Approche perceptive pour la reconnaissance de filets bruités, Application à la structuration de pages de journaux

Aurélie Lemaitre, Bertrand B. Coüasnon, Jean Camillerapp

Approche perceptive pour la reconnaissance de filets bruités Application à la structuration de pages de journaux

Aurélie Lemaitre – Jean Camillerapp – Bertrand Coüasnon

IRISA/INSA
Campus de Beaulieu - 35042 Rennes Cedex

aurelie.lemaitre@irisa.fr

Résumé : *Dans le domaine de la reconnaissance de documents, les filets peuvent servir de base à l'extraction de la structure. Cependant, dans le cas de documents anciens ou bruités, ces filets sont plus difficile à détecter. En effet, ces documents peuvent être mal imprimés, dégradés par de mauvaises conditions de conservation ou déformés lors de la numérisation. Les méthodes trouvées dans la littérature se basent sur une forte connaissance à priori de la longueur et de l'épaisseur des lignes pour pouvoir regrouper les pixels composant une même ligne.*

Nous proposons une nouvelle approche basée sur un mécanisme utilisé par l'oeil humain : la vision perceptive. En effet, la combinaison des visions qu'on peut avoir d'une même image à plusieurs résolutions permet de construire des lignes, sans avoir de connaissance à priori spécifique sur leur nature. Afin de valider notre méthode, nous l'appliquons à l'analyse de la structure de pages de journaux et montrons qu'elle permet d'améliorer les résultats.

Mots-clés : Vision perceptive, filets, multirésolution, structure

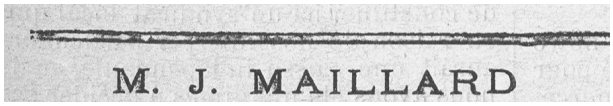
1 Introduction

Les filets sont des éléments structurels qui peuvent servir de base pour la reconnaissance de documents fortement structurés, tels que les formulaires, les tableaux ou les pages de journaux. Ainsi, dans le concours de reconnaissance de journaux proposé lors d'ICDAR'01 [GAT 01], deux des trois méthodes basent leur approche sur la reconnaissance des lignes horizontales et verticales.

Cependant, dans le cas des documents anciens ou abîmés, la détection des filets est plus complexe. En effet, de mauvaises techniques d'impression peuvent produire des lignes avec des bavures ou partiellement effacées ; de mauvaises conditions de conservation du document font apparaître des tâches liées à des pliures du papier ou des déchirures ; enfin, l'étape de numérisation introduit parfois du biais ou de la courbure dans le document. Des exemples de cas difficiles sont présentés figure 1.

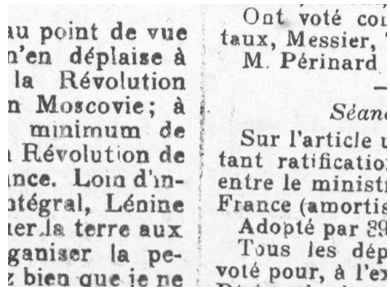
Dans ces conditions, l'extraction de filets est particulièrement compliquée. Les méthodes classiques basées sur la projection ou la transformée de Hough ne sont pas très appropriées. En effet, elles utilisent uniquement une analyse globale de l'image, et sont très sensibles au biais et à la courbure. Par conséquent, d'autres méthodes ont été propo-

(a) Ligne épaisse mouchetée de blanc

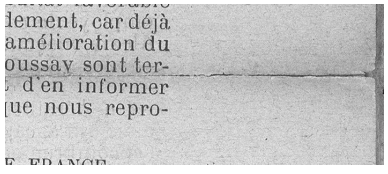


(b) Lignes doubles qui se recouvrent

(c) Ligne discontinue



(d) Ligne fine légèrement effacée



(e) Ligne due à une déchirure du papier

FIG. 1 – Exemple de lignes difficiles à détecter dans des documents anciens

sées. Ainsi, Gatos *et al.* proposent dans [GAT 99] une méthode spécifique basée sur la transformation en niveaux de gris d'une image binaire. Cependant, cette méthode requiert des informations à priori sur la longueur et l'épaisseur des lignes contenues dans le document. Hajdar *et al.* [HAD 01] détectent des lignes discontinues en se basant sur un regroupement de composantes connexes. Là encore, ils ont besoin de connaître une distance maximale entre deux composantes appartenant à une même ligne. Les travaux de Liu *et al.* [LIU 01] sont aussi basés sur un seuil de distance. Dans ces méthodes, les auteurs utilisent une forte connaissance à priori sur l'épaisseur et la longueur des lignes étudiées.

Nous proposons une approche basée sur un mécanisme utilisé par l'oeil humain : la vision perceptive. En effet, quand nous regardons un document, notre cerveau est capable de combiner une analyse à différent niveaux de vision (de près, de loin) pour produire une interprétation du document. Les approches multirésolutions ont déjà été utilisées en reconnaissance de structure [DÉF 95] [EGL 06], mais sont souvent dédiées à un type de documents donné. Dans le cas de la reconnaissance de filets, nous allons montrer qu'utiliser une telle stratégie permet de s'affranchir de connaissances à priori spécifiques à la nature à la fois des lignes mais aussi des documents étudiés.

Dans ce contexte, Xi *et al.* [XI 05] proposent une approche basée sur des ondelettes qui utilise une vision globale du document pour obtenir un positionnement approximatif des lignes, puis une vision locale pour en préciser le contenu. Cependant, cette méthode ne permet pas de gérer des traits courbes ou un biais de plus de 2°, cas fréquents dans des documents d'archives.

Nous présentons donc une nouvelle approche basée sur la vision perceptive, applicable sans connaissance a priori sur le type de documents ou sur la nature des filets, capable de gérer les défauts des documents d'archives que sont le bruit, le biais et la courbure.

Dans une première partie, nous présentons notre stratégie de reconnaissances des lignes basée sur la vision perceptive. Puis, dans la section 3, nous décrivons l'implémentation de cette stratégie au sein d'une méthode générique. Enfin, nous validons notre approche en l'appliquant à la structuration de pages de journaux.

2 Mécanisme perceptif

Pour simuler la vision à différentes distances de l'image, nous utilisons une pyramide multirésolution construite par filtre passe bas à partir de l'image. Expérimentalement, nous avons constaté que l'utilisation de trois niveaux de résolution est satisfaisante. Dans ce cas, en effet, les différences de perception entre résolutions sont significatives, sans être trop importantes. L'analyse est donc basée sur les trois niveaux suivants :

- La résolution *haute* représente la vision de près. C'est l'image initiale (environ 300 dpi).
- La résolution *moyenne* représente une vision intermédiaire. Elle est construite à partir de l'image initiale dont on a divisé les dimensions par 4 (environ 75 dpi).
- La résolution *basse* correspond à la vision de loin. Elle est construite à partir de l'image initiale dont on a divisé les dimensions par 16 (environ 20 dpi).

A chacune des ces résolutions, la perception des segments est différente (figure 2). C'est cette variation de perception qui va permettre d'élaborer notre mécanisme de reconnaissance des lignes.

2.1 Segments perçus aux différentes résolutions

Vision globale En regardant un document à basse résolution (figure 2(b)), les éléments perçus comme des segments sont :
- les lignes épaisses, même si elles sont dégradées (figure 1(a)),
- les lignes multiples (figure 1(b)) qui apparaissent comme un seul segment,
- certaines lignes de texte en caractères gras, ou plus foncées (segment 25 en bas à droite de la figure 2(b)).

Les filets fins sont trop clairs pour être perçus à ce niveau de résolution.

Vision intermédiaire En regardant ce document à résolution moyenne (figure 2(c)), les éléments perçus comme des segments sont :
- des morceaux de filets multiples (figure 1(b)),
- des morceaux de filets fins (figures 1(c) et 1(d)),
- des morceaux de filets épais (figure 1(a)),
- des parties rectilignes de lettres majuscules.

Vision locale En regardant ce document à résolution haute (figure 2(d)), on peut percevoir les segments suivants :
- des morceaux de filets multiples (figure 1(b)),
- des morceaux de filets simples (figures 1(c) et 1(d)),
- des éléments causé par du bruit (figure 1(e)).

Les segments épais peuvent ne pas être perçus si le bruit est trop important, par exemple dans le cas de mouchetage blanc (figure 1(a)).

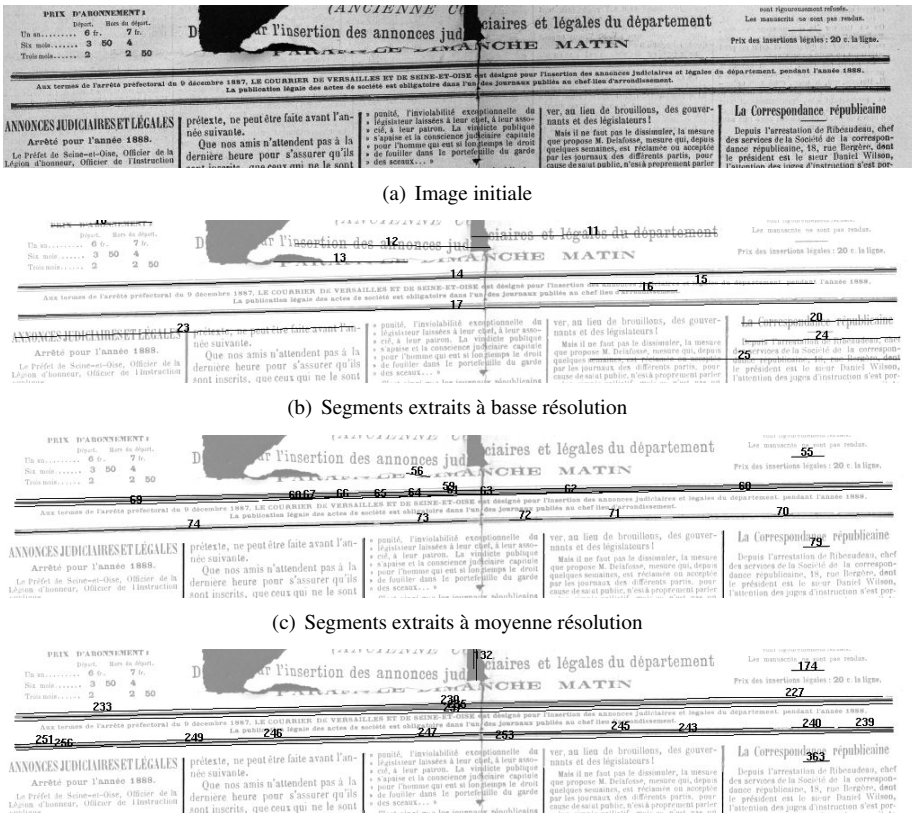
Ces constatations sont regroupées dans le tableau 1. On distingue deux groupes de lignes : les « vrais » filets qui sont des éléments structurels utiles, c'est à dire les lignes épaisses, fines et multiples, et les « fausses » lignes qui sont du bruit dans l'analyse (lignes de texte, caractères, bruits liés à la mauvaise qualité du document).

Résolution	Basse	Moyenne	Haute
Lignes fines	Non	Oui	Oui
Lignes épaisses	Oui	Oui	Non
Ligne multiples	Oui	Oui, 2 lignes	Oui, 2 lignes
Lignes de texte	Oui	Non	Non
Lettres	Non	Oui	Non
Bruit	Non	Oui	Non

TAB. 1 – Pour chaque type de ligne, résolution à laquelle elles sont généralement visibles en tant que segments

2.2 Stratégie d'analyse

Le tableau 1 montre que la perception qu'on peut avoir d'un élément aux diverses résolution permet de déterminer le type de ligne étudiée. Nous en déduisons donc une stratégie d'analyse qui combine les visions aux différentes résolutions pour construire une ligne résultat. Le principe de base est



(a) Image initiale

(b) Segments extraits à basse résolution

(c) Segments extraits à moyenne résolution

(d) Segments extraits à résolution haute

FIG. 2 – Segments extraits aux différents niveaux de résolution (reportés dans un même référentiel pour plus de lisibilité)

que la vision à la résolution inférieure permet d'émettre une hypothèse sur la présence d'une ligne. Cette hypothèse peut être confirmée par la présence de segments aux résolutions supérieures. Cette stratégie est décrite sur la figure 3.

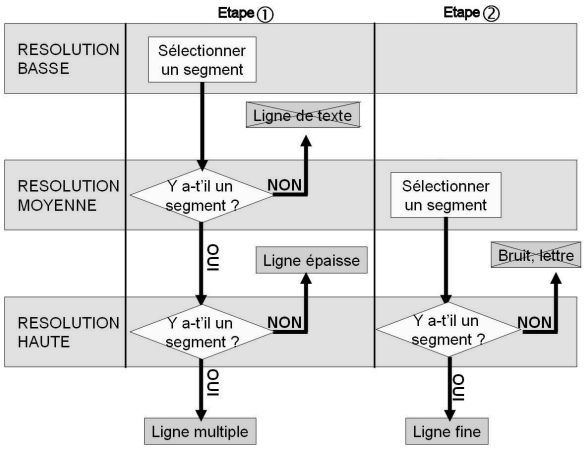


FIG. 3 – Stratégie de combinaison des segments détectés aux trois résolutions

L'analyse est réalisée en deux étapes. On recherche d'abord les segments à basse résolution et leur correspondance à moyenne et haute résolution. Puis, on s'intéresse aux segments restants à moyenne résolution et à leur correspondance à haute résolution.

La première étape consiste à sélectionner un segment à basse résolution. A partir de ce segment, on teste la présence de segments associés à moyenne résolution. Si aucun segment n'est trouvé, il doit s'agir d'une ligne de texte. Dans le cas contraire, la présence à haute résolution de segments associés permet de construire une ligne multiple ou épaisse. Le même mécanisme est proposé pour reconnaître les lignes fines, perceptibles aux résolutions moyenne et haute.

Il est important de noter que c'est la position du segment vu à basse résolution qui va déterminer la zone de recherche pour les résolutions supérieures. De plus, la vision à faible résolution fournit des connaissances sur la nature de la ligne : biais, courbure, épaisseur, longueur, qui vont servir à définir le contexte d'agglutination pour la construction de la ligne finale.

3 Implémentation dans une méthode générique

Pour implémenter la stratégie proposée ci-dessus, il nous semble nécessaire de disposer des outils suivants :

- un extracteur de segments,
- un mécanisme permettant de naviguer entres les différentes résolutions de l'image,
- un mécanisme capable de mettre en correspondance les éléments extraits aux différentes résolutions.

Tous ces critères sont disponibles dans la méthode DMOS développée par Coüasnon, que nous avons enrichie par de nouveaux opérateurs de multirésolution.

3.1 La méthode DMOS

La méthode DMOS [COÜ 01] (Description et MOdification de la Segmentation) est une méthode générique pour la reconnaissance de documents structurés. Cette méthode est basée sur le langage grammatical EPF (Enhanced Position Formalism), qui permet d'effectuer une description bidimensionnelle des éléments contenus dans un document. Une fois cette description réalisée pour un type de document donné, l'analyseur associé est produit automatiquement par compilation. Grâce à ce principe, la connaissance liée à un type de documents est complètement externalisée de la méthode, ce qui assure son caractère générique. Cette méthode a été appliquée pour de nombreux types de documents [COÜ 01] [LEM 07] et validée à grande échelle (plus de 500 000 pages traitées).

Travailler dans le contexte de cette méthode nous permet d'avoir un langage de description (EPF) pour exprimer la stratégie décrite précédemment.

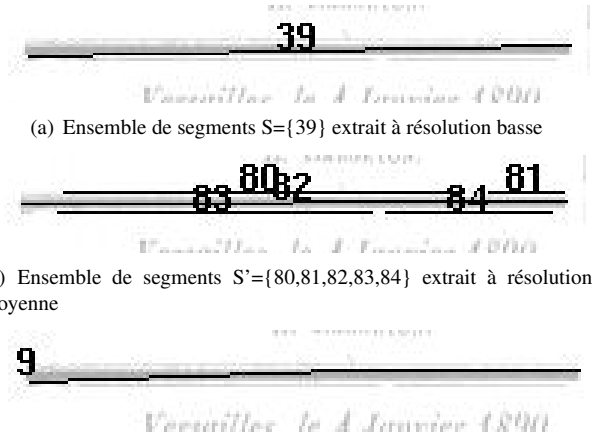
3.2 Les outils multirésolution

Nous avons présenté dans [LEM 07] une première introduction des outils de multirésolution dans DMOS. Ainsi, il est possible d'analyser un document à partir de plusieurs résolutions, et des ensembles de segments extraits à chacune de ces résolutions. Ces segments sont extraits grâce à un détecteur de segment basé sur un filtrage de Kalman [LEP 95]. Cet extracteur est particulièrement adapté aux segments bruités, en biais ou incurvés.

La capacité de changer de résolution en cours d'analyse est offerte par un nouvel opérateur du langage EPF, qui a été créé spécifiquement (voir [LEM 07]).

Nous introduisons un nouvel outil pour mettre en correspondance des éléments issus de plusieurs résolutions (figure 4). Nous proposons le concept de *ligne abstraite* et l'opérateur de *recalage*. Une *ligne abstraite* L est construite à partir d'un ensemble de segments S, extrait dans une résolution A (figure 4(a)). La ligne L possède alors des caractéristiques de position, d'épaisseur, de biais et de courbure qui permettent d'émettre l'hypothèse d'une zone de recherche dans la résolution B (figure 4(b)). Les segments S' trouvés à cette résolution permettent d'ajuster L (figure 4(c)), dans le but de compenser les bruits de quantification liés au changement de résolution (A vers B). Ceci est réalisé par l'outil de *recalage*.

Les *lignes abstraites* ainsi produites par le système peuvent alors servir de base pour la description d'un type de document plus complexe. Dans ce contexte plus large d'un document complet, le concept de *ligne abstraite* permet de masquer l'analyse multirésolution utilisée pour la détection des lignes, et donc de simplifier la description. Ainsi, l'utilisateur de DMOS et du langage EPF n'a pas besoin de tenir compte des variations d'apparence des lignes. Ceci assure la généricité de notre approche qui pourra être utilisée pour des documents de types variés.



(a) Ensemble de segments S={39} extrait à résolution basse

(b) Ensemble de segments S'={80,81,82,83,84} extrait à résolution moyenne

(c) Ligne abstraite L=9 construite à partir de S puis ajustée par recalage en fonction de S'

FIG. 4 – Exemple de recalage d'une ligne abstraite (reporté dans un même référentiel pour plus de lisibilité)

4 Application aux journaux anciens

Afin de valider notre méthode basée sur la vision perceptive, nous présentons un exemple d'utilisation des *lignes abstraites* pour la description de la structure de pages de journaux.

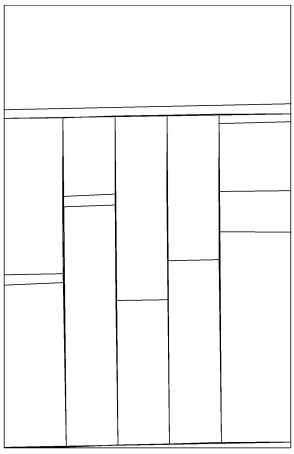
4.1 Principe de validation

Le but de cette application est de montrer l'apport de la vision perceptive pour la reconnaissance des lignes. Nous proposons donc d'extraire dans des pages de journaux les « cases » qui sont séparées par des filets horizontaux et verticaux. Nous mettons en place une grammaire de description des pages de journaux dans le langage EPF. Cette description consiste en un découpage récursif de la page, selon les *lignes abstraites* trouvées. Un exemple de segmentation produite est présentée sur la figure 5. Cette grammaire fonctionne indépendamment de la méthode utilisée pour détecter les lignes.

Afin de valider nos travaux, nous comparons deux méthodes pour la détection de ces lignes : une approche multirésolution qui est l'implémentation de notre stratégie perceptive, et une approche monorésolution qui se contente d'extraire les segments dans l'image à résolution initiale.

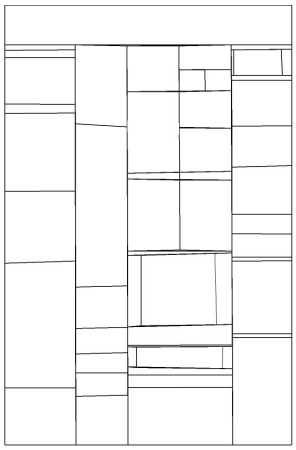
Notre base de validation est constituée de pages de journaux de 1859 à 1944, issues de 4 périodiques différents provenant des Archives Départementales des Yvelines : le « Journal de Mantes », « La Concorde de Seine et Oise », « Le Courrier de Versailles et de Seine et Oise » et « Le Progrès ». Nous avons constaté que certaines pages de journaux présentaient plus de traits difficiles à reconnaître que d'autres. Ainsi, dans les premières pages de journaux (figure 5(a)), l'épaisseur des filets est relativement constante, alors que les dernières pages (figure 5(c)) contiennent davantage d'encarts publicitaires avec des lignes d'épaisseurs variable. Nous avons donc créé deux bases : une base constituée de 179 premières pages de journaux, dans laquelle nous avons établi manuellement une vérité terrain constituée de 4148 cases, et une base contenant 79 dernières pages de journaux et 3480 cases.

(a) Exemple de première page de journal



(b) Cases extraites dans la première page

(c) Exemple de dernière page de journal



(d) Cases extraites dans la dernière page

FIG. 5 – Segmentation de pages de journaux avec notre approche perceptive

4.2 Métrique utilisée

La comparaison entre les résultats et la vérité terrain est en fait un problème de sur et de sous-segmentation. Nous utilisons donc la métrique proposée par Silva [SIL 07], basée sur les notions de complétude et de pureté, qui semble bien adaptée à ce problème de segmentation.

Lorsqu'une case présente dans le vérité terrain est trop segmentée par la méthode (sur-segmentation), elle est *incomplète*. L'*incomplétude* est la proportion de cases attendues trouvées de manière incomplète par rapport au nombre total de cases attendues.

Lorsque deux cases de la vérité terrain sont regroupées en une seule dans le résultat produit (sous-segmentation), cette case reconnue est dite *impure*. L'*impureté* correspond au taux de cases impures reconnues, par rapport au nombre total de cases reconnues.

Les taux d'*incomplétude* et d'*impureté* doivent être les plus petits possible.

4.3 Résultats

Les résultats obtenus sur la base des premières pages sont présentés dans le tableau 2, et ceux sur la base des dernières pages dans le tableau 3.

Les apports de la multirésolution sont plus marquants sur la base de dernières pages. En effet, c'est sur ces pages que les filets sont particulièrement difficiles à reconnaître. Sur cette base, la version multirésolution, basée sur notre approche perceptive, diminue l'impureté (sous-segmentation) de 45%, tout en diminuant l'incomplétude (sur-segmentation) de 20%.

Version	Cases	Incomplétude	Impureté
Monoresolution	4148	10.46%	10.73%
Multiresolution	4148	10.17%	7.87%
Gain		- 3%	- 33%

TAB. 2 – Application de l'extraction de filets pour le découpage de 179 premières pages de journaux en cases (filets de bonne qualité)

Version	Cases	Incomplétude	Impureté
Monoresolution	3480	17.06%	11.34%
Multiresolution	3480	13.70%	6.23 %
Gain		- 20%	- 45%

TAB. 3 – Application de l'extraction de filets pour le découpage de 79 dernières pages de journaux en cases (filets variés et dégradés)

Un des gains principaux dans l'approche multirésolution concerne la vision des filets mouchetés comme celui présenté sur la figure 1(a). En effet, ce type de filet n'est pas visible à résolution haute, et donc non détecté dans la version monorésolution. Au contraire, si on avait basé la version monorésolution uniquement sur la résolution basse, ces filets mouchetés auraient été visibles, mais les filets fins n'auraient pas été détectés. Notre approche permet donc de s'affranchir de ces problèmes.

Ces résultats sont à utiliser uniquement dans un but de comparaison de deux méthodes de détection de lignes, et non pour évaluer les performances d'une segmentation de pages de journaux. En effet, on pourrait améliorer la grammaire de description des pages de journaux, en tenant compte de certaines spécificités liées à la presse, si on voulait obtenir de meilleurs résultats quant à la segmentation en cases.

5 Conclusion

Nous présentons dans cet article une nouvelle approche pour la reconnaissance des lignes dans des documents anciens, abîmés ou mal imprimés. Cette méthode, basée sur le principe de la vision perceptive, combine des visions de l'image à des résolutions différentes. L'étude de l'image à la résolution inférieure permet d'émettre une hypothèse sur l'existence et la nature d'une ligne (position, épaisseur, courbure). Ces caractéristiques permettent de guider l'analyse à une résolution supérieure, et la présence de segments dans cette résolution permet de confirmer l'hypothèse de la ligne. Les segments issus des différentes résolutions sont alors combinés pour former une *ligne abstraite*.

Cette ligne abstraite peut alors servir de base pour la description d'un type de documents plus complexe. C'est la cas de l'application présentée pour le découpage de la structure de pages de journaux. Dans le cadre de cette application, nous montrons que notre méthode permet de diminuer l'impureté de 45% et l'incomplétude de 20% pour les pages ayant des filets dégradés.

Grâce au caractère générique de la méthode DMOS, qui sépare la connaissance du système, notre approche peut être appliquée pour des types variés de documents, d'autant plus qu'aucune connaissance à priori n'est requise sur l'épaisseur ou la longueur du segment, contrairement aux approches trouvées dans la littérature.

Références

[COÜ 01] COÜASNON B., DMOS : A generic document recognition method to application to an automatic generator of musical scores, mathematical formulae and table structures recognition systems, *International Conference on Document Analysis (ICDAR'01)*, 2001, pp. 215-220.

[DÉF 95] DÉFORGES O., PIQUIN P., VIARD-GAUDIN C., , BARBA D., Segmentation d'images de documents par une approche multirésolution. Extraction précise de lignes de texte, *Traitement du signal*, vol. 12-6, 1995, pp. 527-539.

[EGL 06] EGLIN V., Approches perceptives et cognitives en analyse automatique d'images de documents, *Revue TSI technique et Sciences Informatiques, numéro spécial "Document Numérique"*, vol. 25/4, 2006, pp. 523-551.

[GAT 99] GATOS B., MANTZARIS S. L., CHANDRINOS K. V., TSIGRIS A., , PERANTONIS S. J., Integrated Algorithms for Newspaper Page Decomposition and Article Tracking, *International Conference on Document Analysis (ICDAR'99)*, 1999, page 559.

[GAT 01] GATOS B., MANTZARIS S., , ANTONACOPOULOS A., First International Newspaper Segmentation Contest, *International Conference on Document Analysis (ICDAR'01)*, 2001, page 1190.

[HAD 01] HADJAR K., HITZ O., , INGOLD R., Newspaper Page Decomposition Using a Split and Merge Approach, *International Conference on Document Analysis (ICDAR'01)*, 2001, page 1186.

[LEM 07] LEMAITRE A., CAMILLERAPP J., , COÜASNON B., Contribution of Multiresolution Description for Archive Document Structure Recognition, *International Conference on Document Analysis (ICDAR'07)*, 2007, pp. 247-251.

[LEP 95] LEPLUMEY I., CAMILLERAPP J., , QUEGUINER C., Kalman filter contributions towards document segmentation, *International Conference on Document Analysis (ICDAR'95)*, 1995, pp. 765–769.

[LIU 01] LIU F., LUO Y., HU D., , YOSHIKAWA M., A New Component Based Algorithm for Newspaper Layout Analysis, *International Conference on Document Analysis (ICDAR'01)*, 2001, page 1176.

[SIL 07] SILVA A. C. E., New Metrics for Evaluating Performance in Document Analysis Tasks - Application to the Table Case, *International Conference on Document Analysis (ICDAR'07)*, 2007, pp. 481-485.

[XI 05] XI D., LEE S. W., Extraction of reference lines and items from form document images with complicated background, *Pattern Recognition*, vol. 38, n° 2, 2005, pp. 289–305.