Segmentation de documents composites par une technique de recouvrement des espaces blancs

Yves Rangoni, Abdel Belaid

Segmentation de documents composites par une technique de recouvrement des espaces blancs

Yves Rangoni, Abdel Belaïd

LORIA
Campus scientifique
BP 239
F-54506 Vandœuvre-lès-Nancy Cedex
{rangoni, abelaid}@loria.fr

RÉSUMÉ. *Nous présentons dans cet article une méthode pour la segmentation de documents composites. Contrairement à la majorité des publications, nous nous focalisons sur des documents à structure non-Manhattan qui sont généralement créés par montage. Les pages à traiter contiennent donc plusieurs sous-documents qu'il faut isoler. Nous nous inspirons d'une technique par recouvrement d'espaces blancs proposée par Baird et al. ainsi qu'une suite de pré-traitements et post-traitements spécifiques à ces documents particuliers. Les évaluations sont faites sur des documents administratifs d'origines diverses qui nous sont fournis par une société partenaire. Ne disposant pas de documents de vérité, nous avons comparé nos résultats à ceux d'OCR commerciaux que notre méthode surpasse.*

ABSTRACT. *We present here a method for the segmentation of composite documents. Unlike most publications, we focus on non-Manhattan layouts which are usually created by compositing. Therefore, the pages to be processed contain several sub-documents which have to be isolated. We draw inspiration from the white space cover technique introduced by Baird et al. and a suite of pre- and post-processings specific to these particular documents. The evaluations are made on administrative records coming from various sources and provided to us by our industrial partner. As we do not have any groundtruth documents we compared our results with those obtained by a commercial OCR which is outperformed by our method.*

MOTS-CLÉS : *Analyse d'image de document, segmentation géométrique, recouvrement par espaces blancs, pré-traitements*

KEYWORDS: *Document image analysis, geometrical segmentation, white space cover, preprocessings*

1. Introduction

Nous traitons dans cet article d'un problème de segmentation physique de document. Par segmentation physique on parlera du découpage géométrique d'une page en plusieurs régions dites homogènes. Bien que ce travail soit pensé comme une étape préliminaire à la segmentation logique du document (ajout d'information sur le contenu des zones : titres, paragraphes, photographies, etc. ainsi que l'ordre de lecture), nous nous arrêtons donc à la subdivision en zones plus ou moins grandes d'une page de document.

La segmentation physique est un problème étudié depuis longtemps et deux grandes familles de stratégies s'affrontent pour résoudre le problème : les approches ascendantes et celles descendantes. Les méthodes les plus souvent citées et toujours utilisées sont :

– le XY-cut de Nagy (Nagy *et al.,* 1992). Une approche descendante qui subdivise récursivement le document en s'aidant des profils des projections verticales et horizontales ;

– le white-space-cover de Baird (Baird *et al.,* 1990) et le white streams de Pavlidis (Pavlidis *et al.,* 1991). Des approches aussi descendantes qui utilisent un recouvrement par rectangles maximaux de l'arrière plan du document ;

– le smearing de Wahl (Wahl *et al.,* 1982), approche ascendante, consistant à connecter par morphologie mathématique les composantes connexes proches ;

– le Docstrum de O'Gorman (O'Gorman, 1993), une approche ascendante, qui part des composantes connexes de la page, sépare celles du texte et celles du bruit, puis effectue un regroupement par un KNN et une mesure d'angle entre chaque centroides pour détecter les alignements et donc les lignes de texte ;

– la méthode de (Kise *et al.,* 1998) utilisant un diagramme de Voronoï. La tesselation sert à construire des graphes de composantes connexes qui ensuite, en se basant sur l'aire et la distance, permettent de retrouver les frontières des régions de texte.

La difficulté majeure ici sera de faire face à la grande diversité de types de document à segmenter : tant au niveau du contenu que de la qualité de numérisation (Fig. 1). Ces documents sont majoritairement de type administratif (factures, formulaires, extraits d'actes de naissance ou des lettres). Beaucoup font intervenir des annotations manuscrites faites a posteriori ou contiennent par nature un mélange d'écritures imprimées et manuscrites.

L'autre grande difficulté sera de traiter ce que nous appelons des documents composites. Ce sont des documents créés artificiellement par montage de petits documents indépendants sur la même page blanche (Fig. 5 et Fig. 6). Ils sont quasiment exclusivement produits par des particuliers avec tous les problèmes inhérents à ce type de manipulation : orientation, occlusion partielle, débordements, montage par photocopies successives et additive des éléments. Les variations de qualité de numérisation, de contraste, et de bruit sont assez extrêmes. L'enjeu à long terme sera de localiser, extraire et redresser tous les sous-documents d'une page afin de les présenter de la

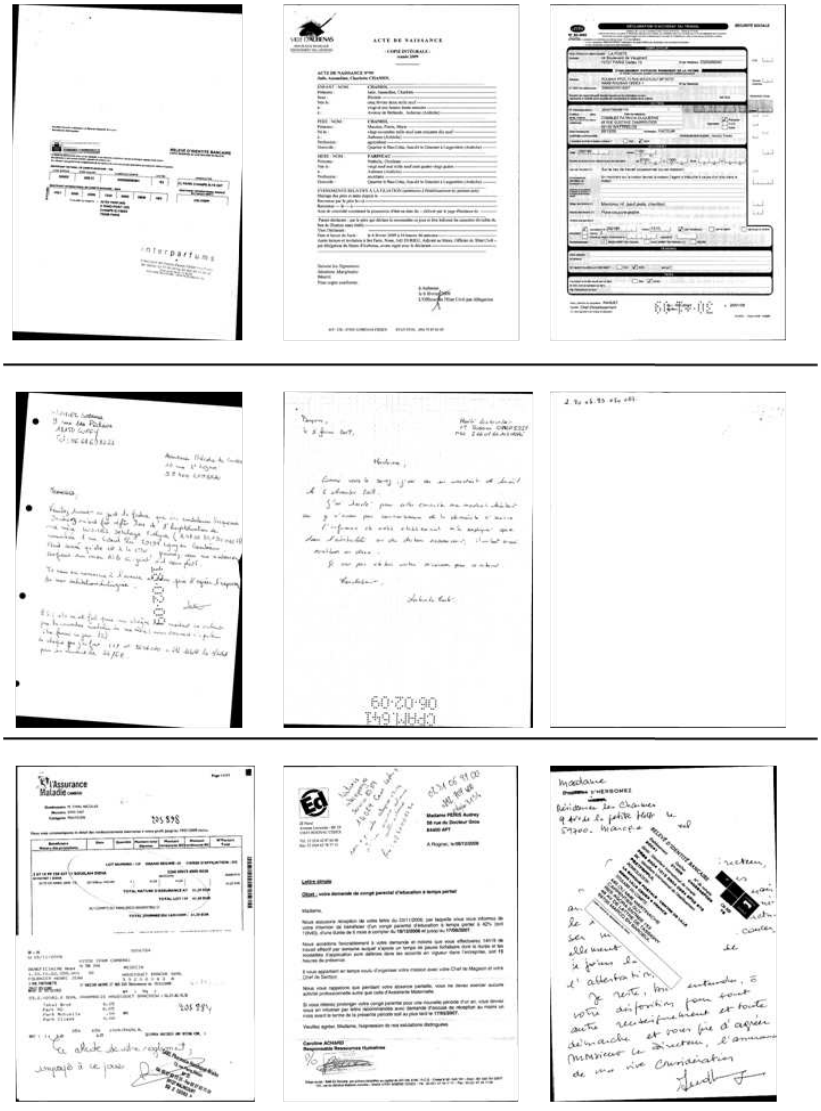

Figure 1. *Exemples de documents de travail*

meilleure manière possible à des outils de reconnaissance dédiés de documents isolés. Les documents sont considérés comme aléatoires, étant donné qu'il n'y a pas de contrainte forte sur leur contenu ou sur leur structure et nous devons aussi les traiter comme un flux aléatoire dans le sens où nous ne possédons strictement aucune information préalable sur le type de document à traiter.

Le travail que nous présentons ici est l'étape intermédiaire : nous nous contenterons de segmenter ou sur-segmenter ce type de document composite et donnerons une segmentation au niveau bloc/paragraphe pour les documents plus simples, unis et à structure physique de type Manhattan.

Il n'y a, à notre connaissance, aucune publication traitant de documents aussi complexes que ceux que nous devons manipuler. À la lecture de l'état de l'art et de nos propres tests, nous avons estimé que la méthode utilisant le recouvrement d'espaces blancs avait le plus de potentiel. En effet, c'est une des seules qui n'a pas besoin de modèle ni d'information sur la taille ou la fonte du texte et qui peut s'appliquer sur des documents non-textuels.

Dans (Shafait *et al.,* 2008) les auteurs proposent une évaluation de cette méthode et les autres citées précédemment sur la base publique UW-III. Ils présentent les points forts et faibles de chacune d'elles. Ils concluent en recommandant le XY-cut ou le smearing pour des documents simples, sans bruit, à structure Manhattan et pour les autres, les méthodes Docstrum ou Voronoï. Notons que toutes les optimisations, les paramètres à fixer et les conclusions concernent des structures Manhattan.

Ce travail ne met pas plus en valeur la méthode des espaces blancs qui est jugée comme moyenne. Il a toutefois était remis en question par (Nagy *et al.,* 2009), argumentant que sans pré-traitements, et plus particulièrement sans élimination des bords noirs, les méthodes par projection sont fortement handicapées. De manière générale, et comme le suggère la réponse (Shafait *et al.,* 2009), tout bruit marginal est pénalisant pour n'importe quelle méthode et qu'éliminer ce bruit est lui-même un problème très difficile. C'est pourquoi nous avons apporté un soin particulier au pré-traitement du document (Sec. 3) tout comme à l'application et à la modification de l'algorithme par recouvrement afin de rester stable sur des documents fortement bruités.

2. Recouvrement par espaces blancs

La segmentation par recouvrement d'espaces blancs (S-WSC) telle que présentée par Baird est constituée de deux grandes étapes. La première, dite algorithme d'énumération (WSC), consiste à recouvrir la page par des rectangles maximaux, des rectangles qui ne peuvent pas être plus étendus car rencontrant des obstacles ou les bords de la page. Dans le cadre de la segmentation géométrique de document, il considère dans ses travaux qu'il est plus facile de séparer le texte par l'espace blanc l'entourant que par le texte lui-même. Ainsi, le recouvrement se fait sur le fond du document (l'arrière plan), les zones de texte devenant elles les obstacles (le premier plan). Afin de ne faire aucune supposition forte sur le document, les obstacles considérés sont les composantes connexes. L'algorithme énumère tous les rectangles maximaux qui peuvent être crées autour des composantes connexes puis sont triés suivant un ordre partiel dépendant de l'aire ($area$) et du ratio ($aspect$) des rectangles obtenus. En considérant les N premiers rectangles de la liste, on crée un recouvrement partiel du fond (et par conséquent on entoure les zones de texte). Plus N est grand, plus la segmentation est

fine. Hormis une approximation grossière de la taille des caractères, la méthode n'a besoin d'aucun autre paramètre pour trouver un recouvrement. En jouant sur l'aire et le ratio des rectangles ($alar$), il est très aisé de fournir des séparations fines (par exemple au niveau de la ligne) ou plus grossières (au niveau des paragraphes).

Une fois un recouvrement obtenu (*cover set*), on considère que l'arrière plan du document est recouvert partiellement par l'union des rectangles du *cover set*. La segmentation est donc constituée des zones non contenues dans le *cover set*.

Nous avons retenu cette méthode car c'est la seule qui n'a besoin d'aucun paramètre pour fonctionner et fournir un découpage géométrique de la page. Elle a surtout le grand avantage de ne pas se baser sur des présupposés de l'écriture, de sa forme, de sa taille, elle peut même fonctionner dans le cas du manuscrit. Ceci est d'autant plus important que dans les documents composites, des informations tangibles sur le contenu de chaque document sont beaucoup plus difficiles à obtenir. Parfois il n'y a quasiment aucune information textuelle (montage de pièces d'identité par exemple). De plus, nous devons aussi traiter des documents "normaux" (pleine page) qui, s'ils ont une structure Manhattan, doivent être subdivisés au niveau paragraphe ou légèrement plus gros. Ne sachant à l'avance le type de document, il nous a semblé que seule une méthode similaire au WSC pouvait répondre à notre objectif.

3. Prétraitements communs

L'étape de prétraitement est cruciale pour la segmentation imprimé/manuscrit et de manière générale conditionne toutes les étapes futures de rétro-conversion. Cet article considère cette étape comme déjà effectué. Nous décrivons brièvement ce que nous avons appliqué pour se prémunir des défauts les plus récurrents :

– l'inclinaison du document numérisé par rapport au document d'origine ;

– des bordures noires autour du document ;

– du bruit type poivre et sel créé par la numérisation de fond grisé ou à motif sur le document d'origine ;

– des zones ou traits noircis à cause d'impuretés présentes sur la vitre du scanner ;

– du bruit dû à l'apposition de plusieurs documents sur la même page ;

– la présence d'informations non textuelles considérées comme du bruit (logos, cadres, bordures de tableaux, etc.).

Le filtrage est constitué des étapes suivantes :

– suppression des bordures par un système de règles sur la forme et la position des composantes connexes ;

– premier filtrage du bruit par un kfill modifié (Chinnasarn *et al.,* 1998) ;

– détection de l'inclinaison par la méthode RAST (van Beusekom *et al.,* 2010) ;

– second kfill modifié (Chinnasarn *et al.,* 1998) sur le document redressé.

Après nettoyage de l'image 2, une estimation de la répartition des composantes connexes est faite. Trois grandes classes sont construites : celle de la ponctuation, celle des caractères, celle des composantes de très grande taille. Pour des documents

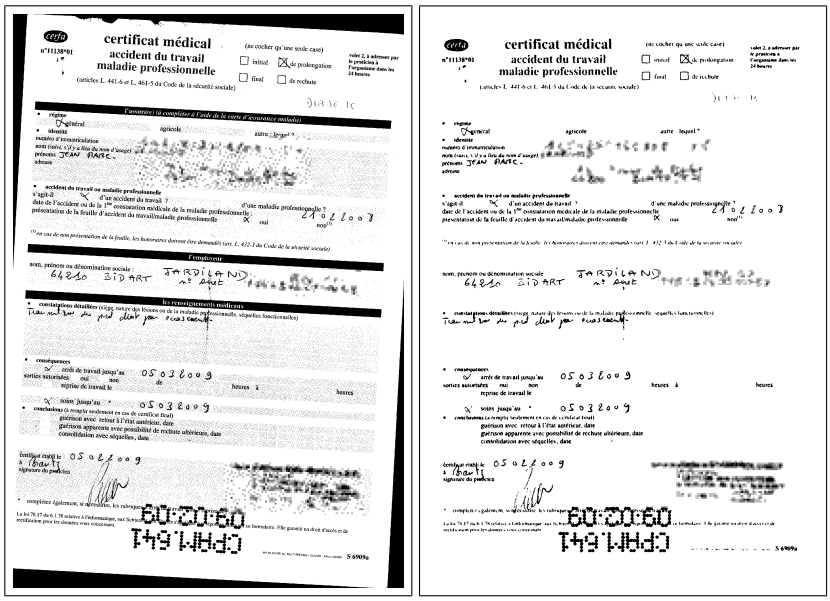

Figure 2. *Exemple d'application de la procédure de nettoyage*

propres, à forte prédominance textuelle, nous connaissons les valeurs attendues. Au moindre problème lors de l'extraction des statistiques, nous utilisons une valeur par défaut fixée à 7 pixels pour la hauteur d'un caractère (*xheight*). Nous retenons uniquement les composantes connexes d'aire supérieure à $2 \times xheight$ et inférieure à la moitié de la surface de la page comme obstacles.

4. Couverture

Contrairement au travail de Baird, nous ne pouvons pas faire autant de "bonnes" suppositions quant à la forme des rectangles à trouver. S'il est clair qu'un rectangle extrêmement fin, proche d'un segment de droite, ne sera probablement pas un bon candidat, il est moins évident de juger d'autres rectangles suivant leur ratio (qu'il soit à 1 ou extrêmement petit ou grand). La clé de tri proposée par Baird mélangeant aire et ratio ($key(cover) = area(cover) \times (1 + alar(cover))$) est beaucoup trop dirigée structure Manhattan. Elle l'est d'autant plus si on utilise sa fonction de pondération W sur la fonction $alar$. En particulier, W pénalise les rectangles proches d'une forme carrée et aurait tendance à n'accorder que très peu de poids aux longs rectangles allongés ce qui, dans les deux cas, nous a posé problème sur des documents Manhattan et non-Manhattan. Comme il est difficile d'établir une fonction de pondération comme l'a fait Baird sur des documents d'une même classe, nous préférons ne pas tenir compte aussi tôt dans le processus de paramètres déjà dépendants de la classe de document. Nous optons donc pour une clé ne faisant intervenir uniquement l'aire des rectangles.

Pour ne pas privilégier des formes dégénérées, nous préférons les écarter directement du *cover set* et de ne plus jamais les considérer. Nous avons donc établi une fonction déterminant ce qu'est un "bon" rectangle en suivant les idées de Baird et al. ; en étant plus strict sur ceux considérés comme très mauvais et en laissant un maximum de rectangles même légèrement dégénérés avoir une clé de grande valeur.Très précisément, on fixe un intervalle $[min\_aspect, max\_aspect]$ qui borne le ratio largeur sur hauteur, des bornes minimales min_width et min_height pour la largeur et hauteur et un $alar$ minimum ($alar(w,h) = abs((log(w/h)/log(2))))$ borné par $logmin\_aspect$.

Durant les expérimentations, nous avons utilisé $min\_width = min\_height = 30$, $logmin\_aspect = 2$ et pas de contraintes pour min_aspect et max_aspect. Plutôt que d'utiliser la méthode de balayage de Baird, nous avons ajouté une fonction de recherche de "pivot" comme proposé par (Breuel, 2002) qui détermine quel est le meilleur obstacle à l'intérieur d'une zone à découper. La présence d'un nouvel obstacle donne toujours lieu à la création de quatre sous-rectangles autour de l'obstacle. Nous prenons celui de centricité maximale ce qui revient à choisir de préférence un obstacle petit et placé au centre de la zone.

Le critère d'arrêt chez nous est soit une aire minimale atteinte par un *cover*, soit par un nombre suffisant de covers atteint. Ceci est beaucoup plus simple à mettre en œuvre que la méthode par apprentissage proposée par Baird. Il nous faudrait sinon créer des règles pour chaque classe de document et surtout savoir déterminer la classe d'un document lors de la segmentation. D'ailleurs, dans la première version, (Baird *et al.,* 1990) proposait de stopper quand les rectangles candidats atteignaient un ratio de plus de 16, laissant sous-entendre que la segmentation en était au niveau de la ligne. Nous préférons dans notre cas aussi une règle géométrique simple, facile à contrôler, d'autant plus que notre segmentation n'a pas besoin d'être très fine.

5. Pré-traitements / post correction pour l'algorithme de recouvrement

5.1. *Pour les documents fortement textuels*

Pour les documents dont il est possible de supposer la présence majoritaire de contenu textuel avec uniquement des statistiques sur les composantes connexes, nous effectuons une correction de la couverture. On partira du postulat qu'il est très pro-

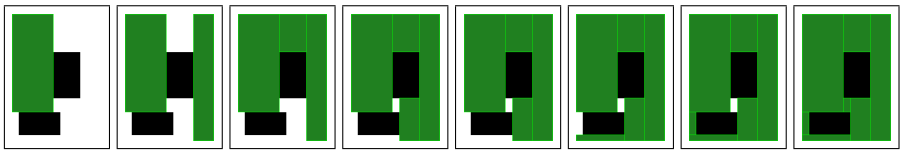

Figure 3. *Cas simple de l'application d'un recouvrement par espaces blancs sur deux obstacles : toutes les étapes permettant le recouvrement*

bable d'avoir un seul et même document avec une structure Manhattan. On va alors privilégier dans la couverture trouvée par l'algorithme de recouvrement toutes les grandes bandes verticales qui permettront de repérer les gouttières. Ce sera principalement les marges verticales, ainsi que les gouttières de séparation de colonnes de texte. On choisit aussi de préférence les rectangles reposant sur un maximum d'obstacles (c'est à dire des rectangles qui toucheront a priori un maximum de caractères et ce sur plusieurs lignes). Ensuite, on effectue un traitement similaire pour les gouttières horizontales, en privilégiant les plus longues et celles traversant si possible de part en part deux gouttières verticales. Si le document a pu être redressé lors de la phase de prétraitement, il est possible d'obtenir un recouvrement parfait jusqu'au niveau de la ligne, mais nous nous arrêtons de toute façon avant (Fig. 4).

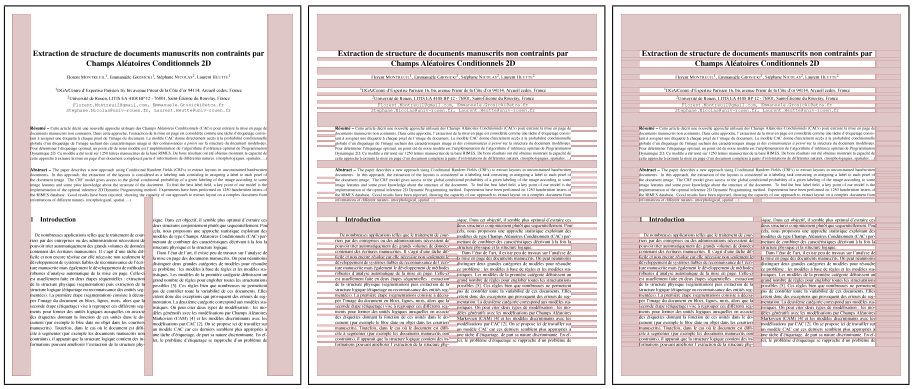

Figure 4. *Repérage des gouttières verticales sur l'image de gauche, repérage des gouttières horizontales sur l'image du milieu, combinaison des deux sous-recouvrements sur l'image de droite*

5.2. *Pour les documents compliqués*

Pour tous les autres documents, on peut faire deux hypothèses principales. Soit la page était un seul et même document mais rendu totalement "illisible" à cause d'un bruit trop important, soit la page contenait plusieurs documents ou beaucoup trop de non-texte. Dans les deux cas, il convient de revoir le prétraitement. Nous nettoyons plus brutalement le document, en lui faisant subir une sorte d'étape de quantisation. C'est le but de l'algorithme 1. On dilate une première fois l'image pour récupérer d'éventuels caractères cassés et qui auraient pu jouer un rôle dans un nombre anormal de composantes connexes. L'image est réduite par un facteur de 8 (ce qui représente approximativement une image de 1M pixels si l'image d'origine était numérisée en 300 dpi). L'image est lissée par un filtre gaussien avec une variance relativement forte, ce qui permet de renforcer les zones de données et de supprimer les petites composantes connexes isolées de l'image d'origine. Intervient alors une binarisation par seuil de l'image pour séparer le bruit de l'information. Afin d'éviter au maximum

de perdre de l'information si la binarisation a été trop forte, nous proposons de dilater l'image avec un élément structurant rectangulaire allongé horizontalement (on fait l'hypothèse que la plupart des documents ont du texte parallèle à l'axe des abscisses). Reste à agrandir cette image avec le même facteur que celui de la réduction et on construit ainsi un masque qui sert à convoluer l'image d'origine. On créé une nouvelle image dont le bruit disparait, la disparition laissant place à des zones rectangulaires blanches (idéales pour le WSC), et renforçant certains bords ou zones d'information qui auraient pu être découpés inutilement par le WSC. La méthode de recouvrement est appliquée sur cette nouvelle image (Fig. 5).

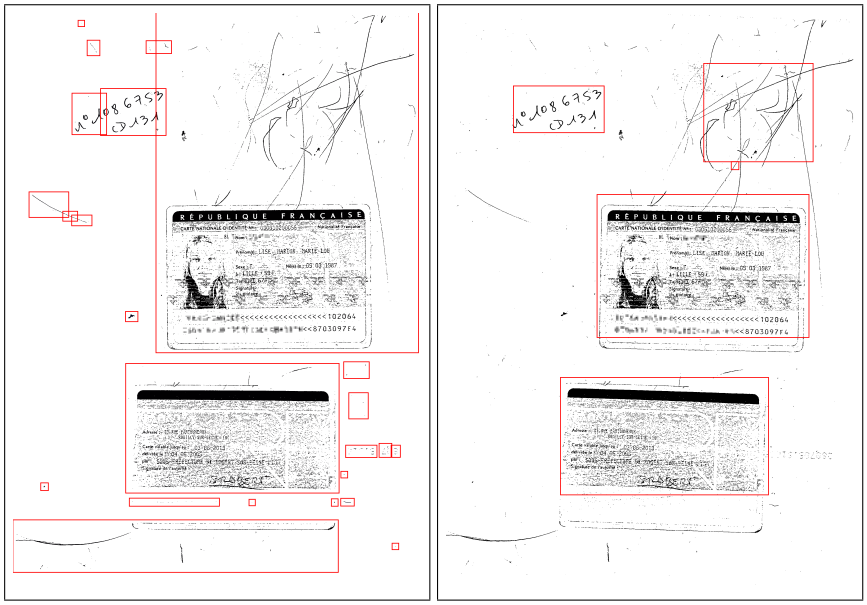

Figure 5. *À gauche, résultat de la segmentation sans le pré-lissage. À droite, résultat de la segmentation avec le pré-lissage*

Algorithm 1 Lissage d'images bruitées

function LISSAGE(image)
  $masque \leftarrow dilatation(image)$
  $masque \leftarrow zoom(masque, zf)$
  $masque \leftarrow filtregaussien(masque, sigma)$
  $masque \leftarrow binarise(masque, bi)$
  $masque \leftarrow dilatation(masque, (d1, d2))$
  $masque \leftarrow zoom(masque, 1/zf)$
  return $image * masque$
end function
$(zf, sigma, bi, dil) \leftarrow (0.125, 5, 24, (9, 3))$

6. Segmentation par recouvrement

L'algorithme de recouvrement donne pour ainsi dire le négatif de la segmentation que l'on voudrait obtenir. Selon Baird, c'est le cas et ceci est vrai pour des documents très simples (idéalement des PDF convertis informatiquement en image). Dans la pratique, les documents à manipuler sont bien plus compliqués et bruités que ceux manipulés dans (Baird *et al.,* 1990) et (Baird, 1994) (et d'ailleurs dans la majeure partie des publications, même récentes (Shigarov *et al.,* 2011)). Nous ajoutons une dernière étape pour construire la segmentation finale. Nous effectuons à nouveau l'algorithme de recouvrement, mais cette fois-ci en utilisant le *cover set* comme obstacle. On peut ainsi intervenir facilement sur la forme géométrique présupposée des zones de texte ou des sous-documents à segmenter. On peut raffiner la segmentation à l'intérieur des blocs si l'une des étapes présentées en 4.1 s'est mal déroulée, et enfin, on considère les composantes connexes de l'image d'origine, avant le nettoyage, afin d'étendre si besoin est, certaines zones. Il va de soi que si les documents de départ sont idéaux, nous n'allons pas plus loin que ce qui est proposé par Baird et al., par contre, pour des documents complexes, nous pouvons rattraper un certain nombre d'erreurs.

7. Expérimentation

Les tests s'effectuent sur une base de plus de 500 pages fournies par une société partenaire. Dans cette base, il y a 32 "classes" de document, qui vont de photocopies de cartes grises, à des formulaires de la sécurité sociale, ou encore des ordonnances de médecins. La variabilité à l'intérieur de chacune de ses classes est très forte. Pour étalonner les différentes méthodes et fixer les seuils que nous utilisons, nous avons extrait 32 documents paraissant couvrir cette variabilité. Les résultats sont jugés quantitativement, il n'y a pas de document de vérité. Les résultats fournis par un OCR commercial[1] nous servent d'étalon pour juger de la qualité des segmentations. Des exemples sont donnés respectivement en Fig. 6. Nous avons paramétré notre méthode afin qu'elle ait tendance à fournir une sur-segmentation. L'objectif à long terme étant d'utiliser ses résultats suivis d'une méthode de regroupement qui servira à extraire tous les sous-documents d'une image composite.

8. Conclusion

Nous avons présenté dans cet article une méthode pour la segmentation de documents composites difficiles. Elle s'appuie sur la technique de recouvrement d'espaces blancs proposée par Baird. Nous avons simplifié et adapté la méthode pour que les recouvrements s'effectuent de manière satisfaisante sur des documents n'ayant pas une structure Manhattan. Nous avons aussi apporté une méthodologie pour nettoyer les documents afin que la méthode de recouvrement puisse être appliquée de manière

1. http ://www.nuance.com/for-individuals/by-product/omnipage/index.htm

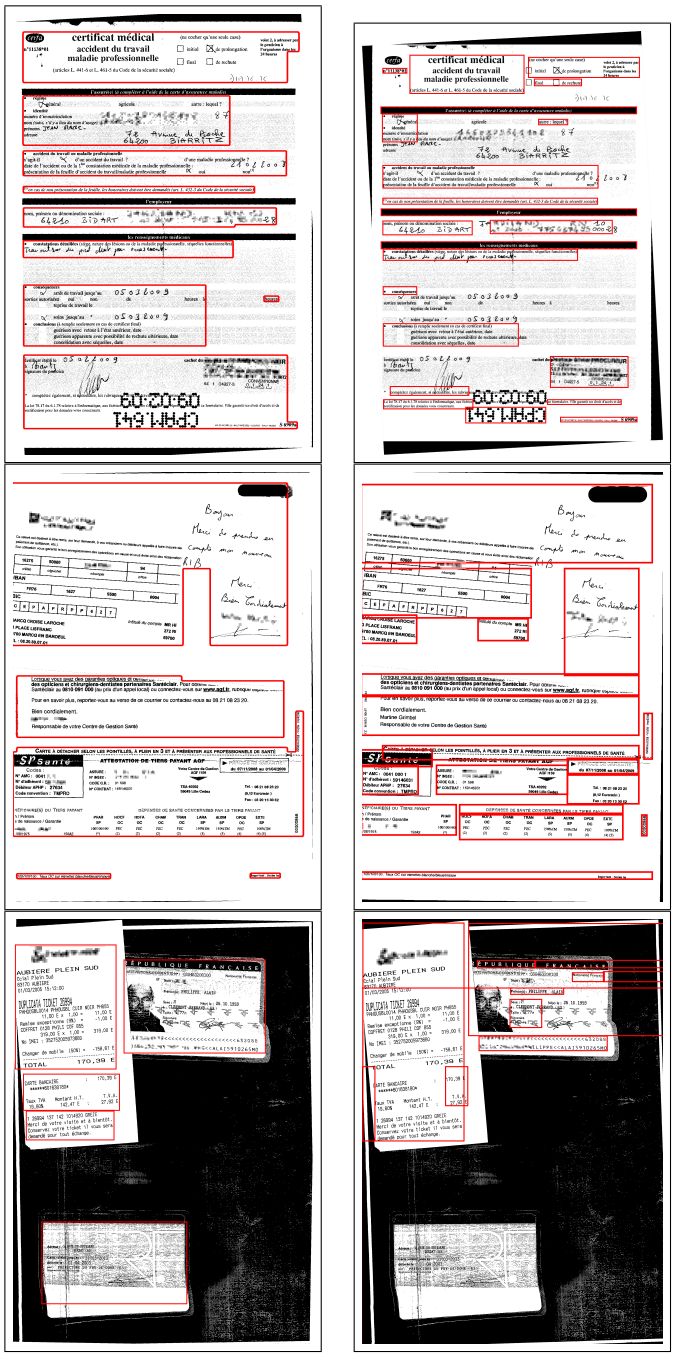

Figure 6. *Résultats obtenus avec la méthode proposée dans cet article dans la colonne de gauche. Résultats obtenus avec l'OCR OmniPage 16 dans la colonne de droite*

optimale. Retrouver alors une segmentation de la page est alors une étape beaucoup plus aisée. En faisant la distinction entre documents à structure Manhattan et non-Manhattan, la méthode arrive à surpasser celles que les OCR commerciaux peuvent obtenir et ce quelque soit la classe de document à traiter. Ce travail est une étape intermédiaire dans un processus plus complet de segmentation, il nous reste désormais à concevoir l'étape de regroupement qui permettra d'extraire tous les documents composites en s'aidant de la sur-segmentation que nous avons proposée.

9. Bibliographie

Baird H., « Background structure in document images », *Document Image Analysis, Word Scientific*, p. 17-34, 1994.

Baird H., Jones S., Fortune S., « Image segmentation by shape-directed covers », *International Conference on Pattern Recognition*, vol. 1, p. 820-825, 1990.

Breuel T. M., « Two Geometric Algorithms for Layout Analysis », *In Workshop on Document Analysis Systems*, p. 188-199, 2002.

Chinnasarn K., Rangsanseri Y., Thitimajshima P., « Removing salt-and-pepper noise in text/graphics images », *The Asia-Pacific Conference on Circuits and Systems*, p. 459-462, 1998.

Kise K., Sato A., Iwata M., « Segmentation of page images using the area Voronoi diagram », *Computer Vision Image Understanding*, vol. 70, p. 370-382, 1998.

Nagy G., Seth S., Viswanathan M., « A prototype document image analysis system for technical journals », *Computer*, vol. 25, p. 10-22, 1992.

Nagy G., Seth S., Viswanathan M., « Comment : Projection methods require black border removal », *Pattern Analysis and Machine Intelligence*, vol. 31, p. 762, 2009.

O'Gorman L., « The document spectrum for page layout analysis », *Pattern Analysis and Machine Intelligence*, vol. 15, p. 1162-1173, 1993.

Pavlidis T., Zhou J., « Page segmentation by white streams », *International Conference on Document Analysis and Recognition*, p. 945-953, 1991.

Shafait F., Keysers D., Breuel T., « Performance evaluation and benchmarking of six-page segmentation algorithms », *Pattern Analysis and Machine Intelligence*, vol. 30, p. 941-954, 2008.

Shafait F., Keysers D., Breuel T., « Response to "Projection methods require black border removal" », *Pattern Analysis and Machine Intelligence*, vol. 31, p. 763-764, 2009.

Shigarov A., Fedorov R., « Simple algorithm page layout analysis », *Pattern Recognition and Image Analysis*, vol. 21, p. 324-327, 2011.

van Beusekom J., Shafait F., Breuel T., « Combined orientation and skew detection using geometric text-line modeling », *International Journal on Document Analysis and Recognition*, vol. 13, p. 79-92, 2010.

Wahl F. M., Wong K. Y., Casey R. G., « Block segmentation and text extraction in mixed text/image documents », *Computer Graphics and Image Processing*, vol. 20, p. 375-390, 1982.