Segmentation et analyse interactives de documents anciens imprimés

Interactive segmentation and analysis of historical printed documents

J.Y. Ramel, S. Leriche
Lab. d'Informatique, École Polytechnique de l'Université de Tours, 64, avenue Jean Portalis 37200 Tours

Manuscrit reçu le 19 mai 2004

Résumé et mots clés

Après avoir caractérisé les spécificités de mise en page dans les ouvrages imprimés anciens, nous montrons par une campagne d'expérimentations que les méthodes ascendantes et descendantes d'extraction de la structure physique apportent des informations différentes qu'il ne faut pas ignorer lorsque l'on désire segmenter de manière optimale des documents anciens. Les tests réalisés mettent également en évidence les sources d'erreurs des méthodes traditionnelles. Partant de ces constatations, notre proposition consiste à utiliser un algorithme de segmentation hybride basé sur la construction de deux représentations de l'image : une carte des formes qui se focalise sur les composantes connexes présentes dans l'image et une carte du fond qui fournit de l'information sur les espaces blancs séparant les blocs constituant la page. Ensuite, sur la base de la segmentation obtenue à l'aide de cette méthode, une classification des blocs extraits peut être réalisée selon des scénarios que l'utilisateur met en place en fonction de ses besoins. Ces scénarios sont définis simplement grâce à une phase d'interaction entre l'utilisateur et le système et permettent de concevoir des chaînes de traitements adaptées aux différents types d'images que l'on peut rencontrer.

Analyse de la structure physique, segmentation, classification des blocs, documents anciens, stratégie d'analyse.

Abstract and key words

In this paper, we first precise the main error sources from classical methods of structural page layout analysis based on a study of the specificity of old printed books. We show that each type of methods (bottom-up and top-down) provides different kinds of information that should not be ignored to obtain both a generic method and good segmentation results. Then, we propose to use a hybrid segmentation algorithm. We build two maps : a shape map that focuses on connected components and a background map that provides information on white areas corresponding to block separation in the page. Then, using this first segmentation, a classification of the extracted blocks can be achieved according to scenarios built by the user. These scenarios are defined very simply during an interactive stage allowing the users to produce processing sequences adapted to the different kinds of images they can meet and to their needs. The method gives very good results while the setting of parameters is easy and not sensitive to low variations.

Page layout analysis, segmentation, block classification, old documents, analysis strategy.

Remerciements

Nous tenons à remercier l'ensemble des personnes ayant collaboré à ce projet et plus particulièrement M.L. Demonet et S. Busson (du CESR de Tours) pour la confiance et l'aide qu'ils nous ont, et qu'ils continuent à nous accorder.

1. Introduction

Nous présentons dans cet article des travaux effectués dans le cadre d'un projet avec le Centre d'Etudes Supérieures de la Renaissance (CESR) de Tours qui souhaite créer une Bibliothèque Virtuelle Humaniste. Actuellement, n'est accessible que la version image de certains ouvrages scannés ou photographiés page par page. Notre projet vise à mettre à disposition, d'une part, une version texte, et d'autre part, un ensemble de méta-données d'indexation pour faciliter les recherches et rendre la lecture des ouvrages numérisés plus conviviale. Dans un premier temps, nous étudions les spécificités des ouvrages anciens pour tenter d'en déduire des caractéristiques stables utilisables durant l'analyse automatisée de leur structure. Nous décrivons, ensuite, les méthodes d'extraction de la structure physique applicables sur de tels documents en mettant en exergue leurs qualités et leurs défauts. Nous explicitons également les adaptations nécessaires aux traitements des documents anciens avec ces méthodes traditionnelles. La seconde partie de l'article propose une nouvelle méthode hybride d'extraction de la structure physique basée sur la construction de deux représentations du contenu des images correspondant, d'une part, à une cartographie des formes et, d'autre part, à une cartographie du fond. En exploitant ces informations, notre algorithme produit une liste de zones constituant une première proposition de segmentation de l'image et servant de base pour la suite de l'analyse.

Visant un objectif de généricité, l'architecture informatique du système que nous avons réalisé autorise la mise en place interactive de scénarios d'analyse du contenu de la représentation intermédiaire obtenue durant la phase de segmentation. Ensuite, en fonction de ses besoins (localisation des lettrines, des notes en marges, …) et à l'aide d'interfaces conviviales, l'utilisateur (non expert en traitement d'images) construit des scénarios permettant d'étiqueter, de fusionner et de supprimer les zones contenues dans la représentation intermédiaire. Il localise ainsi les entités qui l'intéressent en ignorant les autres régions de l'image. Les scénarios élaborés peuvent ensuite être sauvegardés, modifiés et appliqués sur différentes images lors de traitements par lots.

2. Études préliminaires

2.1. Caractéristiques des ouvrages anciens

2.1.1. Techniques de réalisation

Depuis sa création, le CESR a constitué un fonds d'ouvrages anciens qui compte actuellement environ 3 000 ouvrages, couvrant la période s'étendant du milieu du XIV^e^ siècle au début du XVII^e^ siècle. Les premiers ouvrages remontent aux débuts de l'imprimerie, les polices utilisées, la présentation des pages et l'utilisation de l'espace étaient alors très proches de celles des ouvrages manuscrits. Le fonds du CESR provient de toute l'Europe : France, Allemagne, Italie, Suisse, Hollande. Les langues utilisées, le latin ou le français, amènent un facteur supplémentaire de diversité pour les ouvrages. Les typographies médiévales ont une variabilité de forme importante. Des exemples d'images d'ouvrages anciens sont présentés figure 1.

Pour la mise en page, les contraintes techniques imposent des présentations particulières. La variabilité des mises en page est beaucoup plus grande que celle des ouvrages actuels à cause, soit d'imprécisions, soit de libertés prises par l'imprimeur. La plupart du temps, un corps de texte occupe la majorité de la page, avec des notes en marge de chaque côté du texte. La page peut aussi contenir des zones graphiques de différentes tailles et des lettrines. Dans le texte, on retrouve des structures connues comme les titres et les sous-titres, les paragraphes, les numéros de page, et d'autres structures plus particulières comme les réclames. Les styles employés peuvent alterner un style normal justifié ou aligné à gauche. Une autre particularité provient des faibles séparations entre les blocs de texte (notes en marge et corps du texte par exemple). Enfin, sur certaines pages, les règles de mise en page d'aujourd'hui ne sont pas respectées : par exemple, une illustration peut déborder sur les marges (figure 1). Dans les ouvrages de la Renaissance, les illustrations ont généralement été imprimées à l'aide de plaques de bois ou de métal, gravées avec le motif à reproduire et encrées. Elles sont généralement incluses dans un rectangle blanc souvent entouré de texte.

2.1.2. Techniques d'acquisition

D'autres particularités sont dues aux procédés de numérisation employés. Elles concernent les défauts d'éclairage dans la reliure, la courbure des lignes de texte, l'inclinaison des pages, l'élimination imparfaite des taches. De nombreuses recherches ont été menées, notamment lors du projet Debora, pour corriger ces défauts [Trinh03]. Des solutions commerciales existent et sont même considérées comme satisfaisantes (Book Restorer) bien que des problèmes subsistent. Par exemple, la correction des

Figure 1. Exemples de pages d'ouvrages.

défauts de courbure peut entraîner une dégradation sur les frontières des blocs de texte et des images. Dans ce travail, nous avons choisi de ne pas aborder ces problématiques bas niveau mais de mettre en place des algorithmes surmontant ces difficultés dues aux techniques d'acquisition.

2.1.3. Caractéristiques retenues

Ces quelques remarques nous permettent de dresser une liste de caractéristiques dont il est indispensable de tenir compte lors de la conception d'algorithmes d'extraction de la structure physique de documents anciens. Voici donc les connaissances *a priori* que nous avons utilisées lors de la mise en place de nos algorithmes :

- Caractéristiques liées aux techniques d'imprimerie et aux conventions de structuration de l'époque :
• Mise en page complexe qui peut présenter plusieurs colonnes avec des tailles de corps et d'interligne irrégulières
• Pas de charte éditoriale ou de structure logique identifiable
• Présence de notes en marges imprimées ou manuscrites
• Présence d'indicateurs de repérage : numéros de lignes, de pages, réclames, …
• Utilisation de fontes particulières
• Usage fréquent d'ornements (zones non textuelles) tels que bandeaux, lettrines, enluminures, …
• Disposition fluctuante des illustrations graphiques et des légendes associées
• Faible espacement entre les lignes provoquant des contacts entre les caractères
• Espacement entre caractères, mots et blocs de texte non constant
- Caractéristiques des images liées à la nature des documents :
• Images encore dégradées même après restauration (apparition des caractères du verso)
• Présence de superposition de couches d'informations (tampon, notes manuscrites, …)

2.2. Évaluation des méthodes existantes

La typographie et la technologie de l'imprimerie ont fait énormément de progrès et les ouvrages actuels répondent à des normes bien différentes en matière de présentation. Les logiciels conçus pour reconnaître les documents actuels s'avèrent, en conséquence, bien souvent médiocres pour le traitement des ouvrages de la Renaissance. Les méthodes d'extraction de structures employées par ces logiciels peuvent être classées en trois grandes catégories : les méthodes ascendantes, descendantes et mixtes [Belaid97]. Ainsi, avant de décrire notre système, nous présentons les résultats d'une campagne d'expérimentations qui nous a permis d'évaluer les différentes approches traditionnelles de segmentation sur les documents imprimés anciens.

2.2.1. Méthodes ascendantes

Pour être classée parmi les méthodes ascendantes, une méthode doit travailler à partir des pixels. Dans cette catégorie, on trouve les méthodes se basant sur des techniques de filtrages morphologiques ou différentiels et sur l'étude des composantes connexes de l'image.

Filtrages morphologiques et différentiels

Ce type de méthodes a été beaucoup mis en œuvre et testé dans le cadre du projet Débora pour la fusion des caractères en mots puis en lignes et pour l'élimination du bruit [Lebourgeois03]. La figure 2 illustre les résultats pouvant être obtenus à l'aide de ce type de méthodes. Le classique algorithme RLSA (Run Length Smearing Algorithm) [OGorman95] peut également être utilisé; il a un fonctionnement et des résultats similaires à ceux des méthodes morphologiques.
Il est également possible de travailler directement sur les images en niveaux de gris par filtrage différentiel. L'idée est d'utiliser des filtres permettant d'agglomérer les variations d'intensité périodiquement produites par les contours des caractères puis de rechercher des alignements horizontaux pour les lignes de texte.

Étude des composantes connexes

Ces méthodes considèrent chaque page comme un ensemble de composantes connexes. Sur la figure 3, on peut constater que la taille, la proximité et la position relative des composantes connexes peuvent être utilisées pour extraire la structure physique d'une page. Les rectangles circonscrits aux composantes connexes se chevauchent fréquemment dans les graphiques et rarement dans les textes. Ainsi, les zones graphiques correspon-

Figure 2. Filtrage morphologique sur une image binarisée.

Figure 3. Composantes connexes et segmentation.

dent aux composantes connexes de cet ensemble de rectangles circonscrits dont les dimensions (largeur ou hauteur) dépassent un seuil choisi souvent empiriquement par les experts.
Un bloc de texte est un ensemble de petites composantes connexes « proches ». Deux caractères sont dits voisins si la distance les séparant est inférieure à un espace maximal. O'Gorman a proposé d'utiliser uniquement le voisinage entre composantes pour localiser les zones de texte dans une image [OGorman93]. Nous avons adapté cette méthode en cherchant, pour chaque composante, les composantes connexes dans les quatre directions principales du voisinage défini. Une composante n'ayant pas de voisin dans au moins une direction est une composante de contour d'un bloc. Un chaînage circulaire fermé traduit alors la frontière d'un bloc de texte (figure 4).
Comme les méthodes à base de filtrage, ces méthodes sont peu adaptées aux ouvrages anciens à cause de la proximité entre les blocs de texte. Pour la localisation des zones graphiques, les seuils doivent correspondre à une taille légèrement supérieure au plus grand caractère contenu dans l'ouvrage. Si des études statistiques peuvent permettre d'automatiser la sélection de ces seuils, ces méthodes supportent difficilement les variabilités de taille des caractères et de mise en page qui peuvent apparaître fréquemment dans les ouvrages anciens. La mauvaise qualité des images (bruits, tâches, …) pose aussi certains problèmes.
Kiochi Kise [Kise98] propose d'utiliser un pavage de Voronoï plutôt que les composantes connexes pour extraire les blocs de texte d'un document. Le pavage permet de construire un graphe de voisinage simplifiant la localisation des lignes de texte par des regroupements des formes proches.
Les problèmes liés aux méthodes ascendantes viennent des nombreux paramètres difficiles à régler et des temps de calcul souvent prohibitifs. De plus, ces méthodes ne permettent pas de traiter correctement les documents dans lesquels, les espaces entre formes (caractères, mots, …) sont trop fluctuants et les images trop dégradées. Ainsi, comme le montre la figure 2 et les conclusions du projet Debora, il semble impossible de bien segmenter les mots et les paragraphes dans les documents anciens en utilisant uniquement une approche ascendante.

2.2.2. Méthodes descendantes

Nous venons de voir qu'il était difficile de reconstruire les blocs de texte d'une page à l'aide de critères locaux. Partant du point de vue inverse, les approches descendantes tentent de localiser les séparations (espaces) entre les blocs de manière plus globale et à l'aide de plus de connaissances *a priori*. En effet, deux blocs de texte sont nécessairement séparés par un espace blanc de surface importante, que ce soit dans le sens horizontal ou vertical.

Projections horizontales et verticales

La recherche de ces séparations blanches est souvent faite horizontalement et verticalement. Une localisation des séparations horizontales (espaces entre les lignes et entre les paragraphes), précédant une recherche des séparations verticales sur chacun de ces blocs horizontaux augmente la robustesse de l'analyse.
La localisation des séparations blanches se fait généralement par analyse de la forme de l'histogramme du nombre de pixels noirs sur les lignes et les colonnes de l'image. On peut considérer une différence importante entre deux valeurs successives dans l'histogramme comme une délimitation entre deux blocs de texte. Le problème est l'estimation des différences significatives dans l'histogramme. En général, on utilise des connaissances *a priori* sur le document en cours d'analyse (nombre de colonnes, marges, …) pour localiser plus facilement les séparations. Sur les documents anciens, nos tests ont montré que la qualité des résultats augmentait en noircissant les rectangles circonscrits aux composantes connexes de l'image avant de calculer l'histogramme.
Parmi les méthodes proposées pour traiter les mises en page complexes des documents composites (journaux, pages de magazines, …), nous avons adapté la méthode Split and Merge proposée par Hadjar [Hadjar01] dans le cadre du « segmentation contest » de ICDAR'01. Notre but est de localiser les zones blanches homogènes dans une image (figure 5). Cette méthode peut être comparée à d'autres techniques à base de pavages plus ou moins évoluées, comme la méthode Recursive XY-Cut [Nagy84] [Akindele93]. Dans tous les cas, il est nécessaire de faire appel à des connaissances *a priori* sur le document pour bien définir le critère d'homogénéité utilisé. De plus, l'analyse du voisinage (réalisée avec l'aide de graphes ou de quadtrees) de chaque région blanche découverte par découpage permet, ensuite, de localiser et de caractériser les blocs contenus dans la page à un niveau plus ou moins fin (figure 5).
Le problème de ce type de techniques concerne la différenciation entre les minima locaux et globaux dans l'histogramme. De plus, certaines séparations correspondent à des minima non significatifs (cas des images mal redressées). Il est donc difficile d'utiliser ces méthodes lorsque la mise en page est variable ou lorsque les pages sont mal redressées.

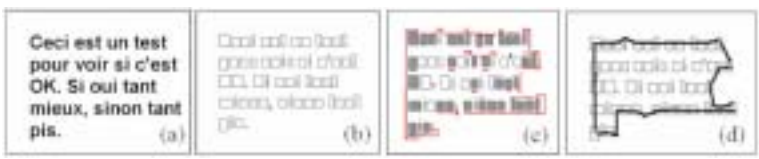

Figure 4. (a) Image initiale, (b) composantes connexes, (c) en gris : composantes de contour, en rouge : côtés sans voisin, (d) chaînage circulaire obtenu.

Figure 5. Split & Merge pour localiser les zones blanches d'un document.

Multi-résolution et analyse de texture

Les méthodes consistant à réduire la résolution de l'image ou à analyser les différentes textures contenues dans l'image sont difficilement utilisables pour atteindre notre objectif de segmentation en blocs, principalement à cause de la proximité entre les différents blocs de texte. Les espacements deviennent indétectables lorsque l'on diminue la résolution et les zones de texte ne présentent aucune différence de texture significative. Ces techniques s'avèrent néanmoins intéressantes lorsque l'on cherche à caractériser globalement à l'aide de différents indices une zone particulière.

Ces méthodes descendantes sont moins sensibles au bruit que les méthodes ascendantes. Lorsque les images sont bien redressées, elles résolvent en partie le problème de proximité entre les blocs et entre les caractères (méthodes à base de pavage). Néanmoins, elles sont difficilement applicables sur les documents ayant une mise en page fluctuante ou non rudimentaire puisqu'elles nécessitent l'utilisation de connaissances *a priori* [Hadjar02] sur le document (nombre de colonnes, critères d'arrêt du découpage, …). Elles ne sont donc pas adaptées au traitement des documents anciens qui ne respectent pas une charte éditoriale précise.

2.2.3. Méthodes mixtes

Les méthodes mixtes proposent d'utiliser les avantages des méthodes ascendantes et des méthodes descendantes par la mise en place d'architectures d'analyse mixant ces deux types d'approches. La plupart des tentatives de mise en place de tels systèmes se sont traduites par l'obtention, soit d'une « usine à gaz », soit d'un système dédié à une classe de documents bien spécifique. Nagy et Al [Nagy93] ont, par exemple, développé une architecture logicielle utilisant une grammaire pour spécifier des règles explicitant comment agglomérer les pixels d'une image pour construire des objets de niveaux sémantiques de plus en plus élevés. La construction et l'étiquetage des blocs sont alors réalisés simultanément en utilisant une approche mixte. Il devient possible de produire des grammaires adaptées aux différentes classes de documents mais la mise en place de ces grammaires reste très laborieuse et complexe et rend ce système peu générique.

3. Une nouvelle méthode de segmentation

Les tests précédents ont mis l'accent sur les limites des méthodes traditionnelles pour traiter des documents anciens ainsi que les raisons de leur échec. Sur cette base, nous avons mis au point une nouvelle méthode exploitant à la fois les atouts des méthodes descendantes et des méthodes ascendantes. Nous nous basons sur la construction d'une carte du fond mettant en évidence les frontières entre les blocs présents dans la page (approche descendante) et sur la carte des formes contenues dans l'image (approche ascendante). Nous proposons ensuite d'utiliser simultanément les informations fournies par ces deux représentations pour obtenir une première segmentation de l'image. Nous résolvons ainsi bon nombre des difficultés mentionnées dans le chapitre précédent.

3.1. Carte des formes

Les composantes connexes fournissent une information pertinente au niveau des formes (graphiques et lettres d'un texte). Elles correspondent à un ou plusieurs caractères, à du bruit ou à une partie de graphique. Leur position, leur taille et le chevauchement de leurs rectangles circonscrits fournissent des informations précises sur la structure des pages. Elles correspondent à une information locale sur chaque forme contenue dans la page. Les composantes connexes représentant les lettres d'un mot sont extrêmement proches. Malheureusement, dans les documents anciens, la mise en forme n'est pas irréprochable. Par exemple, la dernière lettre d'une ligne peut être plus proche d'une note en marge que de la lettre qui la précède sur la ligne. Pour obtenir la carte des formes, après une binarisation de l'image, nous effectuons un suivi des contours des formes permettant d'extraire les rectangles circonscrits de chacune des composantes connexes. La position et la taille des rectangles sont stockées dans une liste constituant la carte des formes (figure 6a). En fonction de ses dimensions, l'un des labels suivants est affecté à chaque forme :

- *Bruit* (composantes connexes de petite taille)
- *Graphique* (composantes connexes de grande taille)
- *Texte* (autres composantes connexes de taille moyenne)

Les seuils utilisés durant cette phase sont choisis par l'utilisateur en fonction de la taille maximale et minimale des caractères

dans l'ouvrage. Il s'agit d'un premier étiquetage qui sera vérifié et qui évoluera par la suite.

3.2. Carte du fond (frontières entre les blocs)

L'image étant binarisée, le fond de la page est représenté par des pixels blancs. Normalement, un grand nombre de pixels blancs sont alignés verticalement (lgb_v) ou horizontalement ($l,\_h$) dans les zones séparant deux blocs (texte, graphiques, …). De même, un grand nombre de pixels blancs sont alignés horizontalement dans les zones séparant deux paragraphes. Par opposition, le nombre de pixels blancs que l'on est capable d'aligner dans un paragraphe entre deux lettres d'un mot ou entre deux mots d'une phrase ou entre deux lignes d'un même paragraphe est faible.
Par conséquent, nous proposons d'associer à chaque pixel de l'image, une valeur correspondant à la somme du nombre de pixels blancs successifs alignés horizontalement plus le nombre de pixels blancs successifs alignés verticalement $Ng(i,j) = Max - [lgb\_v(i,j) + lgb\_h(i,j)]$ pour construire une carte de distances (avec Max = hauteur × largeur de l'image). Il est possible de pondérer $lgb\_v(i,j)$ et $lgb\_h(i,j)$ relativement à la hauteur et à la largeur de l'image de façon à ne privilégier aucune des deux directions.
Nous affectons ainsi une valeur à chaque pixel appartenant au fond de la page à étudier (parties blanches de la carte des formes). Nous obtenons alors, après normalisation à 255 et comme le montre la figure 6b, une carte en niveaux de gris ($Ng(i,j)$). Baird a proposé une méthode se basant (uniquement) sur une information similaire à celle fournie par notre carte du fond puisqu'il propose d'extraire les rectangles blancs de taille maximum dans une page [Baird92]. La méthode de Baird est plus sensible aux problèmes d'inclinaison et de bruits que celle proposée ici puisque la présence d'une petite composante noire peut perturber significativement la construction des rectangles blancs. De plus, dans notre cas, les zones blanches ne constituent qu'une partie de l'information que nous exploitons pour réaliser la segmentation. Comme le montre la figure 6b, notre carte traduit les frontières (plus ou moins marquées) entre les blocs de la page.

3.3. Fusion des informations

3.3.1. Extraction des zones de texte

Pour utiliser simultanément les informations fournies par les cartes des formes et du fond, nous partons de la liste des composantes connexes étiquetées *Texte* pour reconstruire les paragraphes par association de composantes connexes susceptibles d'être des caractères. Pour qu'une association soit réalisée, il faut que les deux composantes soient assez proches et que le segment reliant leurs centres de gravité (G_1 et G_2) ne traverse pas une frontière trop marquée dans la carte du fond (niveaux de

Figure 6. Cartes du fond et des formes.

gris faibles). Cette contrainte multicritère sur laquelle repose notre méthode est traduite sous forme d'un produit et s'exprime donc par :

$$d(G_1,G_2) \times (256 - \underset{(i,j)\in[G_1,G_2]}{Min}[Ng(i,j)]) \leqslant Seuil \qquad (1)$$

où d désigne la distance euclidienne entre les centres de gravité des deux ensembles à associer. Lorsque ce critère est vérifié pour le plus proche voisin d'une composante, on leur attribue un même label (attribution d'un numéro identique à chaque caractère d'un même bloc de texte) sinon l'association est refusée. D'autres manières de définir la contrainte (1) devant être respecter ont été évaluées (somme des niveaux de gris plutôt que Min, distance aux bords des composantes plutôt qu'aux centres de gravités, …) pour ne retenir que celle donnant les meilleurs résultats.

La recherche des voisins se fait successivement horizontalement (fusion horizontale) puis verticalement (fusion verticale) de manière itérative et stoppe lorsqu'aucune fusion ne peut avoir lieu. Il est possible d'utiliser des seuils adaptés pour chaque type de fusion (horizontale et verticale). Les zones obtenues ne sont pas forcément rectangulaires puisqu'elles correspondent simplement à un rassemblement de formes portant un même label. Les zones construites peuvent donc avoir une forme quelconque. Deux exemples de résultat sont fournis figure 7. Le seuil utilisé est le même pour les deux images. Afin de visualiser la segmentation obtenue, chaque label regroupant les formes d'une zone donnée correspond à un niveau de gris différent. Dans l'exemple a) tous les blocs ont été correctement segmentés malgré leur proximité. Dans l'exemple b), la mise en page est beaucoup plus complexe (présence de lettrines, images, légendes, titres) et notre méthode a tendance à sur-segmenter l'image. Mais, comme nous allons le voir, cette sur-segmentation provoquée par le choix d'un seuil initial de fusion strict n'est que temporaire puisque l'algorithme de fusion peut être ré-appliqué plusieurs fois avec des paramètres différents (par insertion de la phase de fusion dans les scénarios d'analyse composés par les utilisateurs).

Figure 7. Exemples de segmentations en zones.

3.3.2. Séparation Texte/Graphique

Les parties graphiques ne correspondent pas toujours à une seule composante connexe. La plupart du temps, elles se traduisent par la superposition des rectangles circonscrits de plusieurs composantes connexes étiquetées *Texte* ou *Graphique*. Dans notre traitement, lorsque plusieurs rectangles circonscrits de composantes connexes étiquetées *Graphique* se chevauchent, ils sont fusionnés afin de produire un seul rectangle englobant l'ensemble de la zone graphique détectée. Il est également possible que des composantes connexes étiquetées Texte soient présentes à l'intérieur d'une zone étiquetée *Graphique*. Ces composantes sont alors ré-étiquetées *Texte_Graphique* et doivent faire l'objet d'une analyse particulière pour déterminer si elles appartiennent effectivement à la classe *Texte* ou à la classe *Graphique*. Pour cela, l'algorithme que nous avons mis au point utilise une nouvelle fois la notion de cartographie du fond pour effectuer la classification. Ce traitement est en fait réalisé juste avant la construction des zones de texte que nous avons étudiée dans le chapitre précédent. L'algorithme utilisé pour discerner les composantes *Texte* des composantes *Graphique* est le suivant :

1. ré-étiquetage des zones *Texte* situées à l'intérieur de zones *Graphique* en zones *Texte_Graphique*
2. coloriage des zones étiquetées *Texte* et *Texte_Graphique* en blanc dans l'image binarisée
3. calcul de la carte du fond (de la même manière que précédemment mais après suppression du texte)
4. les zones *Texte_Graphique* positionnées sur des parties foncées de la carte du fond sont conservées, les autres sont supprimées car elles sont considérées comme des morceaux de zones *Graphique*.

Les zones supprimées sont situées dans des régions fortement texturées de l'image binarisée traduisant la présence de parties graphiques dans l'image initiale (voir figure 8).

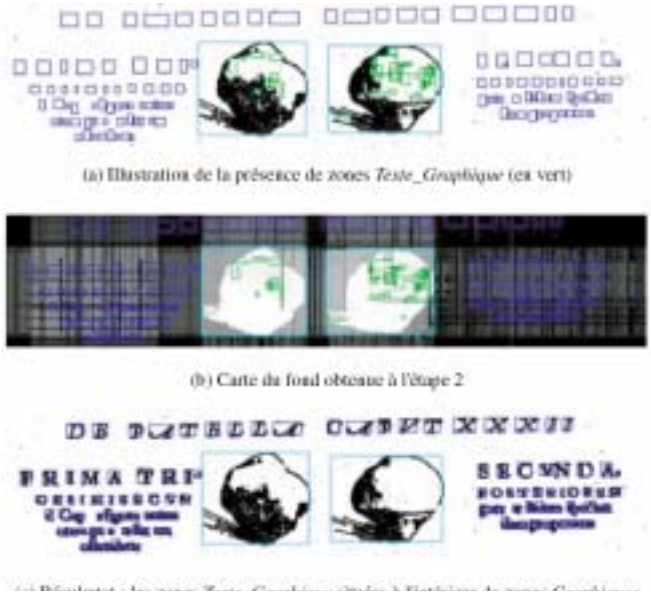

Figure 8. Illustration de la séparation Texte/graphique.

4. Analyse de structure interactive

Comme le montre la figure 7, l'étape de segmentation produit une représentation intermédiaire de l'image comportant un ensemble de zones étiquetées *Texte*, *Graphique*, *Texte_Graphique* ou *Bruit*. À ce stade du traitement, il est possible d'offrir à l'utilisateur la possibilité de construire lui-même des chaînes de traitements (que nous nommerons scénarios d'analyse structurelle) permettant de faire évoluer progressivement le contenu de cette représentation intermédiaire en fonction des caractéristiques des images à analyser et des objectifs visés. Le but est d'obtenir, à terme, un étiquetage le plus précis possible du contenu des pages numérisées d'un ouvrage de manière automatique par application d'un scénario construit de manière interactive. Notre objectif rejoint alors celui des chercheurs travaillant actuellement sur la mise en place de plateforme permettant la définition automatique de scénarios de traitement d'images [Saidali02].

4.1. Mise en place interactive des scénarios d'analyse

Une fois la segmentation initiale de l'image effectuée, l'architecture du système que nous proposons permet de poursuivre l'analyse de manière interactive. Pour cela, nous avons conçu un ensemble d'interfaces permettant à l'utilisateur de construire à sa guise des scénarios d'analyse structurelle provoquant l'évolution progressive du contenu de la représentation intermédiaire (segmentation initiale) préalablement obtenue.

Les outils mis à la disposition de l'utilisateur pour construire les scénarios sont :

- un éditeur de règles permettant de faire évoluer les étiquettes attribuées aux différentes zones (*Texte*, *Texte_Graphique*, *Graphique*, *Bruit*). Ces règles sont appliquées de manière séquentielle selon une stratégie définie par l'utilisateur qui peut, par exemple :
  - ne s'intéresser, dans un premier temps, qu'aux zones *Graphique* ou au contraire qu'aux zones *Texte*,
  - ou choisir d'étiqueter d'abord les zones facilement caractérisables à l'aide de règles simples pour ensuite se baser sur un contexte plus riche pour extraire des objets moins facilement identifiables [Ramel98].
- l'application des algorithmes de fusion horizontale ou verticale avec différents seuils et sur un type de zones défini par l'utilisateur (par exemple pour fusionner uniquement les zones étiquetées *Titre* et découpées au départ en plusieurs blocs comme dans l'exemple de la figure 7b, partie centrale)
- la suppression des zones d'un type particulier (par exemple les zones étiquetées *Texte* si l'utilisateur ne s'intéresse qu'aux lettrines)

Pour construire son scénario, l'utilisateur effectue (à la manière d'une macro) les actions successives devant être enregistrées sur une image type qu'il a sélectionnée ; ces actions sont appliquées sur l'image et il visualise le résultat de chacune d'elles en temps réel et peut ainsi valider leur intérêt. Les actions sont traduites sous forme littérale dans une liste visible en permanence par l'utilisateur (voir figure 9). L'utilisateur peut, ensuite, agir sur cette liste pour modifier son scénario en réordonnant ou supprimant certains traitements. Une fois le scénario considéré comme correct, celui-ci peut être enregistré dans un fichier pour archivage ou pour application sur un ensemble plus conséquent d'images (un ouvrage complet) lors d'un traitement par lot. Actuellement les règles utilisables pour faire évoluer les étiquettes des zones contenues dans la représentation intermédiaire concernent (figure 10) :

- la **position géographique** des zones
- les **relations de voisinage** entre zones identifiées
- la **forme et le contenu** des zones

4.2. Positionnement géographique

Il est possible d'identifier un certain nombre d'invariants sur le positionnement géographique des objets présents dans une image permettant de leur attribuer un label. Le but n'est pas d'utiliser des règles extrêmement contraignantes car la mise en page peut être variable (il est abusif de dire que le centre de gravité d'une note en marge gauche se trouve entre le pixel d'abscisse 205 et 213). Cependant, on peut dire raisonnablement que la position du centre de gravité d'une marge est très certainement située dans le premier tiers de la largeur de la page. Le système DMOS [Couasnon02] utilise une grammaire pour décrire la structure type des documents à analyser à l'aide de règles de voisinage entre objets comparables à celles que l'on propose d'utiliser ici. La technique est implémentée dans le logiciel FormuRead qui permet d'extraire automatiquement la structure de formulaires d'incorporation militaire du XIX^{e} siècle. Malgré leur dégradation, 60 000 formulaires de recrutement militaire du XIX^{ème} siècle ont été testés avec succès (99.6 %).

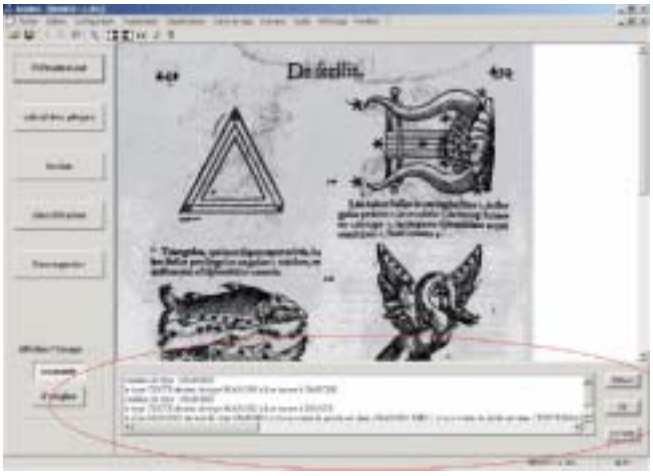

Figure 9. Vue du logiciel Agora avec un scénario.

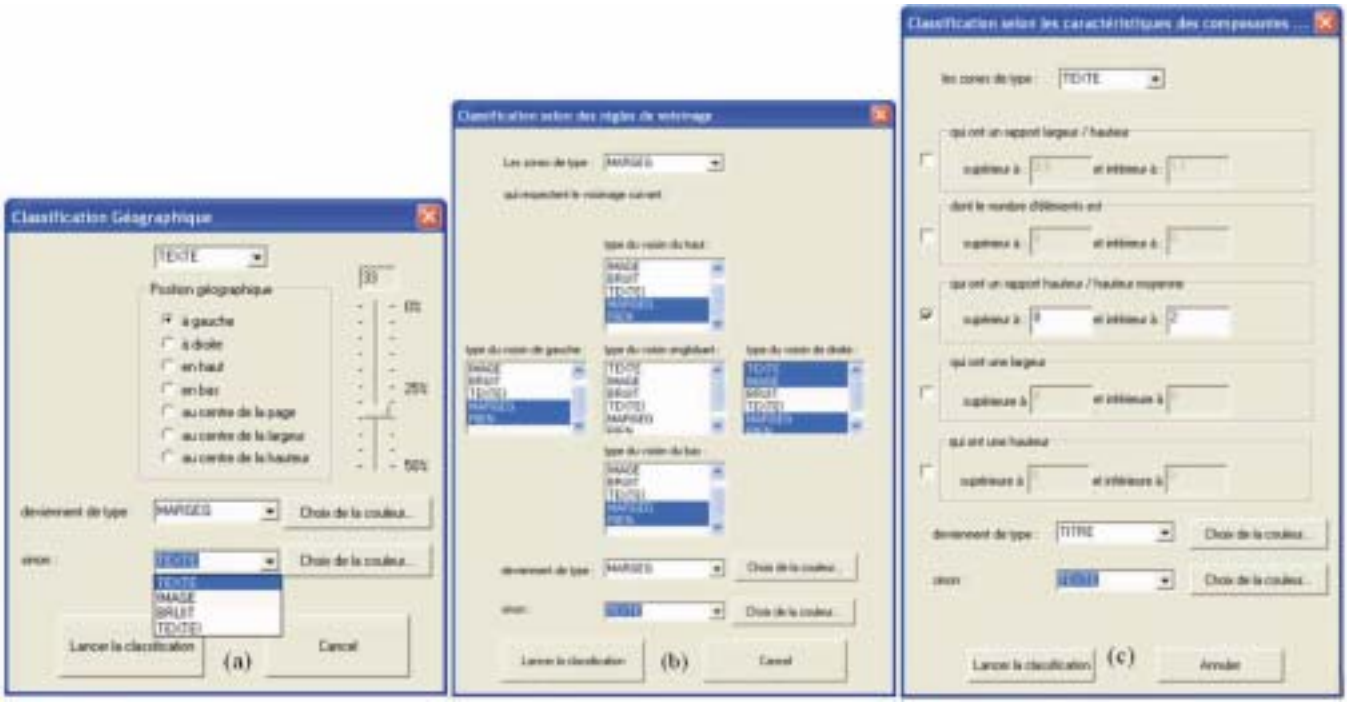

Figure 10. Interfaces pour la mise en place de règles sur la position (a), sur le voisinage (b), sur la forme (c).

Dans notre cas, l'interface proposée (voir figure 10a) permet de tenir compte de la position géographique (gauche, droite, haut, bas, centrage) du centre de gravité d'un type de zones pour faire évoluer son étiquette. Bien sûr ce type de règles ne constituera qu'un premier indice (étiquette provisoire) pour isoler correctement un type d'objet donné.

4.3. Étude du Voisinage

Il est également possible d'insérer dans un scénario d'analyse structurelle des règles concernant les relations de voisinage entre les zones pour attribuer une nouvelle étiquette à une zone (figure 10b). Des exemples de règles pouvant être mises en place pour extraire différents éléments sont fournis figure 11.

4.4. Forme et contenu

Une autre interface permet d'utiliser des critères sur la forme et le contenu d'une zone pour faire évoluer son étiquetage. La liste de ces critères est visible figure 10c. Le ratio Hauteur/Largeur permet, par exemple, de localiser les zones carrées (lettrines) ou les zones beaucoup plus larges que hautes (bandeaux). Le nombre d'éléments correspond au nombre de composantes connexes constituant la zone et peut s'avérer intéressant pour identifier certains objets. Le ratio Hauteur zone / Hauteur moyenne est très efficace pour localiser les zones textuelles ne comportant qu'une seule ligne de texte (titres, légendes, …).
Pour terminer cette partie, rappelons qu'en plus de ces règles permettant de faire évoluer les étiquettes de zones, il est possible d'insérer dans les scénarios des règles de fusion et de sup-

Figure 11. Exemples de règles de voisinage.

pression de zones. Des exemples de scénarios construits par des utilisateurs non spécialistes du traitement d'images sont visibles dans la partie suivante.

5. Expérimentations et résultats

Notre logiciel a été fourni au CESR qui l'utilise actuellement de manière intensive pour traiter, analyser, indexer et mettre en ligne son fonds ancien. Une formation à l'utilisation du logiciel (segmentation et création de scénarios) a été dispensée aux utilisateurs potentiels (chercheurs, historiens, informaticiens) afin qu'ils puissent produire des scénarios et, par conséquent, tester notre système d'analyse interactive d'images. Cette collaboration entre le CESR et le laboratoire a d'ores et déjà permis d'améliorer et de compléter les interfaces du logiciel pour mieux répondre aux besoins du terrain. Plusieurs expérimentations ont été menées, soit par le personnel du CESR, soit au sein de notre laboratoire et ont abouti aux résultats présentés dans ce chapitre.

5.1. Segmentation initiale

Nous appelons « segmentation initiale », le résultat obtenu après la première application de l'algorithme de fusion des composantes connexes *Texte* qui fournit en sortie un ensemble de blocs étiquetés *Texte, Graphique, Texte_Graphique* ou *Bruit*. Cette segmentation n'étant pas définitive (puisque évoluant durant les scénarios d'analyse incrémentale), nous n'évaluons dans cette partie que sa résistance à l'inclinaison des images. Les résultats obtenus après l'application complète d'un scénario attesteront, quant à eux, des performances globales de cette méthode de segmentation.

Nous avons testé la résistance de notre algorithme de segmentation aux défauts d'inclinaison sur des images de documents anciens et sur des images de périodiques actuels. Les résultats obtenus sont résumés dans les figures 12 et 13 et dénotent une tolérance suffisante aux défauts d'inclinaison pour les documents anciens et une tolérance forte pour les documents plus structurés. Si les documents ont été mal positionnés lors des prises de vue, il est préférable d'appliquer un prétraitement de correction des défauts géométriques avant d'utiliser l'algorithme de segmentation. Lorsque les prises de vue sont correctes, les résultats obtenus sont très satisfaisants.

5.2. Analyse structurelle par scénarios

Lors d'une première expérimentation menée au laboratoire, un scénario permettant l'étiquetage de sept types d'objets a été mis en place : classes « Marge Gauche (MG) », « Marge Droite (MD) », « Lettrine (Ltr) », « Numéro de Page Droite (NPD)», « Numéro de Page Gauche (NPG) », « Titre Principal (TP) » et « Titre (T) ». L'échantillon de test comportait 250 pages d'un ouvrage réputé complexe fourni par le CESR. L'exemple de page utilisé pour illustrer cet article a d'ailleurs été tiré de cet ouvrage. Le résultat obtenu est fourni figure 14. Le travail fastidieux de vérification de la classe affectée à chaque zone a été effectué de manière visuelle sur chaque page analysée. Le tableau 1 présente les résultats obtenus avec ce scénario lors du traitement par lot qui a duré 9h20 sur une machine Athlon barton 2.5 Ghz avec 768 Mo de RAM.

Les résultats finaux dépendent de la segmentation initiale et de la robustesse des règles constituant le scénario. L'objectif visé par l'utilisateur qui a conçu ce scénario était de n'avoir aucune mauvaise classification (prise de risque minimale).

Le scénario utilisé était donc réfléchi mais n'implémentait pas de « règles redondantes » qui auraient pu autoriser une prise de risque supplémentaire sur certains critères. En effet, les étiquettes générées à un instant donné peuvent être validées ensuite à l'aide d'autres règles. Avec ce scénario, 92 % des *Marges* sont extraites correctement sans aucune mauvaise détection. Ce scénario pourrait être complété afin de tenter de localiser à nou-

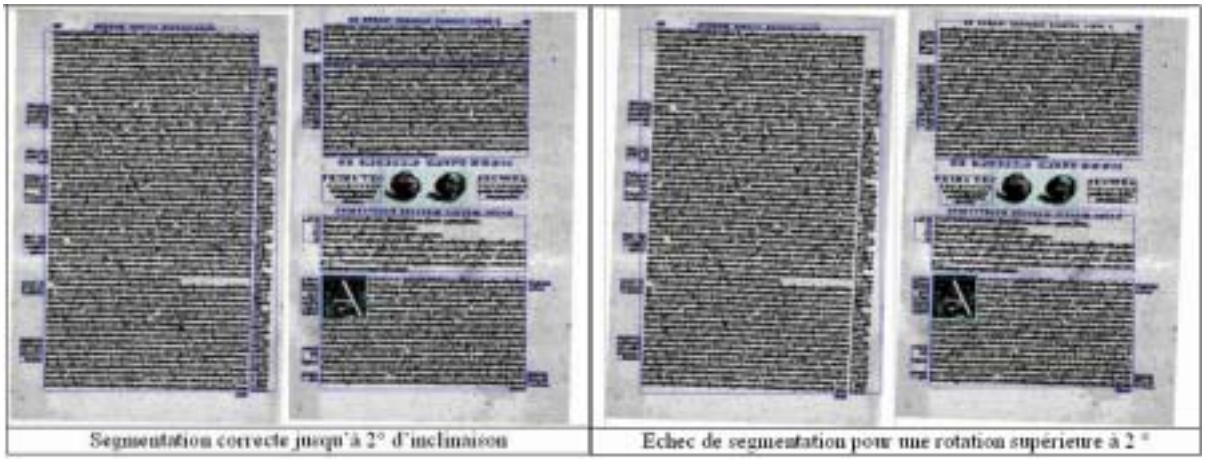

Figure 12. Sensibilité à la rotation sur des images de documents anciens.

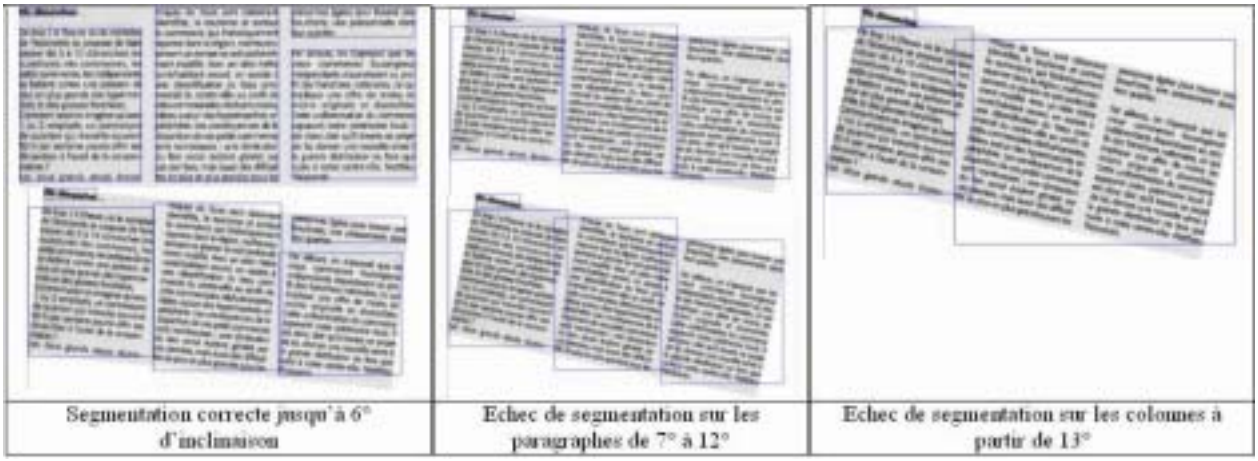

Figure 13. Sensibilité à la rotation sur des images de documents actuels.

Tableau 1. Résultats obtenus avec le scénario 1.

Type	Détectées	Fausses détections	Non détectées	Taux de détection (%)
Marges	207	18	0	92
Numéros de page droite	97	32	7	74
Numéros de page gauche	106	30	10	76
Titres principaux	223	27	0	90
Lettrines	80	1	0	99

veau les *Marges* manquantes une fois les autres objets de haut niveau localisés (évolution du contexte pour faciliter de nouvelles extractions). Toujours avec le même scénario, 99 % des *Lettrines* ont été bien classées ce qui est satisfaisant puisque l'étiquetage des *Lettrines* utilise le résultat de la classification des *Marges*.

De nombreuses autres expérimentations ont été réalisées au CESR de Tours par les personnes ayant suivi la formation à l'usage du logiciel. Outre les taux de reconnaissance obtenus, il est intéressant de remarquer la manière dont sont structurés et appliqués les scénarios d'analyse produits.

Au CESR, l'ensemble des tests a été effectué sur des fichiers issus d'une capture numérique avec l'appareil Nikon D1 au format TIF et en niveau de gris (taille image : 1 200 × 2 000 pixels). Les paramètres de segmentation utilisés sont les suivants :

- binarisation automatique
- hauteur minimale d'une grande composante : 60
- largeur minimale d'une grande composante : 60
- hauteur maximale d'une petite composante : 5
- largeur maximale d'une petite composante : 5
- Seuil de fusion horizontale : 500
- Seuil de fusion verticale : 500

Le premier résultat concerne la segmentation initiale qui ne produit aucune erreur d'étiquetage : Les blocs *Texte* et *Graphique* sont tous détectés correctement sur l'ensemble des ouvrages.

Un premier scénario a été exécuté sur 1 452 images provenant de cinq ouvrages différents. L'objectif de ce scénario (visible ci-dessous) était uniquement l'identification des différents types de zones graphiques présents :

- *Type marge Gauche : Texte à gauche à 15 %*
- *Type marge Droite : Texte à droite à 15 %*
- *Fusion verticale des marges à 100 000*

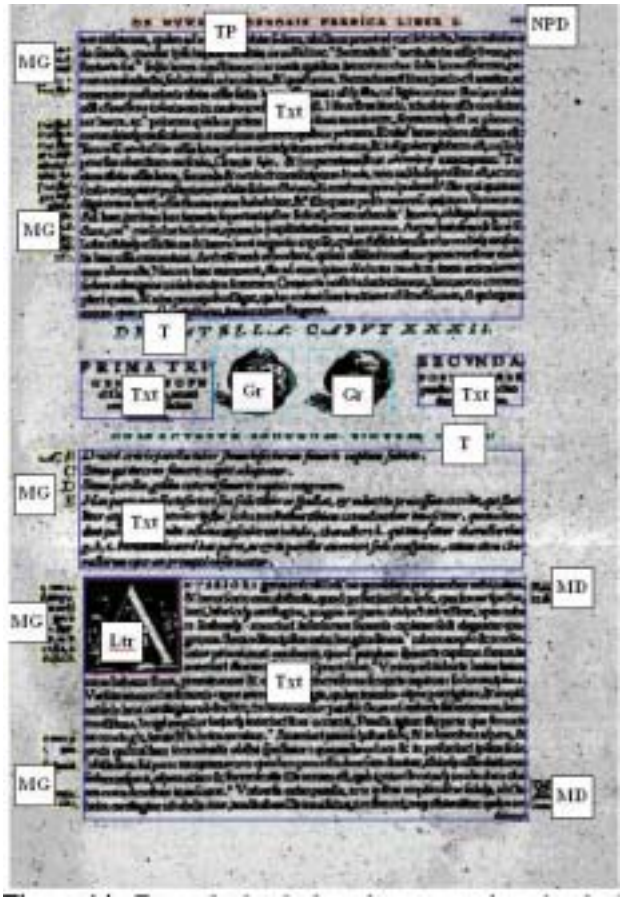

Figure 14. Exemple de résultat obtenu avec le scénario 1.

- Type Lettrine : image dont voisin de gauche Marge Gauche ou rien, voisin de droite texte, voisin du bas texte. + carrée : rapport largeur/hauteur 0,8<<1,2
- Type bandeau : image rapport largeur/hauteur 3<<10
- Type Portrait : image au centre de la page + si portrait en haut ou en bas à 20% redevient de type image
- Type Fleuron : image voisin du haut texte, voisin du bas rien + centrée sur la largeur et 50% en bas - Suppression du texte et des marges

Tableau 2 : Résultats obtenus.

Type	Détectées	Fausses détections	Non détectées	Taux de détection (%)
Bandeau	81	4	8	90.6
Fleuron	36	2	0	100
Lettrine	294	56	11	95.6
Portrait	89	31	0	100

Les images ont été traitées en 11h22 et les résultats obtenus sont fournis tableau 2.
Compte tenu de la diversité des ouvrages (taille, mise en page…) et de la simplicité du scénario, les résultats paraissent très corrects. Néanmoins, pour obtenir de meilleurs taux de détection et pour extraire des objets plus ambigus, il est souvent préférable d'adapter le scénario à chaque ouvrage. Ainsi, la suite des tests a été effectuée sur des images provenant d'un même ouvrage dont quelques pages sont présentées figure 15.
Un second test a été réalisé sur 180 pages de l'ouvrage sélectionné après prétraitement des images (correction d'éclairage 2D pour les tâches et l'ombre de la reliure, redressement des blocs et correction géométrique de la courbure) :
- Type bandeau : image rapport largeur/hauteur 3<<10
- Type Lettrine : image dont voisin de droite texte, voisin du bas texte ou rien. + carrée : rapport largeur/hauteur 0,7<<1,3
- Fusion verticale du texte à 2000
- Type marge Droite : Texte à droite à 25%
- Fusion verticale des marges à 100000
- Fusion horizontale du texte à 2000
- Type pagination : texte en haut à 10% + nombre d'éléments 0<<4
- Type signature : texte en bas à 25% + nombre d'éléments 0<<6
- Si type signature a du texte en dessous, à gauche ou à droite ou si il se trouve à gauche à 50%, il redevient de type texte
- Fusion verticale du texte à 3000
- Fusion horizontale du texte à 3000
- Type Titre : hauteur en pixel 0<<100 + texte en haut à 15% pour inclure le titre courant et la pagination
- Fusion horizontale du titre
- Suppression du type titre si le nombre d'éléments est inférieur à 8
- Suppression du type titre si le rapport largeur/hauteur est inférieur à 1
- Suppression du type titre s'il se trouve à gauche à 10%
- Suppression du type texte
Ce scénario a duré 2h32 et a produit les résultats suivants :
- Détection des Bandeaux à 100%
- Détection des Lettrines à 96% * Pagination détectée à 79%
- Identification de 100% des Titres mais avec 4% d'éléments parasites (tâches, bruits, …)
Un dernier test a été lancé sur les 1202 images de l'ouvrage non pré-traitées dans le but d'étiqueter les blocs *Texte* uniquement. Le scénario suivant a été utilisé :
- Suppression des images
- Fusion horizontale du texte à 3000
- Type marge gauche ou droite : texte à gauche ou à droite à 25%
- Fusion verticale des marges
- Si le type marge à un rapport largeur/hauteur 4<<30 redevient de type texte
- Type signature : texte en bas à 25% dont le nombre d'éléments 0<<8
- Fusion verticale du texte à 3000

Figure 15. Exemple de pages de l'ouvrage sélectionné.

- Type Titre : hauteur/hauteur moyenne en 0<<2 + texte en haut à 15 % pour inclure le titre courant et la pagination + texte dont le nombre d'éléments est inférieur à 50

- Fusion horizontale du type Titre

Après 10h52 de traitement, l'ensemble des *Titres* et des *Signatures* a été détecté correctement. L'extraction des *Marges* a fonctionné mais a produit beaucoup d'éléments parasites (ombres, tâches …)

Pour conclure cette partie expérimentation, on pourra noter que le temps moyen de traitement d'une image est de 30 secondes. Les résultats obtenus sont meilleurs sur des images pré-traitées. La binarisatrion automatique (par Sauvola [Sauvola00]) augmente le temps de traitement de 25 à 35 % ainsi que le taux de réussite en détection et en précision. Enfin, il est préférable de ne pas trop fusionner les blocs *Texte* au départ afin de pouvoir, dans certains cas, isoler de petits éléments (*Signature* ou *Pagination* par exemple) pour ensuite réaliser des fusions supplémentaires plus tard dans le scénario d'analyse.

6. Conclusion

Dans la première partie de cet article, nous avons mis en évidence les sources d'erreurs des méthodes traditionnelles de segmentation à l'aide d'une caractérisation des spécificités de mise en page dans les ouvrages anciens. Nous avons noté que chaque type de méthodes (ascendante et descendante) apportait des informations différentes qu'il ne faut pas ignorer pour atteindre une segmentation de qualité. Pour cela, notre proposition consiste à utiliser un algorithme hybride basé sur la construction de deux représentations de l'image : la carte des formes qui se focalise sur les composantes connexes et la carte des frontières qui fournit de l'information sur les espaces blancs séparant les blocs constituant la page. L'analyse conjointe du contenu de ces deux cartes permet d'aboutir à une segmentation initiale de l'image. Les résultats obtenus avec cette méthode sont très satisfaisants, le réglage des paramètres nécessaires est facile car peu sensible aux variations.

La seconde originalité de notre approche réside dans l'opportunité que nous offrons aux utilisateurs de pouvoir construire de manière interactive des scénarios d'analyse structurelle d'images. Partant d'une segmentation initiale, le but est de pouvoir faire évoluer cette représentation des images de manière progressive pour aboutir à une caractérisation la plus fine possible de son contenu en fonction des objectifs visés et du type d'images à analyser.

Le CESR a pu traiter plusieurs ouvrages complets à l'aide du prototype que nous lui avons fourni et a ainsi augmenté les possibilités offertes aux usagers de sa bibliothèque virtuelle. Même si le système a produit quelques erreurs, le gain de temps par rapport à un traitement manuel est considérable, cela tout en fournissant aux spécialistes de l'histoire du livre un outil qu'ils ne pensaient pas réalisable.

Références

[Akindele93] O.T. AKINDELE, A. BELAID. « Page Segmentation by Segment Tracing ». *In Proc. of the 2nd International Conference on Document Analysis and Recognition*, 1993. p341-344

[Baird92] H BAIRD. « Background structure in document images », In *Advances in Structural and Syntactical Pattern Recognition*, ed. H. Bunke. 1992. p253-269.

[Belaid97] A. BELAÏD, « Conception automatisée de modèles de page en vue de leur utilisation en reconnaissance de documents », Workshop on Electronic Page Models (LAMPE'97). 1997.

[Couasnon02] B. COÜASNON, J. CAMILLERAPP, « DMOS, une méthode générique de reconnaissance de documents : évaluation sur 60 000 formulaires du XIX^{e} siècle », *in Actes du Colloque International Francophone sur l'Écrit et le Document (CIFED'02)*, Hammamet, 2002, p. 225-234.

[Hadjar01] K. HADJAR, O. HITZ, R. INGOLD, « Newpaper page decomposition using split and merge approach », *Proceedings of the 5t^{h} International Conference on Document Analysis and Recognition.* 2001. p. 1186-1191.

[Hadjar02] K. HADJAR, O. HITZ, L. ROBADEY, R. INGOLD, « Configuration REcognition Model for Complex Reverse Engineering Methods: 2(CREM) », *Proceedings of the 5th International Workshop on Document Analysis Systems*, 2002. p. 469-479.

[Kise98] K. KISE, AKINORI SATO, MOTOI IWATA. « Segmentation of page images using the area Voronoi diagram », *Computer Vision and Image Understanding archive*, Volume 70 (3), Special issue on document image understanding and retrieval, 1998, p. 370-382.

[Lebourgeois03] F. LEBOURGEOIS, H. EMPTOZ, E. TRINH, « Compression et accessibilité aux images de documents numérisés - Application au projet Debora », *Document Numérique*, Vol 7 n°3-4. 2003, p. 103-127.

[Nagy84] G. NAGY, S. SETH, « Hierarchical representation of optically scanned documents », *In 7th International Conference on Pattern Recognition (ICPR)*, 1984, p. 347-349.

[Nagy93] G. NAGY, S. SETH, M. KRISHNAMOORTHY, AND M. VISWANATHAN, « Syntactic Segmentation and Labeling of Digitized Pages from Technical Journals », *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(7), 1993. p737-747.

[OGorman93] L. O'GORMAN, « The Document Spectrum for Page Layout Analysis», *In IEEE Transaction On PAMI*,15 (11). 1993. p. 1162-1173.

[OGorman95] L. O'GORMAN, R. KASTURI, « Document Image Analysis », *IEEE Computer Society Press*, Los Alamitos, CA, 1995.

[Ramel98] J.Y. RAMEL, N. VINCENT, H. EMPTOZ, « Extraction contextuelle d'entités graphiques dans les dessins : du plus simple au plus complexe…. » *Colloque International Francophone sur l'Ecrit et le Document*, Quebec (Canada). May 1998. p. 453-462.

[Saidali02] Y. SAIDALI, N. BAUDOUIN, E. TRUPIN, M. HOLZEM, J. LABICHE, « ACTI_VA : Plate-Forme interactive pour l'Acquisition de Connaissances Traiteur d'Images de document », *5^{e} Colloque International sur le Document Électronique* CIDE, Hammamet, Tunisie, 20-23 Octobre 2002, p. 195-209.

[Sauvola00] J. SAUVOLA, M. PIETIKAINEN,: « Adaptive Document Image Binarization », *Pattern Recognition*, Vol. 33, p. 225-236, 2000

[Trinh03] E TRINH, « De la numérisation à la consultation de documents anciens ». Thèse de doctorat en Informatique, Insa de Lyon, Juin 2003

Jean-Yves **Ramel**

Jean-Yves Ramel a obtenu sa thèse de doctorat en Informatique en 1996. Maître de conférences à l'INSA de Lyon puis maintenant à l'Ecole Polytechnique de l'Université François Rabelais de Tours, ses activités de recherche actuelles concernent principalement le domaine de l'écrit et du document. Ces travaux portent plus particulièrement sur les architectures et stratégies d'analyse d'images permettant la mise en place d'algorithmes coopératifs d'extraction d'information dans les documents anciens et graphiques.

Stéphane **Leriche**

Stéphane Leriche est ingénieur en Informatique diplômé de l'école Polytechnique de l'Université de Tours. Il a obtenu son Master de recherche en Informatique en 2004. Durant le cadre de son stage au sein du Laboratoire d'Informatique de Tours, il a travaillé sur l'analyse de la structure des documents anciens imprimés. Il est actuellement ingénieur application au sein de la société SERLI Informatique à Poitiers.