Segmentation et classification des zones d'une page de document

Jean-Marc Vauthier, Abdel Belaïd

▶ **To cite this version:**

Jean-Marc Vauthier, Abdel Belaïd. Segmentation et classification des zones d'une page de document. CIFED-CORIA, Mar 2012, Bordeaux, France. hal-00779232

HAL Id: hal-00779232
https://inria.hal.science/hal-00779232

Submitted on 23 Jan 2013

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Segmentation et classification des zones d'une page de document

Jean-Marc Vauthier, Abdel Belaïd

LORIA
Campus scientifique
BP 259
F-54506 Vandoeuvre-lès-Nancy Cedex
{jean-marc.vauthier, abelaid}@loria.fr

RÉSUMÉ. *Cet article propose une méthode de segmentation de documents complexes en zones d'intérêt en s'appuyant à la fois sur le contenu textuel et la forme. Le contenu textuel correspond aux sorties lisibles validées par un dictionnaire et des expressions régulières adaptées aux données bruitées. Ceci permet en parallèle de localiser des textes d'intérêt (adresses, numéros de téléphone, formules de politesse, etc.). Le contenu non lisible est regroupé en régions physiques en prenant en compte la taille et l'éloignement des composantes connexes en vue de l'identification de zones spécifiques, comme des logos, des signatures et des tampons. Pour cela, des descripteurs morphologiques sont appliqués. Cette classification s'appuie sur une méthode de boosting modifiée associée à des arbres de décision. La modification a porté sur le calcul de la probabilité d'appartenance d'un individu à une classe. Par rapport à l'action actuelle des OCRs qui classent le texte, les tableaux et les images, les résultats de notre méthode accroissent non seulement ces performances mais elle permet aussi à des zones à faible consensus comme, les annotations manuscrites, les logos, les tampons et surtout les signatures d'être reconnues.*

ABSTRACT. *This paper proposes a methodology for complex document segmentation based on textual content and shape. The textual content corresponds with printed text and it is verified by text-word analysis using dictionary and regular expressions variable that are adapted to noise. This allows knowing where the interested expressions are placed (address, phone number etc.) The non-textual content is segmented in zone considering size and distance between connected components in order to classify zones like logo, signature, and table. To make that, features are extracted like run length, Bi level Co-occurrence… This classification is based on a modified boosting method and decision trees. The modification is about the calculation of the probability to draw training data. Compare to OCRs that are able to classify text, tables and pictures, our methodology increases the performance and allows the detection of other zones like handwritten text, logo, signature, table and tampon.*

MOTS-CLÉS : *Segmentation de documents, OCR, boosting, classification.*

KEYWORDS: *document segmentation, OCR, boosting, classification.*

1. Introduction

La classification automatique de documents devient une étape importante de la gestion des courriers entrant dans les entreprises. Leur nombre est de plus en plus important et la nécessité d'obtenir rapidement le document recherché est primordiale. Pour cela, il faut constituer des dossiers à partir des documents entrants. La solution que nous avons retenue pour répondre à ce problème repose sur la modélisation des documents afin d'effectuer un raisonnement à partir de cas. Pour modéliser ces documents, il faut au préalable effectuer leur segmentation, puis classer les zones définies. C'est cette première partie qui sera exposée dans cet article.

La difficulté majeure de la segmentation et de la classification provient de la grande diversité des types de documents à traiter et de leur qualité très variable. Il s'agit de procès-verbaux, factures, photographies, feuilles d'assurance maladie ou encore de constats amiables. Ils contiennent pour la plupart principalement du texte dactylographié mais aussi des signatures, des annotations manuscrites, des logos, des tampons, etc. Certains ne contiennent aucun texte dactylographié, sont très bruités à cause de leur numérisation et leur orientation n'est pas toujours bonne. De plus, leurs dimensions varient du fait qu'ils ne proviennent pas tous du même client. Une autre variable est à prendre en compte, il s'agit du temps. La méthode utilisée ne doit pas nécessiter plus de quelques secondes.

Aucun pré traitement n'est réalisé car la qualité des documents étant très variable, il est difficile de l'adapter en fonction du type de bruit présent sans perdre d'information qui peut parfois être assimilé à du bruit. De plus, le flux de document est aléatoire et ne contient aucune information sur les types de documents présents. La seule constante provient du format TIFF des documents et de leur encodage binaire perdant ainsi les informations liées à la couleur et au niveau de gris.

La segmentation n'est pas un thème nouveau (G. Nagy and S. Seth, 1984), de nombreuses méthodes ont déjà été proposées (O'Gorman, 1993). Des méthodes plus spécifiques (A. Antonacopoulos, 2009) permettent la segmentation de pages de magasine ou de journaux. Mais, à notre connaissance, aucune n'est adaptée à notre problématique où se côtoient de nombreux types de documents de qualité très variable. Pour résoudre ce problème, nous avons développé une nouvelle solution utilisant à la fois le contenu textuel et la forme des documents. Ce contenu textuel est validé par un dictionnaire et des expressions régulières adaptées aux bruits. En parallèle, le contenu non textuel est regroupé en régions physiques suivant la taille et la dispersion des composantes connexes dans le but de classifier ces zones. Des descripteurs morphologiques (run length, composantes connexes, Bi-level Co-occurrence, etc.) sont appliqués sur les zones définies puis un classifieur modifié utilisant le boosting et les arbres de décision est utilisé. La modification par rapport aux méthodes de boosting connues porte sur l'évolution des probabilités d'appartenance d'un individu à une classe. Les résultats obtenus par rapport aux OCRs classant le texte, les tableaux et les images accroissent ces performances mais

permettent aussi de classer des zones peu représentées comme les tampons ou les signatures.

2. La méthode proposée

Plusieurs étapes sont nécessaires à la segmentation des documents et la classification des zones. La première étape consiste à analyser chaque document par un OCR. Les mots extraits sont regroupés en ligne puis en paragraphe. La seconde étape a pour but de valider les résultats de l'OCR par une analyse orthographique et une étude des pixels du texte extrait.

Toutes les zones dites « texte dactylographié » sont ensuite supprimées de l'image d'origine afin de ne conserver que les autres types d'information. Un regroupement par composantes connexes suivant leur taille et leur distance permet de segmenter les zones restantes. Elles sont ensuite analysées et classées par un algorithme de boosting modifié pour définir ce qu'elles représentent. Le schéma de la figure I récapitule les différentes étapes.

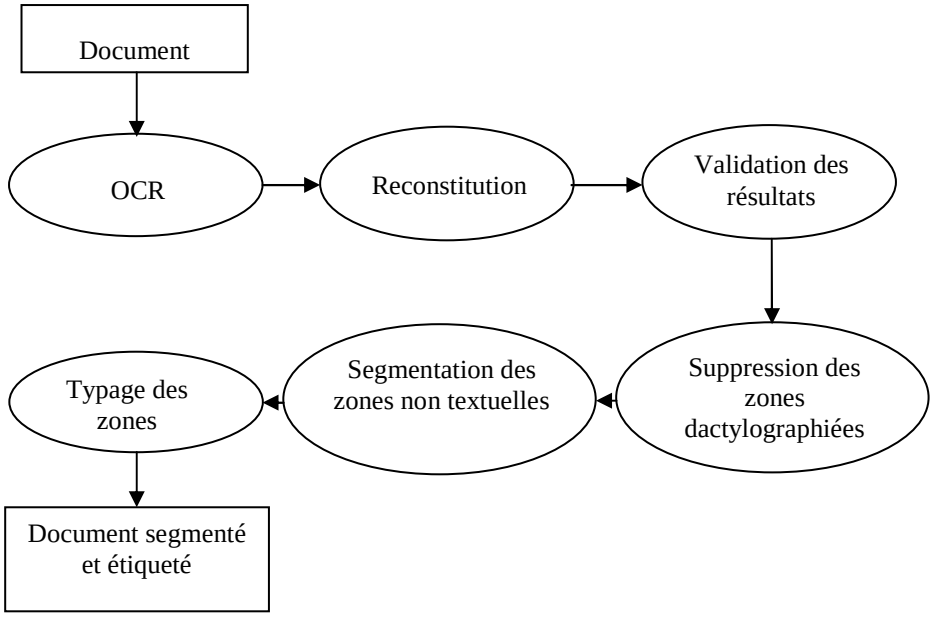

Figure I. *Vue d'ensemble du système de segmentation.*

3. La reconnaissance textuelle

3.1 *Reconstitution des paragraphes*

L'OCR est capable de reconnaitre les caractères d'un document, les mots les lignes et de proposer une segmentation. Cependant, des erreurs peuvent survenir dans ces différentes étapes, surtout si le document ne comporte pas uniquement du texte dactylographié ou si l'organisation du document est complexe. C'est pour cette raison que nous n'utilisons pas l'ensemble des résultats de l'OCR. En effet, seuls les mots et leur position sont conservés. La méthode utilisée se veut simple et rapide.

La reconstruction des lignes s'effectue de la manière suivante : la hauteur moyenne de chaque mot est calculée, puis tous les mots qui sont sur la même ligne horizontale plus ou moins 15% de la hauteur moyenne et qui ont un espacement inférieur à deux fois la hauteur des caractères sont regroupés.

La reconstruction des paragraphes se fait en rassemblant les lignes. Pour cela, leur espacement vertical ne doit pas dépasser deux fois la hauteur moyenne de leurs caractères et leur différence d'alignement gauche ne doit pas être supérieure à 8 fois la hauteur de leurs caractères.

La figure II montre le résultat sur plusieurs exemples de cette segmentation à partir des résultats de l'OCR. Par un souci de simplification de représentation, les zones sont toutes rectangulaires.

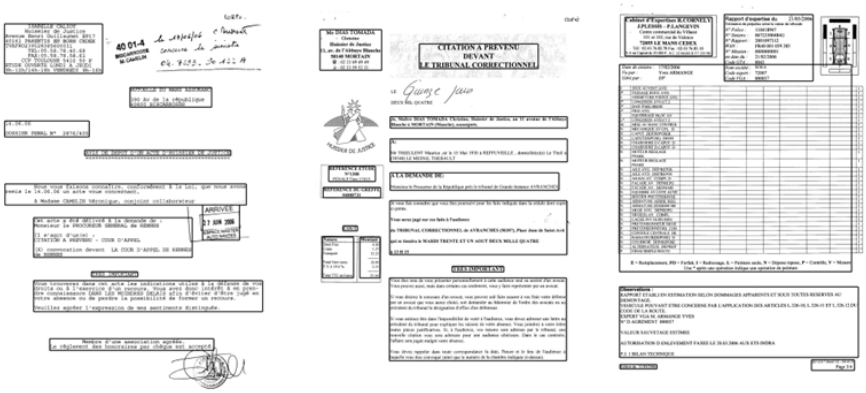

Figure II. *Résultat de la segmentation du texte dactylographié. Chaque zone est délimitée par un rectangle noir.*

3.2 *Validation des résultats*

Les paragraphes obtenus sont analysés afin de valider ou non les choix de l'OCR. En effet, il a tendance à reconnaitre du bruit ou des caractères manuscrits comme du texte dactylographié. Dans ce cas, les caractères retournés ne forment pas de texte ayant du sens. Chaque zone ayant pour objectif d'être étiquetée suivant ce

qu'elle contient, il convient de différencier les zones contenant du texte dactylographié des zones n'en contenant pas.

Deux méthodes sont utilisées. La première consiste à analyser les mots et phrases reconnues par l'OCR. Pour cela, un dictionnaire simplifié a été créé à partir des documents de la base de test. Il comporte les noms communs les plus présents. D'autre part, des expressions régulières permettant de reconnaitre les numéros de téléphone, les adresses, les formules de politesses, les champs « objet » et les dates indiquent la présence de texte dactylographié correctement reconnu. Ces expressions régulières ne sont pas trop strictes sur la détection des différentes informations. En effet, lorsqu'il y a un faible pourcentage d'erreur de reconnaissance de caractères par l'OCR, l'expression régulière doit être en mesure de reconnaitre l'information sans pour autant faire de fausses détections. Ces détections permettent de valider le texte mais ont pour premier objectif de mieux déterminer le contenu d'une zone textuelle. Par exemple, pour les adresses, plusieurs informations permettent de la caractériser : le numéro de rue, le type de rue, le code postal, la ville, le mot « cedex ». Il n'est pas nécessaire de trouver toutes ces informations pour définir que c'est une adresse. Suivant celles qui sont trouvées, on définit ou non si la zone contient une adresse.

La seconde méthode utilisée (Grzejszczak, 2012) a pour but de séparer le texte manuscrit du texte imprimé. Les différentes étapes sont les suivantes :

– des prétraitements :

- kFill pour supprimer le bruit poivre et sel (Chinnasarnet al., 1998)
- suppression des bordures noires
- redressement du document
- second filtrage par kFill (Chinnasarnet al., 1998)

– une segmentation du contenu des zones en « pseudo-mots » par smearing double permettant ainsi une segmentation plus fine et mieux adaptée à chaque ligne du document

– une caractérisation de chaque pseudo-mot par extraction de caractéristiques :

- densité de pixels noirs d'un pseudo-mot
- étude des composantes connexes
- distribution verticale des pixels
- profil supérieur-inférieur
- dérivée maximale du profil horizontal
- longueur et nombre de segments horizontaux
- bilevel co-occurrence
- N x M grams avec N = M = 2, 4 distances et 15 motifs.

– une classification par SVM de la classe d'appartenance de chaque « pseudo-mot » : manuscrit, imprimé, autre

– une phase de post correction utilisant le contexte pour affiner la segmentation. Il s'agit de comparer le type attribué au pseudo-mot considéré avec le type de ses voisins.

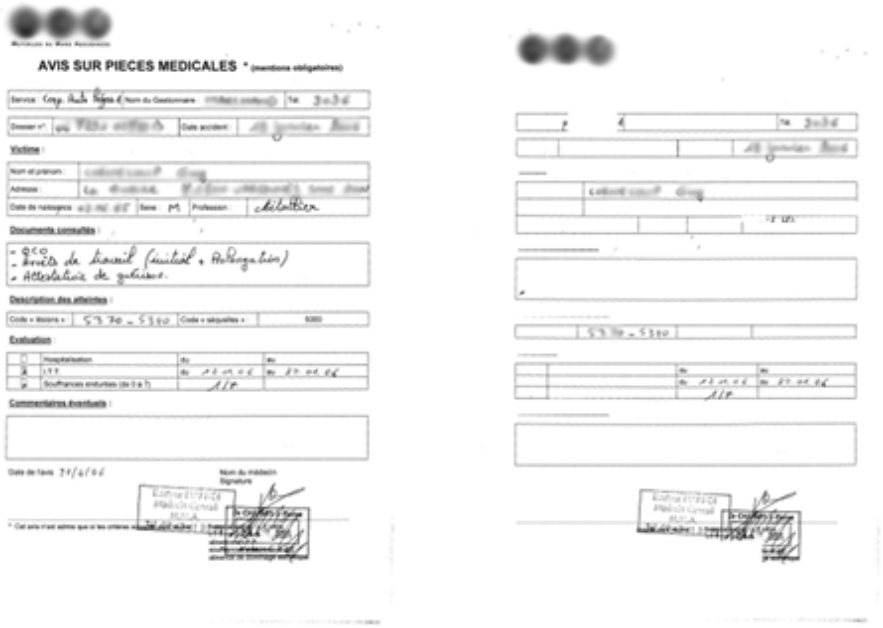

Figure III. *A gauche la page d'origine, à droite, la page après suppression du texte dactylographié. Le texte a été rendu flou volontairement.*

Chaque pixel est étiqueté suivant sa classe présumée d'appartenance. Un vote majoritaire permet de définir si la zone contient du texte manuscrit ou imprimé.

Les deux méthodes précédentes indiquent s'il s'agit ou non d'une zone contenant du texte dactylographié. Si leur avis divergent, c'est la première méthode qui prime si elle détecte des mots ou des expressions, sinon, c'est la seconde. Pour terminer cette étape, les mots contenus dans les zones validées comme zones dactylographiées sont supprimés de l'image d'origine comme le montre la figure III.

4. Reconnaissance des zones non textuelles

4.1 *Regroupement du rejet de l'OCR*

La segmentation des zones non dactylographiées est réalisée en regroupant les composantes connexes restantes. Afin de ne pas prendre en compte une partie du bruit, les composantes connexes ayant une taille inférieure à 4 * 4 pixels sont

ignorées. Cela permet aussi d'augmenter la vitesse de traitement. La formule suivante permet de calculer la distance corrigée entre deux composantes connexes :

$$e_H = \left|c_{H,1} - c_{H,2}\right| - \frac{tailleH_1}{3} - \frac{tailleH_2}{3}$$

$$e_V = \left|c_{V,1} - c_{V,2}\right| - \frac{tailleV_1}{3} - \frac{tailleV_2}{3}$$

$$d = \sqrt{{e_H}^2 + {e_V}^2}$$

avec $c_{H,1}$ (resp. $c_{H,2}$) l'abscisse du centre de la zone 1 (resp. 2) et $tailleH_1$ (resp. $tailleH_2$) la largeur de la zone 1 (resp. 2), $c_{V,1}$ (resp. $c_{V,2}$) l'ordonnée du centre de la zone 1 (resp. 2) et $tailleV_1$ (resp. $tailleV_2$) la hauteur de la zone 1 (resp. 2).

Si la valeur *d* est inférieure à 4% de la hauteur de l'image en pixel, alors les deux composantes sont regroupées. Ce chiffre de 4% provient de l'étude de la dispersion des pixels dans les pages à traiter afin de rassembler au mieux les composantes connexes tout en limitant le bruit. La figure IV montre le résultat de la segmentation sur 3 documents.

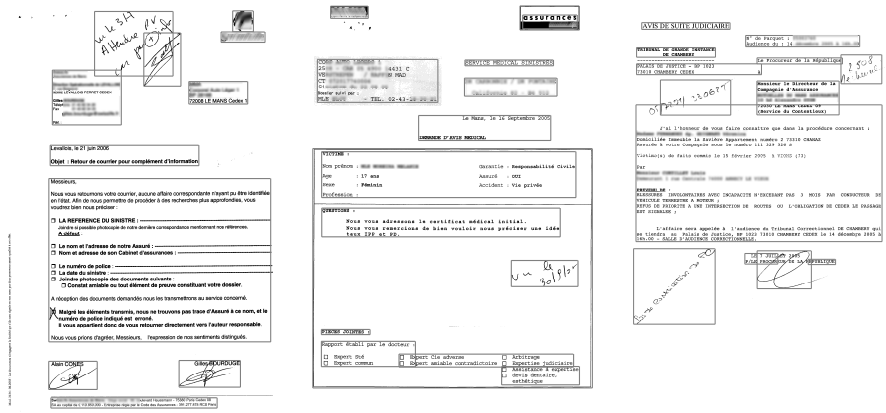

Figure IV. *Trois exemples de segmentation obtenus avec les documents de la base de données.*

4.2 *Classification des zones*

4.2.1 *Calcul de descripteurs*

Les zones conservées sont analysées par extraction de descripteurs. Les petites composantes connexes qui avaient alors été ignorées lors de la segmentation sont rajoutées afin de conserver les détails pour les objets, notamment les tampons, qui sont faiblement imprimés. L'objectif de l'extraction de descripteurs est de définir quel type d'information est contenu dans chaque zone (texte manuscrit, logo, tampon, signature, tableau, bruit). Ces descripteurs proviennent de plusieurs catégories qui sont les suivantes :

– les run lengths en adaptant les résultats d'Y.Wang (Y. Wang, 2006)

– les couples et les triplets de run lengths adaptés à partir de la méthode de Wang (D. Wang, 1989)

– les composantes connexes (taille, inertie, densité …)

– le bilevel Co-occurrence développé par Yefeng Zheng (Y. Zheng, 2004)

Ces différentes catégories fournissent 102 descripteurs. Une réduction du nombre est effectuée pour ne conserver que les plus efficaces. La méthode utilisée reprend l'heuristique de Mark A. Hall (M. A. Hall, 1999) :

$$M_S = \frac{k\overline{r_{cf}}}{\sqrt{k + k(k+1)\overline{r_{ff}}}}$$

Elle produit 43 descripteurs provenant des différentes catégories précédemment citées. A partir de la base d'apprentissage qui contient 3048 éléments, un algorithme de classification est créé. Il s'agit d'un algorithme de boosting adapté à partir de l'algorithme logitboost (J. Friedman, 2000).

4.2.2 *Boosting*

Les algorithmes de type boosting sont apparus dans les années 80. Ils regroupent des méthodes générales capables de produire des décisions très précises à partir de règles de décision dites « faibles ». Le boosting a pour origine le modèle PAC (Probably Approximately Correct) (L.G. Valiant, 1984). Il a montré que tout algorithme d'apprentissage produisant des résultats meilleurs qu'un choix aléatoire peut être boosté afin d'atteindre une erreur aussi faible que voulue, avec une probabilité aussi grande que désirée en un temps polynomial. Ceci a été démontré à deux reprises (R. E. Schapire, 1990), (Y. Freund, 1995). Le principe est d'entrainer plusieurs règles de décision dites « faibles » à partir des exemples de la base d'apprentissage. Il s'agit généralement d'arbres de décision. Le choix de l'exemple se fait de manière pseudo aléatoire car chaque exemple est pondéré par un coefficient. Il s'agit d'un des facteurs d'influence des résultats du classifieur. Son influence est proportionnel à la qualité des données, si elles sont bruitées et hétérogènes dans chaque classe ou non. Enfin, lorsque toutes les règles de décision

ont été créées, la classification d'un nouvel élément se fait par vote pouvant être pondéré de chaque règle.

Pseudo algorithm 1 LogitBoost

$w_i^0 = \frac{1}{2}$ #poids

$F^0(x_i) = 0$ #fonction

Pour i **de** 1 **à** n **faire** :

$p^0(x_i) = 1/2$

p(x) correspond à la probabilité que y = 1, y indique l'appartenance à la classe

Pour *j* **de** 1 **à** *M* **faire** :

Pour *i* **de** 1 **à** *n* **faire** :

$w_i^j = p^{j-1}(x_i)\left(1 - p^{(j-1)}(x_i)\right)$ #poids

$z_i^j = \frac{y_i^j - p^{j-1}(x_i)}{w_i^j}$ #travail

Créer un arbre de régression *f* minimisant $\sum_{i=1}^{n} w_i^j (z_i^j - f(x_i))^2$

Pour i **de** 1 **à** n **faire** :

$F^j(x_i) = F^{j-1}(x_i) + \frac{1}{2} f^j(x_i)$

$p^J(x_i) = \frac{\exp\left(F^j(x_i)\right)}{\exp(F^j(x_i)) + \exp(-F^j(x_i))}$

Sortie du classifieur : $signe(\sum_{j=1}^{M} f_j(x))$

L'algorithme correspond au classifieur LogitBoost pour n classes. M correspond au nombre de règles de décision à créer. L'adaptation majeure pour notre algorithme de boosting appelé PBoost concerne l'évolution de la probabilité *p* dans l'algorithme. Dans notre algorithme, elle se calcule de la manière suivante :

$$p^j(x) = \frac{2^{(F^j(x) - maxF)*0.9}}{2}$$

avec $maxF = \max_{1 \le i \le n} F^j(i)$

4.2.3 *Les arbres C4.5*

Les règles de décision dites « faibles » utilisées par les algorithmes de boosting peuvent être de nature très diverse comme les arbres de décision, les réseaux

bayésiens, etc. Dans notre étude, nous avons utilisé des arbres de décision C4.5 (R. Quilan, 1993). Il s'agit d'arbres capables de gérer les valeurs manquantes, les variables continues et d'effectuer un post élagage. Lors de la construction de l'arbre, les attributs discriminants sont choisis suivant la valeur du GainSplit à chaque nœud pour chaque attribut. Il correspond au rapport du Gain par l'entropie de Shannon. L'attribut ayant le GainSplit le plus important est considéré comme le meilleur candidat pour la division de l'arbre.

$$E(S) = -\sum_{j=1}^{n} f_S(j) log_2 f_S(j)$$

$$Gain(S, A) = E(S) \ - \sum_{i=1}^{m} f_S(A_i) E(S_{A_i})$$

$$GainSplit = \frac{Gain}{E}$$

avec S l'ensemble des valeurs, n le nombre de valeurs différentes pour l'attribut considéré et f_S(j) la fréquence d'apparition de la valeur j dans S, m le nombre de valeurs différentes pour l'attribut A.

4.2.4 *Correction du regroupement*

Lors de la phase de regroupement, certaines parties de l'image sont sous segmentées. Ce résultat est préférable à la sur segmentation car il est plus facile de regrouper deux zones de même type que de détecter qu'une zone contient des informations de type différent. Dans cette étape, il convient de regrouper des zones de même type étant proche, voire superposées. Pour cela, leur alignement et leur recouvrement sont calculés. Suivant ces valeurs, elles sont ou non regroupées.

5. Expérimentation

5.1 *Description des bases de documents*

5.1.1 *Les zones non textuelles*

Afin d'entrainer un modèle pour la classification des zones non textuelles, une base d'apprentissage et une base de test sont nécessaires. Elles contiennent des zones extraites de documents industriels qui seront présentés dans la partie suivante. L'extraction a été réalisée à l'aide de l'algorithme présenté précédemment. Les zones ont ensuite été étiquetées manuellement suivant ce qu'elles contiennent (logo, tampon, signature, écriture manuscrite, tableau et bruit). Comme l'illustre les figures IV à VIII, leur qualité est très diverse et elles sont plutôt bruitées. La base de données comporte 3048 éléments répartis sur les 6 classes.

Figure V. *Exemple de zones « signature »*

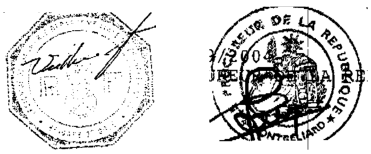

Figure VI. *Exemple de zones « tampon »*

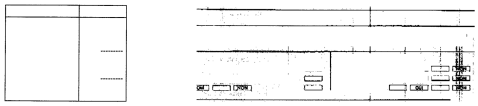

Figure VII. *Exemple de zones « tableau »*

Figure VIII. *Exemple de zones « logo »*

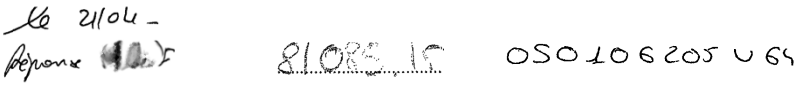

Figure IX. *Exemple de zones « manuscrite »*

5.1.2 *Les pages complètes*

Pour tester notre système, nous utilisons des documents provenant du monde de l'industrie. Il s'agit de cas concret provenant de plusieurs entreprises, leur qualité varie donc d'une page à l'autre. Cependant, ils sont tous au format TIFF, monochromes et scannés principalement à une résolution de 300dpi. Aucune autre information sur le type ou la provenance n'est présente. Pour les tests, 150 pages choisies aléatoirement parmi 2600 sont utilisées. Elles sont toutes analysées par l'OCR Omnipage pour récupérer les mots dactylographiés et leur position. Une segmentation et l'étiquetage des zones ont été réalisés manuellement afin d'obtenir

des documents de vérité. Une tolérance de 10% est accordée pour la segmentation qui, dans certains cas, est difficile à définir.

5.2 *Classification textuelle avec et sans correction*

L'OCR utilisé, Omnipage 16, confond quelques fois le texte dactylographié et d'autres types d'informations, notamment le texte manuscrit. Il est donc nécessaire de vérifier la reconnaissance de l'OCR et de corriger les erreurs. Nous avons effectué les mêmes tests avec l'OCR FineReader 11. Le degré de confusion était encore plus important, c'est pourquoi nous n'avons pas présenté de résultats avec cet OCR.

La figure IX montre des exemples dans lesquels le post-traitement corrige l'étiquette de zones contenant du manuscrit. 82% des mots reconnus à tort par l'OCR sont supprimés par cette méthode d'analyse des mots et d'analyse des pixels. Le principal problème provient d'un mauvais regroupement. En effet, il arrive qu'une même zone contienne du texte dactylographié et un autre type d'information, il devient alors impossible de les séparer. Chaque mot peut être analysé avant regroupement mais étant donné leur taille, les résultats ne sont pas toujours pertinents ce qui rend la classification difficiles à traiter.

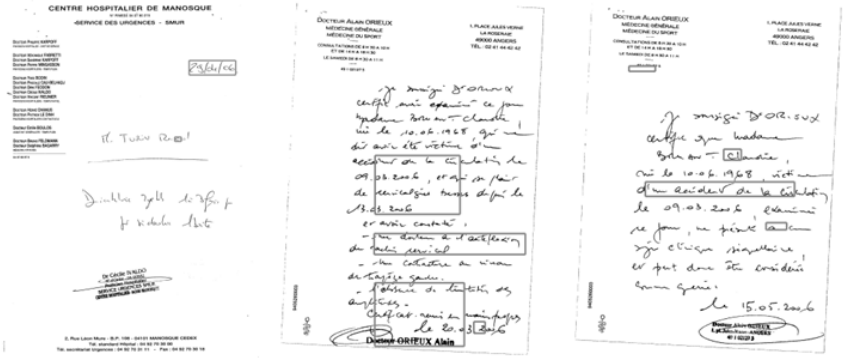

Figure X. *Les zones encadrées en gris sont les zones confondues par l'OCR et corrigées par notre méthode.*

5.3 *Classification des zones non textuelles*

Pour rappel, la classification des zones non textuelles s'effectue suivant 6 catégories : logo, tampon, signature, texte manuscrit, tableau et bruit. Les résultats de classification avec les descripteurs et le classifieur PBoost sont présentés sur la figure X. Le taux de rappel moyen pondéré des 6 classes est de 95,9%. Toutes les classes ont un taux élevé, il n'y a pas de grande disparité.

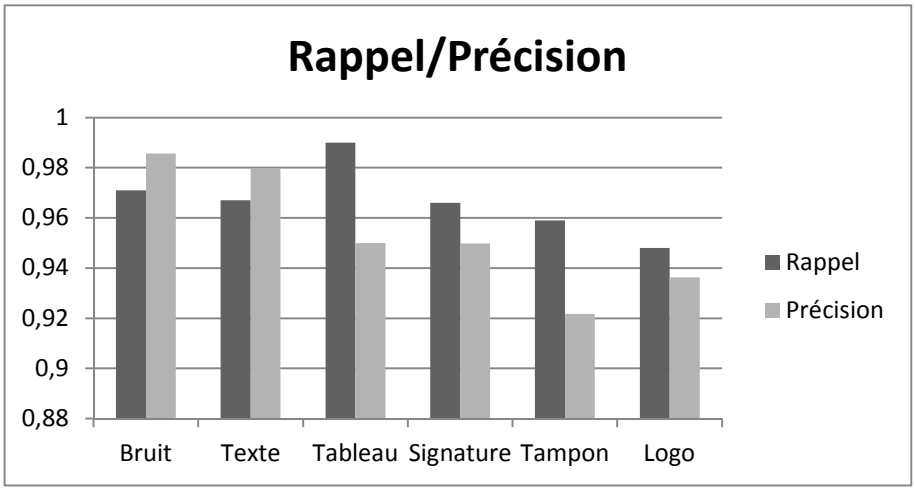

Figure XI. *Résultat de la classification pour les 6 types de zones.*

5.4 ***Comparaison entre LogitBoost et PBoost***

Afin de valider les modifications apportées à PBoost, nous avons testé l'algorithme sur 3 bases de données. La première comporte les valeurs extraites à partir des documents décrits dans la partie 4.1.1. La seconde comporte les informations extraites à partir de la méthode de D. Grzejszczak décrite dans la partie 3.2. Il s'agit d'une base contenant 4200 éléments de textes manuscrits et imprimés. La dernière utilisée est la base de données Mnist (Y. LeCun et al., 1998) en utilisant que les 150 pixels les plus représentatifs obtenus avec l'heuristique de Mark A. Hall cité dans la partie 4.2.1. Le tableau I présente les résultats comparés entre LogitBoost et PBoost avec les mêmes réglages pour ces 3 bases. Ces deux algorithmes sont associés aux arbres de décision C4.5 décrits précédemment.

Base de données	Précision		Rappel		F-mesure	
	PBoost	LogitBoost	PBoost	LogitBoost	PBoost	LogitBoost
Notre base	96,0%	94,9%	95,9%	94,9%	95,9%	94,9%
Grzejszczak	89,6%	88,9%	89,5%	88,9%	89,5%	88,9%
Mnist	93,2%	92,6%	93,2%	92,5%	93,2%	92,5%

Tableau I. *Tableau comparatif entre le PBoost et le LogitBoost.*

Les résultats montrent une amélioration de plus de 1% sur notre base pour la classification de zones après segmentation. Les résultats en utilisant les descripteurs de D. Grzejszczak ou de la base de données Mnist indiquent aussi une amélioration de l'ordre de 0,7%.

5.5 *Comparaison des performances avec les OCRs et résultats*

La comparaison entre notre solution et les solutions commerciales est difficile car elles n'offrent pas le même degré de précision pour la classification. En effet, les OCRs proposent essentiellement 3 catégories qui sont le texte, les tableaux et les images. Nous avons augmenté les possibilités de typage en ajoutant les logos, les signatures, les tampons et le texte manuscrit. La figure XI illustre principalement les erreurs de reconnaissance des zones dactylographiées et la figure XII l'intérêt d'avoir enrichi les types de zones.

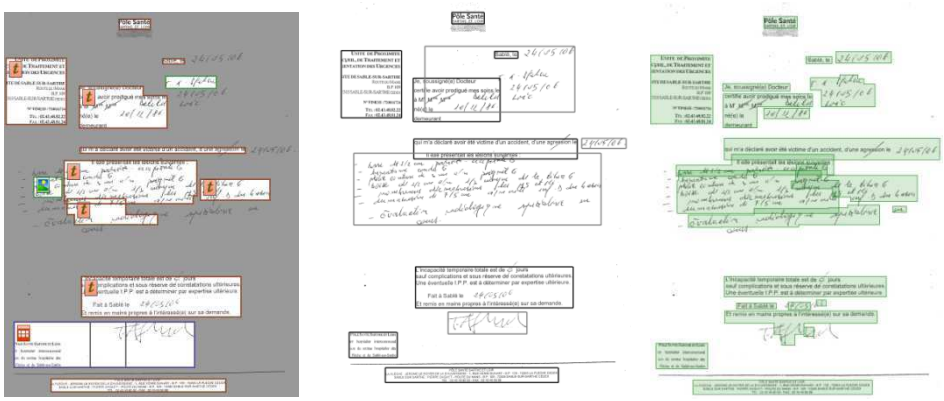

Figure XII. *A gauche, la segmentation proposée par Omnipage ; au milieu, notre segmentation ; à droite, la segmentation de FineReader.*

Nous avons testé notre méthode sur 150 pages annotées manuellement, pour un total de 962 zones. La comparaison pour la segmentation accorde une marge d'erreur de 10%. Si la zone comprend 100px, le nombre de pixels manquants ou présents à tord ne doit pas dépasser 10. Les résultats obtenus sont les suivants : 87,2% des zones sont correctement positionnées, 6,1% sont des zones détectées à tord (sur segmentation ou reconnaissance de bruit) et 6,7% ne sont pas détectées. Parmis les 87,2% des zones correctement positionnées, 3% sont mal étiquetées.

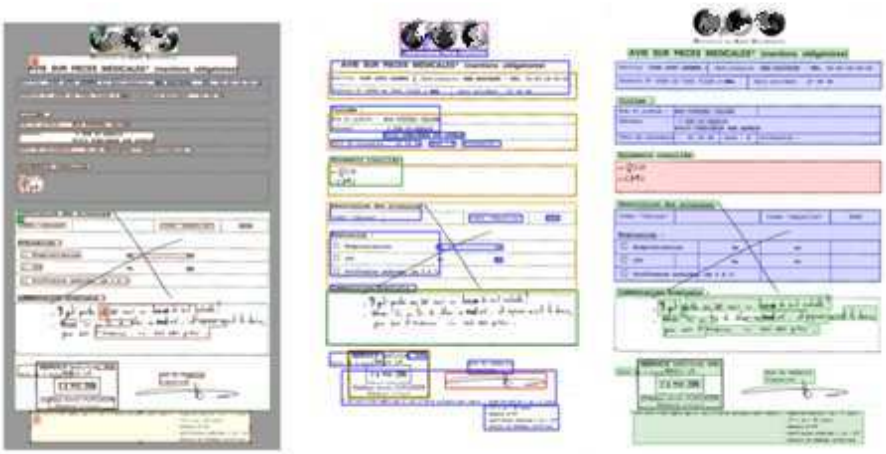

Figure XIII. *A gauche, la segmentation proposée par Omnipage ; au milieu, notre segmentation ; à droite, la segmentation de FineReader.*

6. Conclusion

Nous avons présenté dans cet article une méthode pour la segmentation et la classification des zones de documents complexes. Elle utilise comme point de départ la reconnaissance des mots par un OCR et se base à la fois sur le fond et la forme des documents. Notre méthode de segmentation s'appuie sur le rassemblement des composantes connexes en prenant en compte leur taille et leur éloignement. La classification des zones utilise plusieurs méthodes d'extraction de descripteurs adaptées à nos documents. Elle permet d'obtenir 5 classes : logo, tampon, signature, tableau et texte manuscrit. Enfin, une nouvelle version de boosting a été mise en place afin d'obtenir de meilleurs résultats sur la classification des zones. Elle porte sur la modification de l'évolution de la probabilité de tirage au sort des exemples d'apprentissage. Comparer aux résultats des OCRs, la segmentation fournie par notre méthode permet un meilleur découpage et un typage plus adapté aux zones. Elle est la première étape d'un projet qui a pour but de reconstruire des dossiers. L'étape suivante consiste en la modélisation des documents à partir de la segmentation et de la classification obtenue.

7. Bibliographie

A. Antonacopoulos, S. Pletschacher, D. Bridson and C. Papadopoulos, « ICDAR2009 Page Segmentation Competition », 1370-1374, 2009.

K. Chinnasarn, Y. Rangsanseri, P. Thitimajshima , « Removing Salt-and-Pepper Noise in Text/Graphics Images », *The Asia-Pacific Conference on Circuits and Systems*, p. 459-462, 1998.

Y. Freund. « Boosting a weak learning algorithm by majority ». *Information and Computation*, 256-285, 1995.

J. Friedman, T. Hastie, R. Tibshirani. « Additive Logistic Regression : A statistical view of boosting ». *The Annals of Statistics*, 337-407, 2000.

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

D. Grzejszczak, Y. Rangoni and A. Belaïd. « Séparation manuscrit et imprimé dans des documents administratifs complexes par utilisation de SVM et regroupement », CIFED 2012.

M.A. Hall, « Correlation-Based Feature Selection for Discrete and Numeric Class Machine Learning », *Proc. 17th Int"l Conf. Machine Learning*, pp. 359-366, 2000.

Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. « Gradient-based learning applied to document recognition », *Proceedings of the IEEE*, 2278-2324, November 1998.

G. Nagy and S. Seth. « Hierarchical representation of optically scanned documents », *Proceedings of the 7th International Conference on Pattern Recognition Montreal Canada*, p. 347-349, 1984.

R. J. Quinlan, « Learning with Continuous Classes », *5th Australian Joint Conference on Artificial Intelligence*, Singapore, 343-348, 1993.

R. E. Schapire. « The strength of weak learnability », *Machine Learning*, 197-227, 1990.

L.G. Valiant. « A theory of the learnable », *Artificial Intelligence of Language Processing*, 1134-1142, 1984.

D. Wang and S. Srihari. « Classification of Newspaper Image Blocks Using Texture Analysis », *Computer Vision, Graphics and Image Processing 47*, 327-352, 1989.

Y. Wang, I. Phillips, R. Haralick. « Document zone content classification and its performance evaluation », *Pattern Recognition 39*, 57-73, 2006.

Y. Zheng. « Machine Printed Text and Handwriting Identification in Noisy Document Images », *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 3, 337-353, 2004.