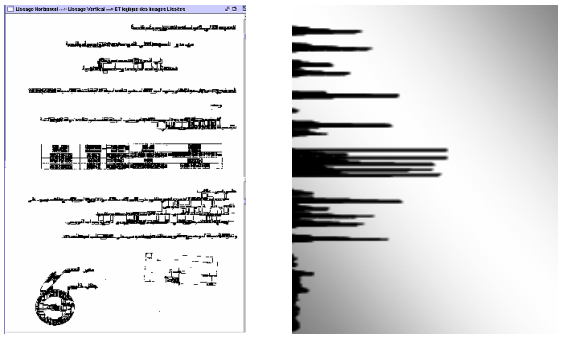


FIG. 3 - Profile de projection horizontale pour un exemple d'image de document après un prétraitement par RLSA

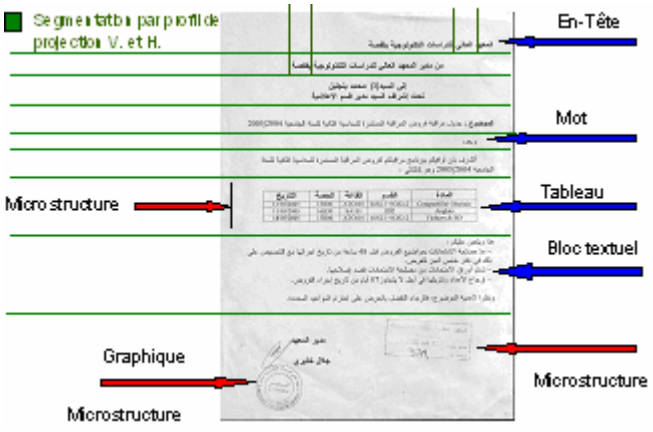


FIG. 4 - Exemples de microstructures d'une image de document

3.3 Identification des microstructures textuelles

Comme le montre la figure 4, une microstructure extraite d'une image de document composite peut être un bloc textuel, un tableau, un graphique, etc. Pour identifier les microstructures textuelles (figure 5), nous appliquons dans un premier temps l'analyse du profil de projection horizontale sur chaque microstructure extraite à partir de l'image du document d'origine (sans prétraitement). La présence de plusieurs séquences de lignes blanches de pixels montre éventuellement des espaces interlignes pour un bloc de texte. Pour s'assurer qu'il s'agit effectivement d'un bloc de texte, nous appliquons dans un deuxième temps l'analyse du profil de projection verticale pour chaque ligne de texte issue de la première analyse. La présence de plusieurs séquences de colonnes blanches de pixels montre certainement des espaces inter mots. Evidemment, la non présence d'espaces interlignes et inter mots confirme qu'il ne s'agit pas d'une microstructure textuelle. La méthode d'identification des microstructures non textuelles sera détaillée dans la section suivante.

FIG. 5 - Exemple de microstructure textuelle

3.4 Identification de graphiques comportant des lignes droites

Une microstructure non textuelle pourrait être essentiellement un tableau ou un graphique (un logo, un tampon, une photo, …). Pour différencier les tableaux des autres types de microstructures non textuelles, nous appliquons la transformée de radon [BRA 95] [LIM 90] puisqu'elle permet de détecter la présence de lignes droites.

Dans le cas d'un tableau (figure 6), la transformée de Radon retourne des pics sous forme de points et qui signalent la présence des lignes droites. La signature des formes des tableaux présente des pics sous formes de lignes verticaux correspondants aux cadres du tableau. Par contre, pour les microstructures qui ne présentant pas des lignes droites, la transformée de Radon retourne une enveloppe externe ne présentant aucun pic et notamment pour le cas d'un cercle (figure 7). Ceci explique l'absence de lignes droites et par conséquent l'absence de tableaux.

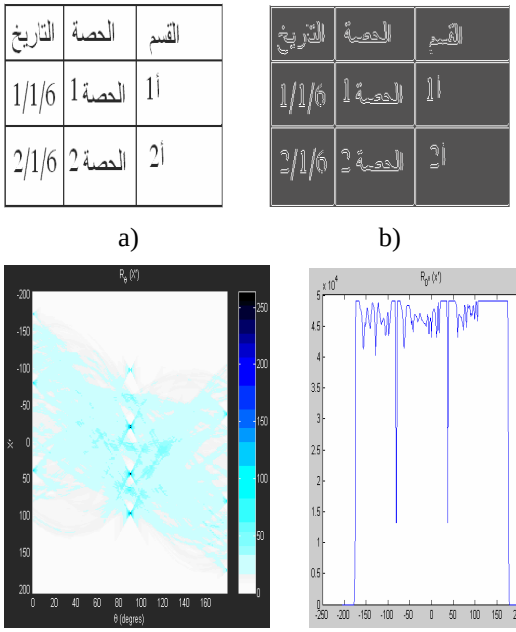


FIG. 6- Phases de détection des lignes droites :
a) Image originale d'un tableau
b) Extraction des bordures
c) Calcul de la transformée de Radon
d) Signature de forme d'un tableau

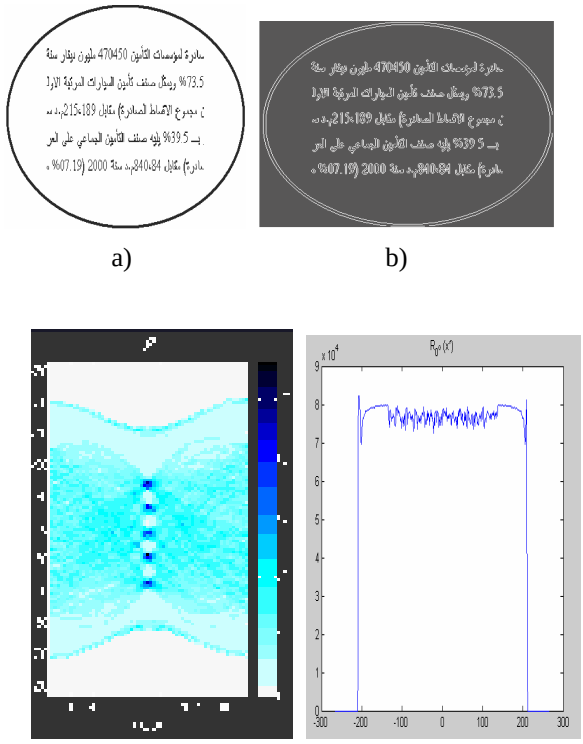


Fig. 7 - Phases de détection des cercles
a) Image originale d'un cercle
b) Extraction des bordures
c) Calcul de la transformée de Radon
d) Signature de forme d'un cercle

Dans notre exemple fig. (8), les pics maximum de la transformée de radon correspondent à θ=90 et x'=0 et θ=90 et x'=-48 etc.

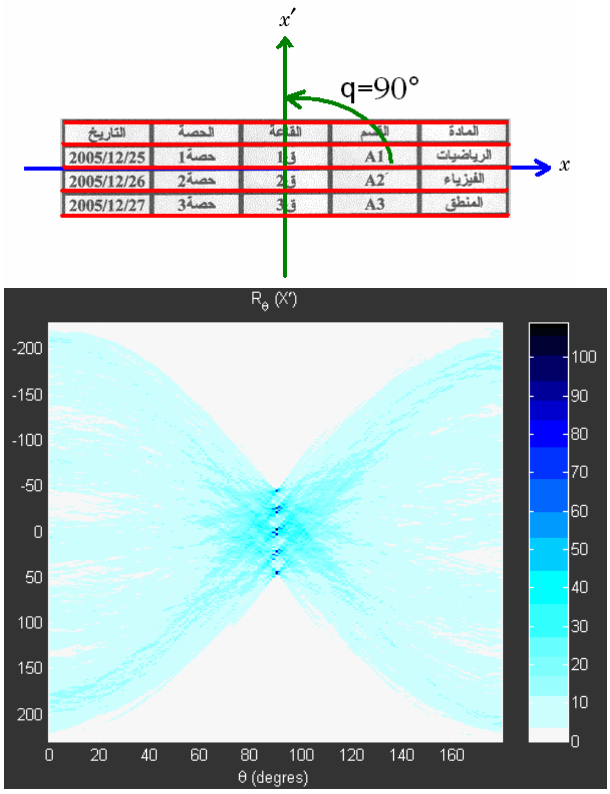


FIG. 8- Calcul des coordonnées des lignes droites

Les lignes droites perpendiculaires à x' sont tracées en rouge et superposées sur les lignes droites de l'image originale, les lignes de la transformée de radon sont tracés en bleu pour l'axe des x et en vert pour l'axe x'.

3.5 Extraction du contenu textuel à partir de graphiques

Pour extraire le contenu textuel des graphiques, nous proposons une technique basée sur la segmentation par approche région.

La séparation en zones de texte et en zones graphiques peut s'avérer très compliquée surtout lorsqu'une zone graphique est constituée d'éléments textuels.

Notre approche d'extraction des entités textuelles figure (9) est réalisée par l'algorithme suivant :

- Binarisation de l'image;
- Lissage horizontal et vertical de l'image binaire par RLSA;
- Et logique entre les deux images;
- Segmentation par alternance des profils de projection horizontaux et verticaux;
- Extraction des entités textuelles;
- Test de l'existence de blocs mosaïques;
- Cas d'un tableau :

- Lancer le module d'étiquetage;
- Repérer les sommets de l'histogramme du profil de projection horizontale;

- Calcul des densités des étiquettes des sommets de l'histogramme; {Etiquettes du cadre}
- Extraction du cadre et des données textuelles;

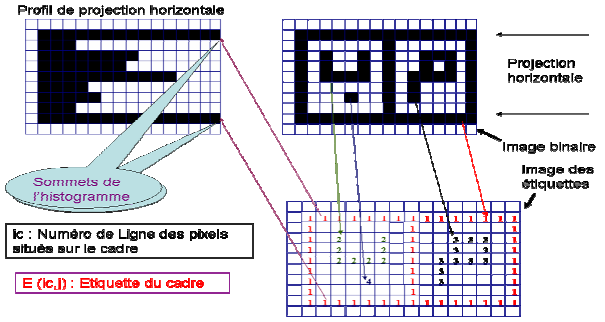


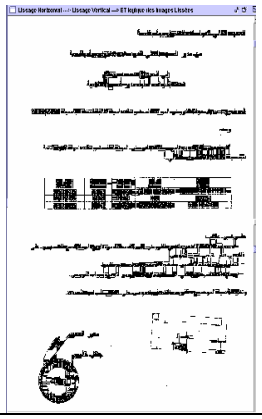
FIG.9 - Etiquetage et extraction des étiquettes du cadre

3.6 Résultats expérimentaux

Nous considérons un document qui ne comporte que du texte comme étant un document simple par opposition à un document composites qui pourra comporter en plus du graphique. Les documents fictifs sont des documents issus d'un logiciel de traitement d'images par contre les documents réels sont des documents administratifs numérisés.

Pour pouvoir expérimenter notre démarche, nous avons constitué une base d'images de documents composites. Cette base contient deux types de documents : 30 documents fictifs (15 documents simples et 15 autres composites) et 20 documents réels (10 documents simples et 10 autres composites). Les figures figure 6a-a, figure 6b-a et la figure 4 montrent des exemples de ces documents.

La figure 10 illustre notre démarche de segmentation sur un document composite réel. La table 1 expose les taux de segmentation correcte obtenus.

Image Originale	Image lissée
	

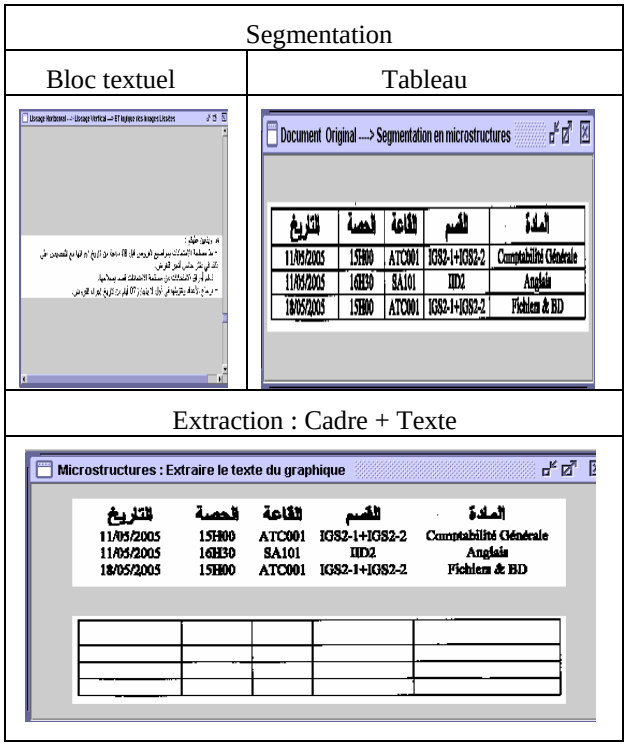


Fig.10 - Démarche de segmentation

	Taux de segmentation correcte	
Documents	Documents simples	Documents composites
Fictifs	99%	89%
Réels	90%	81%

TAB 1 - Taux de segmentation correcte

Notons que la qualité et la fiabilité de notre démarche de segmentation est sensible aux bruits d'acquisition de l'image et aux différentes déformations ainsi le degré de complexité des documents.

4. Conclusion et perspectives

Dans cet article, nous avons proposé une méthode de segmentation d'images de documents composites. Cette méthode consiste à regrouper les objets connexes et proches les uns par rapport aux autres dans des microstructures homogènes (texte, tableau, etc.). Elle extrait ensuite les différentes microstructures du document et sépare celles qui sont textuelles de celles qui sont non textuelles (graphiques). La méthode proposée catégorise les microstructures graphiques (tableau, cercle, etc.) et localise leur contenu textuel. Nous avons expérimenté cette méthode sur 50 documents contenant de l'écriture Arabe (fictif, réel, simple et composite). Le taux de segmentation correcte globale était de l'ordre de 89.75 %.

A ce niveau, notre méthode à sa version actuelle permet de segmenter une image de document composite et d'extraire le contenu textuel.

Notons que la plupart des documents composites manipulés dans les pays Arabes sont des documents bilingues utilisant des alphabets Arabe et Latin. Ainsi, pour pouvoir utiliser concrètement notre méthode pour mettre en place un système de reconnaissance et d'indexation de documents composites bilingues utilisant des alphabets Arabe et Latin, il faudra lui intégrer un module de différentiation entre l'écriture Arabe et l'écriture Latines de natures imprimé et manuscrite ainsi qu'entre les entités textuelles et les chiffres et un module de reconnaissance de l'écriture Arabe et Latine et des systèmes de reconnaissance de chiffres.

References

[AKI 93] AKINDELE O.T, BELAID A., Page segmentation by segment tracing, ICDAR'93, 1993, pp.341-344.

[AKI 87] AKIYAMA T., MASUDA I., A method for document image segmentation based on projection profiles, stroke densities and circumscribed rectangles, Systems and Computers in Japan 18(4),1987,pp-101-111.

[ANT 03] ANTANACOPOULOS A., GATOS B., KARATZAS D., Page Segmentation Competition, ICDAR'2003, 2003, pp. 688-692.

[BEL 04] BELAID A., EMPTOZ H., VIGNAUX G. RTP 33 :" Document et contenu; création, indexation, navigation", action spécifique 96 : " Numérisation et valorisation"

[BRA 95] BRACEWELL, RONALD N., Two-Dimensional Imaging. Englewood Cliffs, NJ: Prentice Hall, 1995, pp. 505-537.

[HAD 03] HADJAR K., INGOLD R., Arabic Newspaper Page Segmentation, ICDAR'03, 2003, pp. 895-899, Août.

[HAD 01] HADJAR K., HITZ O., INGOLD R., Newspaper Page Decomposition using a Split and Merge Approach, ICDAR'01, 2001, pp. 1186- 1189.

[JAI 98] JAIN, A.K., Yu, B., "Document Representation and Its Application to Page Decomposition". IEEE trans. on PAMI, Vol. 29, N. 3, pp. 294-308, 1998.

[KAN 05] KANOUN S., ALIMI M.A., LECOURTIER Y., *Affixal* Approach for Arabic Decomposable Vocabulary Recognition: A Validation on Printed Word in Only One Font, 8th IAPR - International Conference on Document Analysis and Recognition (ICDAR'2005), pp. 1025 - 1029, 29 August – 1 September, 2005, Seoul, Korea.

[KIE 98] KIENINGER, T. G., DENGEl, A., "A paper-to-html table converting system". Actes *Document Analysis Systems* (DAS98), Japan, 1998.

[LEB 00] LEBOURGEOIS, F., EMPTOZ, H., Vigne, H., "RASADE / Reconnaissance Automatique des Structures Associées aux Documents Ecrits", *Conf. Int. Francophone sur l'Ecrit et le Document* (CIFED'00), Lyon, pp. 281-294, 2000.

[LEE 01] Lee, S.-W., RYU, D.-S., "Parameter-Free Geometric Document Layout Analysis".*IEEE-PAMI*, Vol. 23, N. 11, pp. 1240-1256, 2001.

[LIM 90] LIM, JAE S., Two-Dimensional Signal and Image Processing. Englewood Cliffs, NJ: Prentice Hall, 1990, pp. 42-45.

[NAG 00] NAGY, G., "Twenty Years of Document Image Analysis". in PAMI. IEEE-PAMI,Vol. 22, N. 1, pp. 38-62, 2000.

[PAV 92] PAVLIDIS, T., ZHOU, J., "Page Segmentation and Classification". *Graphical Models and Image Processing*, Vol. 54, N. 6, pp. 484-496, 1992.

[TAK 93] TAKASHI S., MICHIYOSHI T., TOSHIFUMI Y., Document image segmentation and text area ordering, ICDAR'93, 1993,pp. 336-340.

[WON 82] WONG K.Y., CASEY R.G., WHAL F.H., Document analysis system, IBM J. Res. Dev., 1982,pp-26.6.647.646.