Segmentation logique d'images de journaux anciens

Thomas Palfray, David Hébert, Pierrick Tranouez, Stéphane Nicolas, Thierry Paquet

Segmentation logique d'images de journaux anciens

Thomas Palfray — David Hebert — Pierrick Tranouez — Stéphane Nicolas — Thierry Paquet

Laboratoire LITIS, UFR de sciences
Avenue de l'universite
76800 Saint Etienne du Rouvray
prenom.nom@univ-rouen.fr

RÉSUMÉ. *Nous présentons dans cet article une méthode destinée à la segmentation d'articles dans des journaux anciens. Celle-ci est capable d'analyser des mises en page complexes de documents dégradés. Cette tâche est accomplie à l'aide d'un modèle de Champs Aléatoires Conditionnels permettant d'étiqueter les zones d'intérêt avec un attribut logique. Celles-ci sont ensuite analysées afin de déterminer la structure et l'ordre logique des articles. La méthode repose sur la génération d'une grille de séparation inter articles que nous appliquons sur le document de manière récursive, ce qui permet d'appréhender n'importe quel type de mise en page. Les résultats de cette méthode sont évalués sur une base d'images issues d'un fond de presse régionale quotidienne. Cette méthode est intégrée dans une chaîne de traitement capable de traiter de grandes quantités de documents et permettant de générer des objets numériques au format METS/ALTO.*

ABSTRACT. *We present a method for article segmentation in old newspapers which deals with complex layouts analyzing of degraded documents. This is done using a model of Conditionnal Random Fields for labelling areas of interest with a logical attribute. these interest areas are then analyzed to determine the structure and the logical order of items. The method relies on the generation of an article separation grid applied on the document recursively, allowing to capture any type of page layout. The results of this method are evaluated on images from daily local press. This method is integrated into a worklow which can process large amounts of documents and to generate digital objects in METS/ALTO format.*

MOTS-CLÉS : *Analyse automatique de mise en page; Extraction d'informations à partir de documents numérisés*

KEYWORDS: *Page layout analysis; Information extraction from document images*

1. Introduction

Au cours des vingt dernières années, les archives et bibliothèques nationales du monde entier ont mis en oeuvre de nombreux programmes de numérisation de leurs fonds patrimoniaux afin de préserver ceux-ci tout en facilitant l'accès du public aux informations qu'ils contiennent. Le cas de la presse ancienne est emblématique de cette volonté de diffusion de l'information historique, du fait de la richesse et de la diversité des informations contenues dans les documents de ce type. Néanmoins, ces documents nécessitent des traitements particuliers pour exploiter pleinement leur contenu informationnel, du fait de leur taille, de leur mise en page particulièrement complexe et de l'évolution de celle-ci au fil des innovations techniques des imprimeries, comme l'illustre les exemples de la Figure 1. De plus, la qualité de conservation de ce type de document est souvent inégale selon les périodes historiques du fait des variations importantes de qualité du papier utilisé, résultant en une copie numérisée parfois très dégradée, voir inutilisable. Afin d'exploiter au mieux ces documents, il est nécessaire de disposer d'une segmentation en articles permettant d'isoler des portions intéressantes d'un journal en vue d'une consultation plus aisée par l'utilisateur par le biais des outils numériques modernes. Conscient de ces difficultés, nous avons développé une nouvelle méthode de d'étiquetage logique des journaux anciens destinées à extraire des métadonnées à partir de ces images numérisées grâce à l'utilisation conjointe d'une méthode de classification de séquence de pixels basée sur les champs aléatoires conditionnels, associé à un ensemble de règles définissant la notion même d'article au sein d'un numéro de journal. Nous présentons dans un premier temps les travaux existant pour ce type de tâche, puis nous décrivons la méthode utilisée dans son ensemble, avant de présenter quelques résultats issus de mesures sur des numéros de journaux extraits du Journal de Rouen. Enfin, nous présentons brièvement la chaîne de traitement complète dans laquelle cette méthode est utilisée, avant de conclure par quelques perspectives.

2. Autres travaux

Depuis 2001 s'organise une compétition de segmentation de page lors de la conférence ICDAR (Antonacopoulos *et al.,* 2009) dans laquelle certains algorithmes proposés peuvent sembler avoir des visées semblables aux nôtres. Néanmoins, la base de documents utilisée pour cette compétition est composée de documents contemporains, ce qui peut impliquer une inefficacité des méthodes proposées lorsqu'on les applique sur les journaux anciens. On peut toutefois citer (An *et al.,* 2010) qui utilisent une méthode de classification au pixel. (Breuel, 2002) utilise une approche basée sur les maximum d'espaces blanc pour délimiter les colonnes et les blocs de texte. Cette méthode est utilisée au sein d'OCROPUS[1]. Bien qu'intéressante dans son approche, elle ne permet pas d'appréhender les difficultés de nos documents qui sont souvent inclinés et déformés et qui contiennent de nombreux séparateurs empêchant la détection de ces

1. http ://code.google.com/p/ocropus/

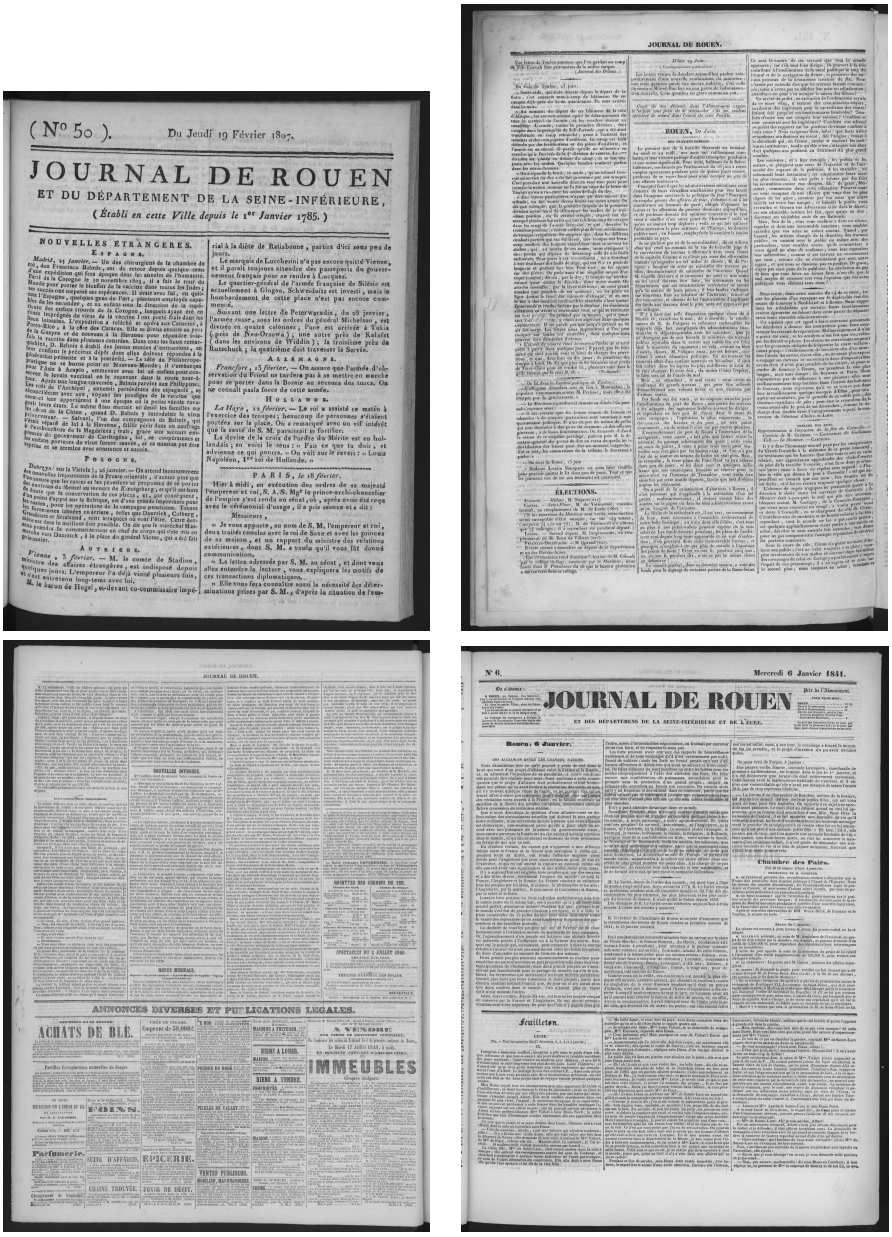

Figure 1. *Quelques exemples de diversité de la mise en page.*

espaces blancs. Une méthode plus intéressante pour répondre à ces difficultés est décrite dans (Lemaitre *et al.*, 2008) et utilise une approche multirésolution pour extraire

les blocs de textes contenus dans des pages de journaux anciens. Malgré son efficacité sur les documents anciens, cette méthode, tout comme les précédentes se contente de délimiter des blocs, sans essayer de les ordonner de manière logique comme nous le proposons ici.

3. Notre approche

Nous travaillons sur une image binaire en entrée du système, la sortie de celui-ci est constituée par un fichier au format METS[2] contenant les articles ordonnés par ordre de lecture. Pour cela, nous voulons retrouver le modèle éditorial du document à l'aide d'indice visuels établis à l'aide de champs aléatoires conditionnels couplés à une analyse de structure par un algorithme récursif. Nous effectuons les tâches suivantes, détaillées dans la suite de cet article :

– Segmentation de l'image par champs aléatoires conditionnels
– Lissage de la segmentation par vote majoritaire
– Extraction des lignes de texte
– Génération d'une grille de séparateurs
– Analyse récursive pour l'extraction des articles et du sens de lecture

3.1. *Segmentation logique de séquence de pixels*

La méthode d'extraction d'articles présentée s'appuie sur une étape de segmentation des document, réalisée grâce à un champ aléatoire conditionnel à quantification multi-échelles (Hebert *et al.*, 2011). Ce système permet d'obtenir une segmentation fine où chaque pixel se voit attribuer une étiquette pour caractériser son rôle dans le document. La segmentation obtenue ne consiste donc pas uniquement en un découpage en blocs physiques car elle apporte également une identification logique aux entités détectées.

3.1.1. *CAC à quantification multi-échelles*

Un Champ Aléatoire Conditionnel à quantification multi-échelles est un modèle de champ aléatoire linéaire discret capable de travailler sur des données continues discrétisées selon plusieurs niveaux de quantification. Une donnée en entrée est quantifiée par un ensemble de fonctions de quantification. Chaque donnée quantifiée est alors donnée en entrée d'un CAC linéaire qui pondère chaque étiquette en tenant compte de toutes les informations quantifiées détectées localement dans l'image. Ce modèle s'écrit de la manière suivante :

2. http ://www.loc.gov/standards/mets/METSPrimerRevised.pdf

$$p(Y|X) = \frac{1}{Z(X)} \exp\left(\sum_{t}^{T}\sum_{k}^{K} \lambda_k f_k(y_{t-1}, y_t, Q_1(o), ..., Q_n(o))\right)$$

X est une séquence de données continues (une séquence d'observations continues), Y est la séquence d'étiquettes associées aux observations de la séquence X. Les fonctions f_k sont des fonctions de caractéristiques (feature functions dans la littérature anglophone des CAC) ici binaires appliquées sur les observations o de la séquence X, quantifiées par n fonctions de quantification $Q_1(o)..Q_n(o)$. Les variables λ_k sont les paramètres du système et pondèrent la connaissance apportée par chaque fonction f_k, ce qui, par extension, permet de pondérer la connaissance apportée par chaque quantification d'une observation donnée. Ce modèle permet de calculer la séquence d'étiquettes la plus probable Y que l'on peut associer à la séquence d'observations X. Pour plus de détails sur ce modèle, le lecteur est invité à consulter (Hebert *et al.*, 2011).

3.1.2. *Le modèle de segmentation*

Le modèle utilisé pour la segmentation des documents est un modèle de séquences de plages horizontales. Une observation est composée de 2 valeurs numériques calculées sur une image binaire : la longueur de la plage horizontale et la longueur médiane des plages verticales s'étalant sur la plage horizontale. Le choix de la valeur médiane a pour but de minimiser l'impact du bruit sur la caractérisation verticale d'une plage de pixels. La figure 2 illustre la caractérisation d'une observation.

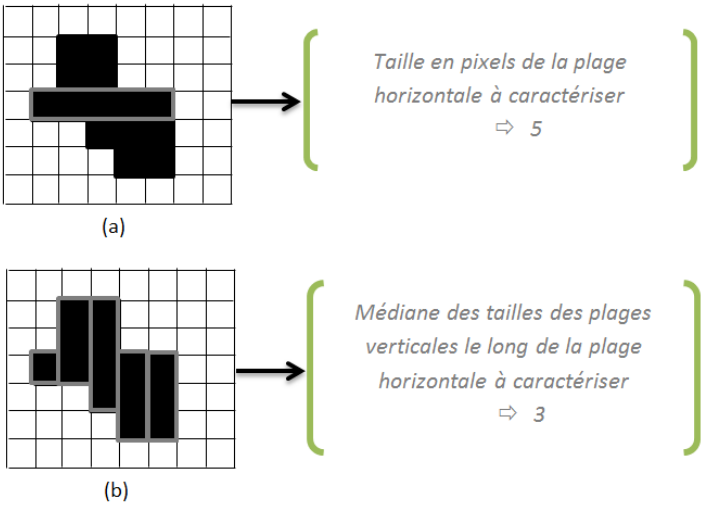

Figure 2. *Caractérisation d'une observation par la longueur de la plage horizontale (a) et la longueur médiane des plages verticales (b)*

Ces caractéristiques possèdent plusieurs avantages : elles sont rapides à calculer, elles permettent d'évaluer un certain contexte et sont adaptées à l'analyse de documents (du fait de leur structure X-Y). De plus, grâce à ces caractéristiques, nous modélisons des séquences de plages et non des séquences de pixels ce qui nous permet de réduire le nombre d'observations et par conséquent le temps de traitement.

Chaque décision locale quant à l'attribution d'une étiquette à une observation est prise en évaluant des fonctions de caractéristiques définies dans un certain contexte. Nous définissons un ensemble de modèles de combinaisons (template ou pattern) pour définir un contexte de 5 observations : si x est la position de l'observation courante, alors la prise de décision quant à l'étiquette à associer s'effectuera en combinant les connaissances apportées par les fonctions de caractéristiques évaluées sur la fenêtre d'observations $[o_{x-2}, o_{x-1}, o_x, o_{x+1}, o_{x+2}]$.

Les étiquettes possibles pour une plage de pixels horizontale sont au nombre de 10. Elles permettent de décrire très précisément l'organisation physique des espaces intra et inter caractères :

– Séparateur vertical
– Séparateur horizontal
– Caractères du titre
– Inter-caractères du titre
– Inter-mots du titre
– Caractères
– Inter-caractères
– Inter-mots
– Bruit
– Fond

Pour l'étiquetage logique des parties d'articles, les étiquettes "Inter-mots", "Inter-caractères" et "Caractères" sont regroupée pour ne former qu'une étiquette "Ligne de texte". De la même manière, les étiquettes "Inter-mots du titre", "Inter-caractères du titre" et "Caractères du titre" sont rassemblées pour ne former qu'une classe "Titre". L'apprentissage du modèle est réalisé sur une base de 9 images étiquetées par les 10 étiquettes énoncées précédemment.

Chaque image est analysée ligne par ligne, l'étiquetage de toutes les lignes est concaténé pour reconstruire une image exploitable, notée $Iseg$, pour la suite des traitements. Un exemple de sortie est visible sur la Figure 3. A l'issue de cette étape on dispose des marqueurs de séparateurs horizontaux et verticaux, des titres et des lignes de texte. Ces marqueurs détectés localement comportent toutefois certaines erreurs. Une analyse complémentaire à l'aide de règles de regroupement fondées sur un modèle éditorial simple et générique va permettre d'analyser la disposition spatiale des différentes entités physiques pour finalement extraire les articles contenus dans le document.

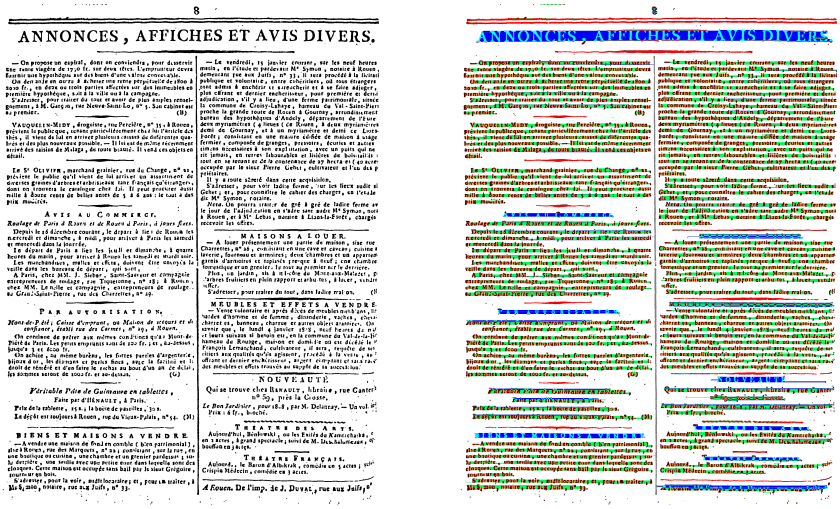

Figure 3. *Exemples de résultat de segmentation sur une page. L'image $Iseg$ est visible à droite.*

3.2. *Post traitement et segmentation logique*

3.2.1. *Lissage de la segmentation par vote majoritaire*

Malgré la capacité du modèle CAC à classer chaque région ou pixel selon une classe du modèle éditorial sous-jacent, des erreurs d'etiquetage apparaissent, comme nous pouvons le voir sur la Figure 3. Nous appliquons un premier algorithme destiné à lisser le résultat de la segmentation en effectuant un vote à la majorité des étiquettes sur les pixels noirs de l'image traitée comme décrit dans l'Algorithme 1. Une illustration des résultats de cette méthode est visible sur la Figure 4.

Algorithme 1 : Vote à la majorité pour les composantes connexes

Entrées : $Iseg$: image résultat de segmentation
Sorties : $Ientitee$: image des entitées extraites
Extraire les composantes connexes constituée de toutes les étiquettes connexes hors étiquette de fond de Iseg;
pour tous les *composantes connexes* **faire**
  $nEi \leftarrow$ le nombre de pixels étiquetés Ei;
attribuer à tous les pixels de CC l'étiquette $argmax_{Ei}(nEi)$;

Une fois l'image $Ientitee$ entièrement construite, nous pouvons extraire les lignes de texte et les articles qui composent le document.

Figure 4. *Résultat de segmentation corrigé par l'algorithme 1. L'image $Ientitee$ est visible à droite.*

3.2.2. *Extraction des lignes de texte*

L'extraction des lignes de texte est réalisée par extraction des composantes connexes appartenant à l'étiquette *texte* de l'image Icc. Malgré la robustesse de notre méthode d'extraction, il peut arriver cependant que plusieurs lignes soient connectées entre elles à cause d'une déformation trop importante du document. Afin de résoudre ce problème, nous proposons de détecter les lignes susceptibles d'être incorrectes en calculant l'aire moyenne de toutes les lignes de textes contenue dans un numéro de journal. Puis, nous considérons que celles dont l'aire est supérieure à l'aire moyenne sont probablement incorrectes. Nous appliquons alors un algorithme spécifique permettant de corriger ces lignes. Celles-ci sont stockées avec les autres en attendant d'être associées aux blocs d'articles extraits par notre méthode de segmentation.

3.2.3. *Extraction des articles à l'aide d'un modèle éditorial*

3.2.3.1. Définition d'un article

Comme nous l'avons dis précédemment, nous définissons sur notre image des zones d'intérêt représentant les entités suivantes : *titre*, *texte*, *séparateurs horizontaux* et *séparateurs verticaux*. La notion d'article répond à des règles de mise en page précises, ainsi nous considèrerons qu'un article commence par une entité *titre* qui est suivie par au moins une entité *texte* et se termine par une entité *séparateur horizontal* ou une autre entité *titre*. A cette définition de base vient s'ajouter le cas particulier

d'articles s'étendant sur plusieurs pages. Nous considèrons ainsi qu'une entité *texte* située en première ou dernière position d'une liste d'articles est reliée à un article de la page précédente ou suivante du numéro de journal en cours d'étude.

3.2.3.2. Définition d'une grille de séparateurs

Les séparateurs horizontaux et verticaux des journaux constituent une information robuste, mais parfois morcellée du fait des dégradations du document, sur laquelle nous souhaitons nous baser pour extraire les articles qui les composent. En effet, les séparateurs verticaux délimitent les colonnes du document, tandis que les séparateurs horizontaux découpent les articles au sein de ces colonnes, tout en délimitant également les différentes portions de la page. Les zones de titre sont également des informations discriminantes puisqu'elles indiquent le début d'un article. On peut donc considérer le document analysé sous la forme d'un arbre de blocs représentant les différentes portions de celui-ci. Ainsi la racine de l'arbre est constitué d'un seul bloc représentant toute la page, le niveau d'en dessous est constitué par les blocs délimités par les séparateurs horizontaux les plus larges etc. On appelle hiérarchie la position d'un bloc ou d'un séparateur dans l'arbre de la structure du document, celle-ci est visible sur la Figure 5. L'ensemble des séparateurs présents sur une page constitue donc une grille logique décrivant la structure du document. Nous souhaitons exploiter cette information afin de définir une liste d'articles articulés logiquement entre eux et classés par sens de lecture, ceci en fonction de leur position dans le document, mais également en fonction de leur position hiérarchique dans l'arbre de la structure de ce document. Chacun de ces articles est composé d'une ou plusieurs cases de notre grille de séparateurs.

3.2.3.3. Génération de la grille de séparateurs et formation des blocs de textes

La première étape de notre méthode de segmentation consiste donc à prolonger l'ensemble des zones d'intérêt étiquetées séparateurs et titre, en vue de générer une grille de séparateurs représentant l'ensemble des *articles* du document. Nous appliquons l'ensemble des règles suivantes et dans cet ordre :

– Création du masque de séparateurs horizontaux et verticaux.

– Connecter les séparateurs verticaux proches dans la direction verticale.

– Prolonger les séparateurs verticaux tant qu'ils ne croisent pas un séparateur horizontal ou un titre.

– Connecter les séparateurs horizontaux proches dans la direction horizontale.

– Prolonger les séparateurs horizontaux et les titres tant qu'ils ne croisent pas de séparateurs verticaux.

Nous obtenons ainsi la grille recouvrant toute l'image traitée. Nous pouvons alors analyser celle-ci en vue d'extraire les articles de journaux. Pour cela, nous génèrons la liste des blocs qui sont les cases de notre grille, nous comparons les coordonnées de ces blocs aux coordonnées des lignes de textes extraites en 3.2.2 ainsi que les lignes de type *titre* et nous ajoutons à ces blocs les lignes de textes incluses dans ceux-ci.

Nous rejettons les blocs ne contenant aucune ligne de texte de la liste. Nous disposons alors d'une liste de cases positionnés sur la grille de séparateurs, il reste à exploiter cette liste de cases pour obtenir la liste ordonnée des articles selon l'ordre de lecture du modèle éditorial. La Figure 6 illustre la notion de sens de lecture au sein d'un document exemple.

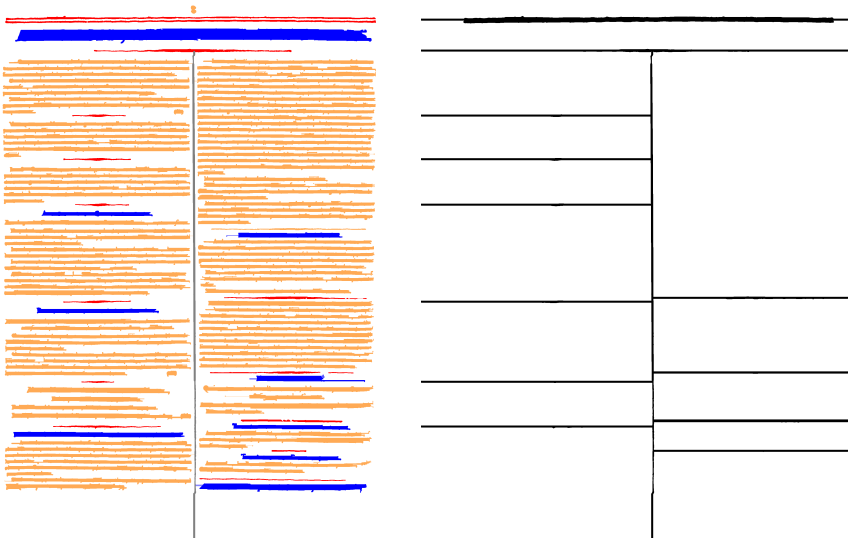

Figure 5. *Exemple de grille de séparateurs issue de l'image précédente.*

3.2.3.4. Extraction des articles et détection de l'ordre de lecture

Ces cases précédemment définies contiennent des articles ou des *demi-articles* : quand le texte finit sur le bas de la section et reprend en haut de la colonne suivante, on appelle demi-article chacune de ces portions. Pour pouvoir les assembler, il faut avoir perçu le sens de lecture au sein de la section qui les contient.

Comme expliqué dans 3.2.3.2, les séparateurs divisent la page en sections, puis celles-ci en sous-sections etc. jusqu'au niveau bloc qui sont les feuilles de cet arbre. En 3.2.3.3 nous avons construit les cases : nous allons à présent reconstruire les autres sections.

Pour chaque section s, on cherche le séparateur horizontal strictement plus large qu'elle, situé au-dessus d'elle, et le plus proche. Ce séparateur horizontal délimite deux sections : la section père est celle parmi ces deux qui contient s. On répète ce processus jusqu'à pouvoir remonter dans cet arbre de toutes les cases de 3.2.3.3 jusqu'à la page.

Au sein de chaque section qui n'est pas une feuille, on ordonne les fils : une section est inférieure à une autre si et seulement si son coin haut gauche est strictement plus

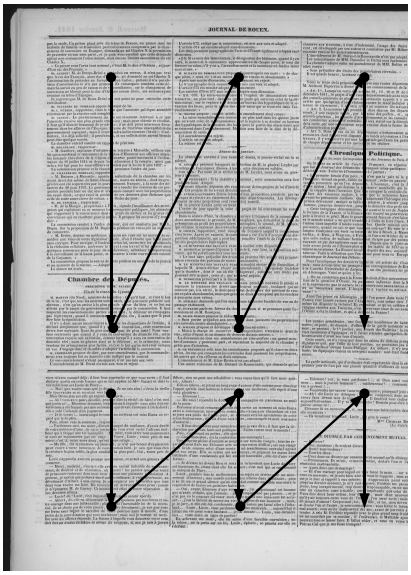

Figure 6. *Exemple d'ordre de lecture.*

à gauche ou de même abscisse mais plus haut. Cette relation d'ordre est calculée avec une précision relative à celle qu'on peut attendre pour le placement des cases, qui sont issues rappelons-le de toute une chaîne de traitements qui ne sont pas exacts au pixel près. Si certains fils d'une section sont des feuilles, cette relation d'ordre induit un sens de lecture de ces feuilles. Si le délimiteur d'une case est un séparateur strictement plus large que lui, et qu'il ne contient pas de titre, c'est qu'il est la deuxième partie d'un demi-article. Il sera alors assemblé avec la case qui le précède immédiatement pour faire un article.

À la fin de cette étape nous avons donc l'ensemble de nos articles, ordonnés dans chaque section.

4. Résultats

Cette méthode a été testée sur une base de 42 images extraites du corpus du Journal de Rouen. Les résultats ont été comptabilisés visuellement en étudiant les images, car nous ne disposons pas de vérité terrain permettant de les vérifier automatiquement ceux-ci. Les résultats de cette méthode sont présentés sur le Tableau 1.

L'analyse des erreurs de notre méthode nous permet d'affirmer que la majeure partie de celles-ci provient d'erreurs d'etiquettage dues au CAC, par exemple dans notre base, nous avons 14 lignes de *texte* etiquetées par erreur comme étant du *titre,* ce qui

Figure 7. *Articles extrait en combinant la grille de séparateurs aux informations sur les lignes.*

Art. présents	Art. détectés	Art. corrects	% correct	% surseg.
226	245	194	85.84%	8.41%

Tableau 1. *Résultats de segmentation logique en articles en utilisant la méthode de grille de séparateurs*

induit la création d'un nouvel article lorsque nous appliquons l'ensemble des règles éditoriales décrites en 3.2.3.1 à chaque fois, provoquant ainsi une sursegmentation de 28 articles là où le système ne devrait en trouver que 14.

5. Chaîne de traitement

Cette méthode fait partie intégrante d'une chaîne de traitement destinée à traiter de grandes quantitées de documents de type journaux anciens. Cette chaîne effectue sur les images de documents les opérations de binarisation et de redressement nécessaires avant d'appliquer aux données la méthode décrite dans cet article. Puis la chaîne utilise les lignes de texte issues de la méthode pour alimenter un système de reconnaissance optique de caractères (OCR) développé au sein du laboratoire LITIS (Ait-Mohand *et al.*, 2010) chargé de reconnaître le texte contenu dans celle-ci. Nous pouvons également choisir d'associer des données d'OCR déjà générée par une application tierce.

En sortie, notre chaîne de traitement génère des fichiers XML au format METS/ALTO contenant la structure logique de l'enchaînement des articles, mais également la structure physique constituée par les lignes de texte détectées et reconnues par OCR. Les fichiers ainsi obtenus peuvent être utilisés à des fins d'archivage ou mis en ligne pour une consultation aisée par le public. Nous disposons à cette fin d'une application web de consultation et de recherche de journaux anciens[3] qui exploite les données extraites par notre système.

6. Conclusion et perspectives

Nous avons présenté dans cet article une méthode de segmentation logique fondée sur l'analyse du résultat d'un étiquetage bas niveau fourni par un champ aléatoire conditionnel, par un ensemble de règles de regroupement exploitant un modèle éditorial générique. Cette méthode est capable de segmenter le texte contenu dans les documents de type journaux anciens multi colonnes avec un minimum de règles éditoriales. Elle donne un taux de segmentation logique de 85.84% sur la base de test de 42 images issues du Journal de Rouen.

Ce premier résultat est encourageant et nous permet de dégager deux pistes principales d'amélioration. La première consiste à améliorer le modèle basé sur les champs aléatoires conditionnels, puisque nous avons vu que la majorité des erreurs provient de cette étape de notre méthode. Dans un deuxième temps nous devrons faire évoluer tant le modèle de champ aléatoire que les règles éditoriales pour prendre en compte les autres éléments structurant du journal comme les figures, les illustrations, les légendes d'images et les tableaux.

Remerciements

Le projet "Plateforme d'Indexation Regionale" (PlaIR) est financé par la région de Haute Normandie et l'Europe dans le cadre des fonds FEDER. Il est soutenu par le CHU de Rouen ainsi que l'Université de Rouen.

7. Bibliographie

Ait-Mohand K., Heutte L., Paquet T., Ragot N., « Adaptation de modèles de Markov cachés-Application à la reconnaissance de caractères imprimés », *Proceedings of CIFED 2010 Conference*, 2010.

An C., Yin D., Baird H., « Document Segmentation Using Pixel-Accurate Ground Truth », *2010 International Conference on Pattern Recognition*, IEEE, p. 245-248, 2010.

3. démonstration disponible sur : http ://plair.crihan.fr

Figure 8. *Résultats de notre méthode sur les exemples montrés sur la Figure 1.*

Antonacopoulos A., Pletschacher S., Bridson D., Papadopoulos C., « ICDAR 2009 Page Segmentation Competition », *2009 10th International Conference on Document Analysis and Recognition*, IEEE, p. 1370-1374, 2009.

Breuel T., « Two geometric algorithms for layout analysis », *Document Analysis Systems V*, vol. 2, p. 687-692, 2002.

Hebert D., Paquet T., Nicolas S., « Continuous CRF with multi-scale quantization feature functions Application to structure extraction in old newspaper », *Document Analysis and Recognition (ICDAR), 2011 International Conference on*, IEEE, p. 493-497, 2011.

Lemaitre A., Camillerapp J., Couasnon B., « Approche perceptive pour la reconnaissance de filets bruités, Application à la structuration de pages de journaux », *in* , A. T. et Thierry Paquet (ed.), *Dixième Colloque International Francophone sur l'Ecrit et le Document*, Groupe de Recherche en Communication Ecrite, France, p. 61-66, 2008.