Analyse d'images de documents anciens : Catégorisation de contenus par approche texture

Nicholas Journet, Rémy Mullot, Véronique Eglin, Jean-Yves Ramel

Analyse d'images de documents anciens : Catégorisation de contenus par approche texture

N. Journet[1], R. Mullot[1], V. Eglin[2], J.Y. Ramel[3]

1 Laboratoire L3i, Université de la Rochelle 1 17042 La Rochelle Cedex 1 - FRANCE

2 LIRIS INSA de Lyon, 69621 Villeurbanne cedex - FRANCE

3 LI-RFAI, 64 Avenue Jean Portalis 37200 TOURS - FRANCE

{njournet,rmullot}@univ-lr.fr, Jean-yves.ramel@univ-tours.fr,

veronique.eglin@insa-lyon.fr

Résumé *: Nous proposons une caractérisation du contenu des ouvrages anciens basée sur une approche texture non paramétrique. Cette démarche se veut générique et adaptable à tout type d'ouvrages en s'appuyant sur l'homogénéité des textures que l'on retrouve dans un ouvrage. En appliquant à plusieurs résolutions 5 algorithmes d'extractions de textures il est possible de caractériser le contenu des pages d'un ouvrage. Cette méthode est appliquée sur des pages d'ouvrages anciens du 16ème siècle.*

Mots-clés *: Approche texture, caractérisation de contenu, méthode non paramétrique, documents anciens*

1 Introduction

La conservation des documents du patrimoine et leur accès au plus grand nombre constitue aujourd'hui un enjeu majeur. Il suffit pour s'en convaincre de voir les offensives des moteurs de recherche comme Google, qui loin de porter ses efforts uniquement sur les documents contemporains, propose au travers de « Google print », une recherche dans des bases documentaires de livres anciens.

Notre contribution se situe dans le cadre de la construction de la bibliothèque humaniste virtuelle, projet porté par le CESR (centre d'Etudes Supérieures de la Renaissance), à Tours. Ce projet propose de mettre en ligne des bases d'ouvrages du 16ème siècle, en version image, avec une indexation permettant de retrouver l'information d'une part par des mots clés correspondant aux métadonnées produites manuellement à partir d'index proposés par les humanistes et d'autre part, par une indexation à partir de l'analyse du contenu des livres. Notre contribution porte plus particulièrement sur ce second point, puisque nous proposons une aide à l'expertise pour une indexation des contenus.

Lorsque l'on parle d'indexation du contenu, plusieurs orientations sont généralement proposées. La première qui semble la plus naturelle consiste à proposer une analyse du texte lui-même. Il faut dès lors employer des moteurs de reconnaissance de caractères et de mots qui fonctionnent de façon performante sur les textes de documents contemporains. Malheureusement, les tests effectués sur les textes anciens donnent des performances décevantes [BOU01], nécessitant des corrections nombreuses et coûteuses.

La seconde approche consiste à analyser la structure du document ou de l'ouvrage. Il s'agit plus ici de retrouver la structure physique et logique donnant une autre forme d'information que le contenu sémantique. Cette structure est riche et permet de naviguer rapidement dans l'ouvrage facilitant ainsi la recherche d'information (sommaires, titres, …).

Cet article présente une démarche originale de caractérisation d'images de documents. A l'instar de ce qui se fait dans le domaine de l'indexation d'images naturelles, nous proposons une démarche axée sur l'extraction d'une multitude d'informations issues d'une analyse de *texture* sur l'image. L'étape suivante consiste à analyser toutes ces informations extraites pour rechercher l'homogénéité de la représentation au sein du même ouvrage. En effet, lors de la conception de l'ouvrage, l'imprimeur a utilisé une même modalité de représentation pour marquer un contenu. Par exemple, les corps de texte sont représentés par une typographie unique, ainsi que par des espaces inter-lignes fixes définissant une signature caractéristique. C'est ce type de caractéristiques sur lesquelles nous allons construire notre approche.

Dans un premier temps, nous allons présenter les spécificités des ouvrages anciens comparés à ceux du monde contemporain et les outils les plus utilisés afin d'extraire les différentes couches d'informations. Puis nous présenterons notre méthode pour finalement présenter une évaluation de la caractérisation du contenu.

2 Les méthodes d'extraction d'information

2.1 Quelques mots sur les projets de traitement de documents anciens

Le partimoine numérisé est la source de travaux de recherche importants. Parmi les projets de grande

envergure on peut citer le projet DEBORA [BAR02] ou encore les projets BAMBI [BOZ97] et METAe : [JOU04] qui ont permis d'offrir des premières solutions en matière de transcription de caractères, annotations des images, navigation dans un corpus… Mais à l'heure actuelle, la valorisation des ouvrages anciens passe également par la création d'outils d'indexation ne portant pas que sur le contenu textuel. Ainsi, les auteurs de [UTT05, PAR05] posent les premières pierres visant à permettre l'indexation de dessins de traits. En ce qui concerne la segmentation ou la dématérialisation de documents anciens, des outils présentés dans [RAM06] ont déjà permis d'indexer plus de 17000 pages d'ouvrages archivés au Centre d'Etude Supérieur de Tours. Dans [COU03] l'auteur présente une plate forme d'indexation d'images d'archives militaires qui a permis d'indexer automatiquement des éléments de structure de plus de 250.000 images.

2.2 Caractérisation du contenu de documents

Les caractéristiques des documents anciens portent avant tout sur une hétérogénéité forte des ouvrages traités. Une harmonisation des présentations et des règles éditoriales a pris plusieurs siècles, ce qui du coup se traduit par une variété de livres où des différences de mise en page, de typographie, de style d'illustrations sont fortement présentes.

A cette spécificité vient s'ajouter des caractéristiques de dégradation (pages jaunies, tâches d'encre, aspect visible de l'encre du verso…) et de défauts de numérisations (défauts de courbure, de lumière…) qui rendent complexe tout traitement de caractérisation ou de segmentation des ces images. Pour amenuiser ces bruits, des processus dédiés au traitement des documents anciens ont été développés [DIG et TRI03].

Traditionnellement, les méthodes d'analyse de la structure sont regroupées en deux classes suivant qu'elles sont pilotées par les données, ou par les modèles. Dans la première classes d'approches, les méthodes sont appliquées dans la plupart des cas sur des documents binarisés et pour lesquels de nombreux seuils sont à régler. Ces seuils sont utilisés pour regrouper les pixels de l'image en entités physiques ([HAD03], [BRE02], [LIE02], [KHE03]). Nos tests inspirés de ces méthodes montrent qu'elles sont souvent inappropriées aux documents anciens du fait de leur hétérogénéité. La Figure 1 illustre les limites d'une approche par composantes connexes. On peut par exemple citer les problèmes liés aux espaces inter-mots irréguliers (point 1), mais aussi les tâches d'encre dont les propriétés sont proches de celles d'un caractère (point 2) ou encore la difficulté de gérer les composantes inclues les unes dans les autres (point 3).

Dans la seconde classe d'approches, la connaissance précise du modèle est nécessaire pour guider les traitements dans cette extraction. Là encore, le cadre applicatif dans lequel nous nous sommes positionnés ne permet pas d'utiliser ces méthodes, puisqu'aucun modèle ne peut et ne doit être défini a priori si on s'intéresse à un corpus hétérogène.

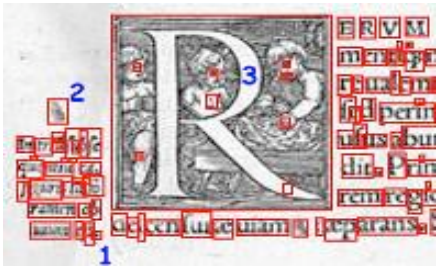


Figure 1: Limites d'une approche classique

Pour s'affranchir de connaissances a priori, les approches textures permettent de construire des modèles par analyse du contenu des pages ([GUP00], [LI00], [ETE97] [JAI92] [BAS04]). Ces approches se basent sur l'utilisation d'outils comme les filtres de Gabor ou les ondelettes et permettent de séparer le texte des illustrations en s'appuyant sur l'étude des fréquences et des orientations des différentes parties de l'image. Nos tests ont montré que l'application de ces filtres, et en particulier les filtres de Gabor, (utilisés jusque là sur des documents contemporains) étaient difficilement directement transposables sur des images de documents anciens. En effet, comme le montre la figure 2, il y a une différence importante entre les fréquences caractéristiques d'une zone de texte et d'une photo ce qui conduit à une une bonne séparation texte/dessin. Cependant, on ne retrouve pas la même caractéristique avec des dessins de traits où le grand nombre de transitions entre le fond et l'encre se caractérise par des fréquences proches de celles d'une zone de texte : cela rend la séparation entre texte et dessin difficile. De plus, comme le souligne [LEE01], la principale difficulté de ces approches portent sur leur complexité algorithmique dûe à une paramétrisation des filtres utilisés pour rechercher des fréquences et des orientations.

Figure2 : Application de bancs de filtres de Gabor inspirés de [BAS04] (Texte en rouge et illustrations en

bleu et fond en vert)

3 Extraction d'attributs de texture

3.1 Principe de notre approche

Le but de notre approche est de réaliser une caractérisation robuste du contenu des documents issus d'ouvrages anciens provenant d'horizons très différents. Pour se faire, nous allons extraire des caractéristiques pour chaque pixel de la page. Ces caractéristiques sont extraites par différentes techniques itérées à différentes résolutions. Il est en effet difficile d'exprimer une caractéristique qui soit générique d'un document à l'autre sans prendre en compte l'expression de l'échelle. Cette contrainte est ainsi très largement estompée avec l'introduction de la multi résolution.

Dans notre cas, 5 cartes sont calculées à 4 résolutions différentes. Chaque carte exprime le résultat d'une analyse spécifique (fréquence, orientation, …) sous la forme d'une carte de valeurs. A l'aide d'une analyse par fenêtre glissante il est possible d'affecter à chaque pixel de l'image des valeurs correspondant aux résultats d'extraction d'attributs texture. Cette analyse est réalisée à 4 résolutions différentes, donnant au final 20 valeurs (cartes) pour chaque pixel.

3.2 Informations textures liées aux orientations

L'orientation est l'une des caractéristiques visuelles impliquée dans la vision préattentive. Le modèle de Itti [ITT00] l'utilise pour caractériser les points de saillance dans les images naturelles pour son côté discriminant. Nous avons choisi d'étudier les orientations des textures à l'aide de la rose des directions proposée par Bres dans [BRE94].

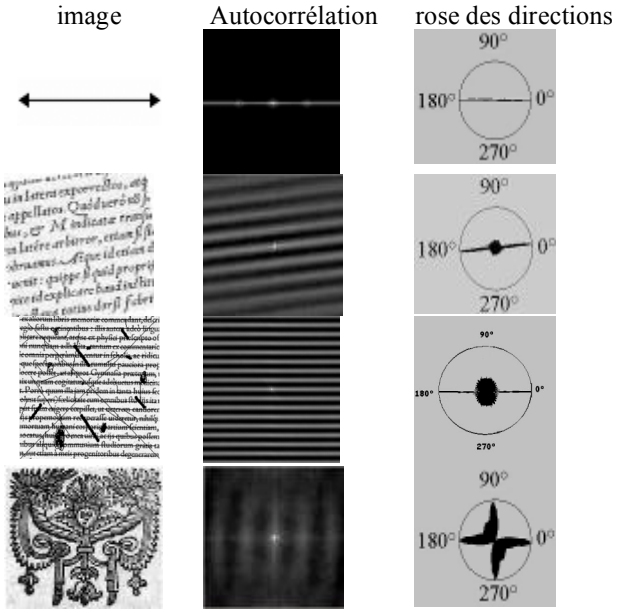


Figure 3 : Exemple de roses des directions

La rose des directions est un diagramme polaire permettant d'interpréter le résultat de la fonction d'autocorrélation appliquée sur une image. Elle indique de manière très précise les orientations présentes. Elle a en plus la particularité d'être peu sensible aux bruits de type « tache d'encre » qui sont fréquentes dans les images que nous traitons. La figure 3 illustre la réponse de cette méthode à différents types d'images.

Dans le domaine de l'analyse d'images de documents la multirésolution permet de percevoir des structures de tailles différentes. Dans notre appliquation nous avons choisi de partir de la taille initiale de l'image et de réduire par 2 les dimensions à chaque changement de résolution. La figure 4 illustre l'intérêt de ce choix pour un calcul de la rose des directions. On voit ainsi très nettement que la forme de la rose est différente selon la taille de la résolution choisie pour analyser l'image.

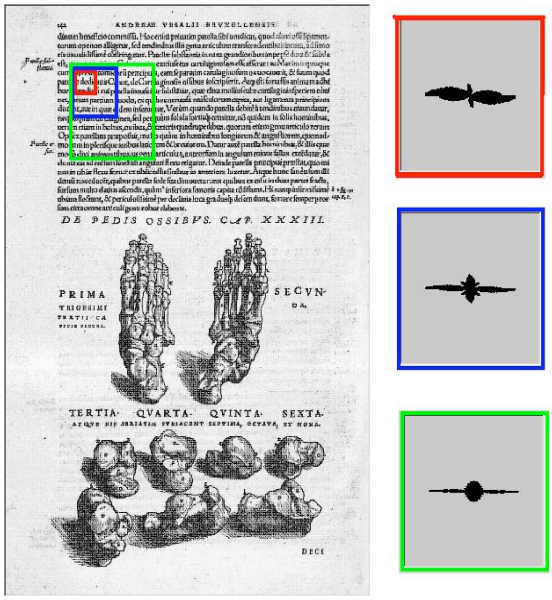


Figure 4 : Intérêt d'un calcul à multirésolution

A partir de ce diagramme, trois caractéristques sont extraites permettant de constituer 3 cartes. La première d'entre elles est tout simplement liée à la valeur de l'orientation principale (formule 1).

$$Carte1^k(i,j) = ArgMax(Rose(i,j)) \quad \textbf{(1)}$$

Avec $Carte1^k(i,j)$ la valeur de l'attribut texture 1 (i,j) pour la résolution k et Rose(i,j) la rose des directions calculée au point (i,j).

Les deuxième et troisième informations de texture extraites sont liées à la valeur de la fonction d'autocorrélation pour l'orientation principale de la rose (identifiée lors du calcul de la carte 1). Du fait de sa définition mathématique issue de la fonction d'autocorrélation, la rose des directions permet de caractériser l'anisotropie ou de l'isotropie de la texture étudiée. C'est pour cela que nous calculons d'une part l'intensité de la corrélation des pixels selon l'angle trouvé par l'équation 1 (équation 2) , et que d'autre part nous calculons l'importance de l'intensité de cette orientation relativement aux autres (équation 3).

$$Carte2^k(i,j) = R(ArgMax(Rose(i,j))) \quad (2)$$

Avec R la valeur de la rose des directions pour un angle donné.

$$Carte3^k(i,j) = \sum_{theta=0}^{theta=359} R(theta) \quad (3)$$

Avec $theta \neq ArgMax(Rose(i,j))$

3.3 Informations liées aux fréquences

Si on a plus couramment l'habitude d'utiliser des outils comme la transformée de Fourier, les filtres de Gabor ou les ondelettes pour caractériser les informations de fréquences [ZHU01], nous avons décidé de ne pas le faire pour caractériser celles des documents anciens. En effet, la phase « délicate » de la paramétrisation associée au fait que les propriétés fréquentielles d'une zone de texte et des dessins de traits sont parfois similaires nous a poussé à explorer d'autres voies pour exprimer les caractéristiques liées aux fréquences. Nous nous sommes notamment inspirés des travaux de [EGL98], [ALL04] et [RAM06] qui détaillent comment il est possible de caractériser différents types de texte ou de séparer le texte des illustrations en étudiant les propriétés des transitions des niveaux de gris des pixels.

Le troisième attribut de texture se calcule en appliquant la formule 4 et permet, pour chaque pixel, d'obtenir une information sur les transitions (en niveaux de gris) entre les pixels de fond et d'encre.

$$Carte4^k(i,j) = \underset{h\in(0,H)}{S}\left(\sum_{l=1}^{L} I(l,j) - I(l+1,j)\right) \quad (4)$$

Avec $I(l,j)$ le niveau de gris du pixel (l,j), L et H la largeur et la hauteur de la fenêtre d'analyse et S la variance des transitions de chaque ligne de la fenêtre.

Le dernier indice texture calculé est inspiré des travaux d'analyse des longueurs de plage. Au travers d'une étude récursive de l'image (4 itérations de l'algorithme classique X-Y cut) où pour chaque itération nous allons calculer la somme des niveaux de gris en colonne et en ligne et finalement en faire la moyenne (eq 5)

$$Carte5^k(i,j) = \frac{\sum_{l=1}^{L} I(l,j) + \sum_{h=1}^{H} I(i,h)}{2} \quad (5)$$

4 Exploitation des indices de texture pour la caractérisation du contenu

4.1 Analyse préalable des indices textures

Avant d'évaluer la pertinence des indices extraits précédemment, nous allons dans un premier temps analyser le contenu de ces données. Après avoir étudié à l'aide de l'ACP près de 200 pages issues de 9 ouvrages différents, nous avons pu arriver à plusieurs conclusions. Tout d'abord il se trouve que l'inertie portée par les 4 premiers axes est toujours très bonne (90% des images analysées ont une inertie cumulée supérieur à 75%) et qu'il est donc possible de réduire les données générées en s'appuyant sur la redondance d'information. Ce choix de réduction est cependant est assujetti à un calcul d'une ACP par image car nos tests ont montré également que les vecteurs propres des variables diffèrent selon le contenu de l'image. La seule corrélation existante quelque soit le cas de figure est entre la carte 2 et la carte 5. Ceci s'explique entre autre par la forte isotropie de nos images (le texte est toujours horizontal et le texte est très fortement présent dans nos images). La carte 2 va donc trouver une forte corrélation horizontale des pixels relatifs à l'encre et la carte 5 va également être sensible aux longues plages d'informations, mais concernant cette fois les plages blanches. Il est donc envisageable de ne calculer qu'un seul de ces deux attributs de texture étant donné qu'ils sont anticorrelés. La dernière analyse est relative à la question de l'échantillonnage. Les images traitées étant de grande taille, il est intéressant de savoir s'il est possible (et dans qu'elle mesure) d'échantillonner les pixels. En se basant sur une analyse successive du cercle des corrélations, des valeurs propres et des vecteurs propres, il a fallu descendre en dessous d'une taille d'échantillon de 0.1% du nombre de pixels de l'image initiale (les pixels sont pris au hasard) avant de voir apparaître une différence significative de ces indices.

4.2 Classification non supervisée des contenus

Effectuer une classification des pixels à partir des indices texture extraits est avant tout l'occasion de voir si la caractérisation du contenu est conforme à notre objectif de séparation de l'information en couches. On rappelle que chaque pixel de l'image dispose de 20 valeurs issues des 20 cartes construites. Notre objectif étant de regrouper les pixels de l'image correspondant à des zones homogènes, cela consiste à regrouper les vecteurs proches au sens d'une métrique. Ce problème est un problème de classification non supervisée pour lequel nous ne connaissons pas a priori les étiquettes des points permettant de construire les classes. La mise en œuvre de cette classification a été confiée à l'algorithme Clara qui permet de manipuler des vecteurs de grande dimension et un nombre important de données. Cet algorithme opère en deux étapes. Dans un premier temps un partitionnement d'un échantillon de pixels est calculé, le nombre de classes étant un paramètre de l'algorithme. Ensuite, chaque pixel restant est classé dans une de ces partitions (sans rejet possible). Clara est décrit en détail dans [KAU90].

La figure 5 montre le type de résultats que l'on obtient à l'issue de ce traitement. La classification a été calculée sur un ouvrage complet. De ce fait, si deux pixels ont la même couleur, cela signifie qu'ils appartiennent à la même classe.

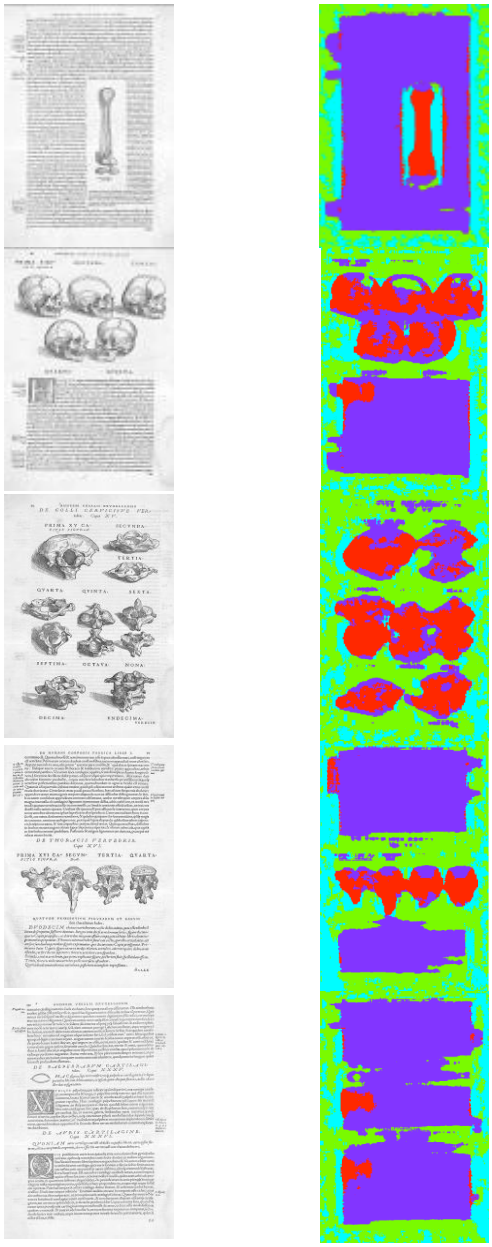


Figure 5 : Résultat d'une classification à 4 classes

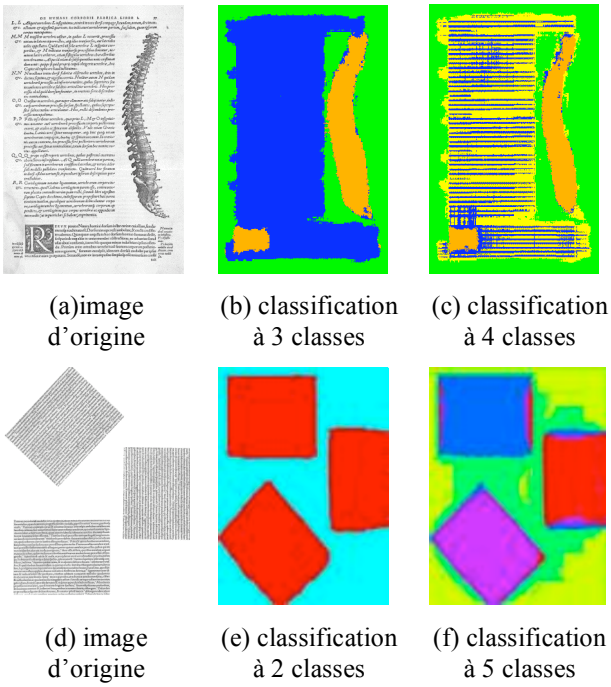
4.3 Analyse des résultats

Les figure 5 et 6 donne donc une première indication sur le pouvoir discriminent des indices textures. Si dans ces exemples le choix du nombre de classes reste arbitraire, on voit néanmoins certaines choses intéressantes apparaître en fonction de ce paramètre (par exemple les espaces interlignes). En effet, les tests visant à séparer les pixels en 3 classes montrent que dans la plupart des cas, le texte est très bien caractérisé ainsi que le fond. Les graphiques, du fait de leur constitution intrinsèque très variable sont parfois moins bien extraits.

En ce qui concerne les erreurs de marquage observées, elles interviennent principalement au niveau des zones de transition entre textes et images mais aussi au niveau des titres contenant de gros caractères isolés, ce qui fait qu'une partie des titres (isolés du corps de texte) sont marqués comme des dessins.

Afin d'apporter une évaluation chiffrée de ce marquage, nous proposons d'évaluer la capacité de l'algorithme Clara à pouvoir séparer les pixels en 3 classes : Texte/fond/dessins.

Pour ce faire nous avons saisie manuellement la vérité terrain de près de 200 pages extraites de 9 ouvrages. Chaque classification est donc comparée aux fichiers de vérité terrain ce qui nous permet au final de chiffrer à 83% la proportion de pixels de dessins bien marqués et à 92% la proportion de pixels de texte correctement identifiés.



(a)image d'origine | (b) classification à 3 classes | (c) classification à 4 classes

(d) image d'origine | (e) classification à 2 classes | (f) classification à 5 classes

Figure 6 : Exemples de classifications

Ces résultats sont à relativiser car ils ne font pas fois de qualité du potentiel de segmentation. Il faudrait pour cela mettre nous même en place notre propre algorithme de classification. Il faudrait également évaluer une capacité à segmenter en « entités » et pas seulement à marquer les pixels. Il faudrait, par exemple, un taux de reconnaissance lié à la capacité à identifier des lignes, des paragraphes, des titres…

5 Conclusion

Nous avons présenté dans cet article, une méthode de séparation des éléments physiques des pages d'ouvrages anciens. Cette méthode se veut peu paramétrique, et adaptable à tout type de documents anciens. La spécificité repose sur l'extraction d'indices textures dédiés aux documents ainsi qu'à la recherche d'éléments physiques homogènes au sein de l'ouvrage lui-même. La caractérisation de ces contenus est composée d'une part de la construction de cartes correspondant à des caractéristiques différentes (fréquences, orientations, ..), et à des résolutions différentes. Cette approche très générique est intéressante puisqu'elle laisse la possibilité d'ajouter des cartes supplémentaires correspondant à des caractéristiques non extraites dans cette version de la

méthode. L'évolutivité du système est donc simple à mettre en place.

À l'issue de cette construction des cartes, une analyse de données montre qu'il est possible de réduire la taille des données traitées en jouant sur la redondance des informations extraites. Une illustration par une caractérisation des contenus sur 3 à 5 classes montre la pertinence de la démarche. Cette première validation nous a permis de passer à la réalisation d'outils d'aide à l'indexation et à la navigation se basant sur ces indices extraits.

6 Bibliographie

[ITT 00] Itti, L., Koch, C. A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research*,40(10-12):1489-1506, 2000.

[SAP78] JM Bouroche, G Saporta. Que sais-je ? n° 1854 : L'analyse de données. Presses universitaires. 1978

[BOU01] F Le bourgeois ,H Emptoz, E Trinh, F Muge, C Pinto I Granado. DEBORA WP 4.3 1 WP4.4 Description du matériel et logiciel de traitement d'image pour la numérisation des collections et leur interprétation. 2001.

[DIG] Digibook, http://www.i2s-bookscanner.com/fr/default.asp

[TRI03] TRINH, E., De la numérisation à la consultation de documents anciens. Thèse de doctorat en Informatique. Insa de Lyon. 175p., 2003.

[HAD03] HADJAR, K., INGOLD, R. Arabic Newspaper Page Segmentation, in ICDAR, pp. 895-900, 2003.

[BRE02] BREUEL, M;T. Two Geometric Algorithms for Layout Analysis, Xerox Palo Alto Research Center, in DAS, pp. 214-222, 2002.

[LIE98] LIE, J., HU, J., WU, L. Page segmentation of chinese newspaper, in Pattern recognition, 2695-2704, 2002.

[CIN98] L. Cinque, L. Lombardi, G. Manzini, A multiresolution approach for page segmentation, Pattern Recognition Letters 19(2): 217-225 (1998)

[KHE03] Khedekar, s.,Ramanaprasad, v., Setlur S., Govindaraju, V. Text - Image Separation in Devanagari Documents, in proc. On ICDAR, vol.2, pp. 1265-1269, 2003.

[GUP00] P.Gupta, N.Vohra, S.Chaudhury, S.Dutt, Wavelet Based Page Segmentation, Proc. of the ICVGIP, pp.51-56, 2000.

[LI00] Jia Li, Robert M. Gray, Context-Based Multiscale Classification of Document Images Using Wavelet Coefficient Distributions, IEEE. Image Processing, vol. 9, pp. 1604-1616, Sept. 2000.

[ETE97] K.Etemad, D.Doermann, R.Chellappa, Multiscale Segmentation of Unstructured Document Pages Using Soft Decision Integration, IEEE Trans. on Pattern Analysis and Machine Intelligence, v.19 n.1, p92-96, 1997.

[JAI92] A.K. Jain, S.Bhattacharjee. Text segmentation using Gabor filters for automatic document processing, Machine Vision and Applications,Vol.5,169 – 184. 1992.

[BAS04] P.Basa P.S. Sabari, R.Nishikanta, P. A G Ramakrishnan, Gabor filters for Document analysis in Indian Bilingual Documents, In Proceedings International Conference on Intelligent Sensing and Information Processing., pages pp. 123-126, 2004.

[LEE01] Lee, S.W., Ryu, D.S., Parameter-Free Geometric Document Layout Analysis, IEEE Tran.on PAMI., Vol. 23, No. 11, p1240-1256, 2001.

[BRE94] S BRES, Contributions à la quantification des critères de transparence et d'anisotropie par une approche globale. PhD Thesis, 1994.

[KAU90] Kaufman, L.,Rousseeuw, P. J. Finding Groups in Data,, Wiley Series in Probability and Mathematical Statistics, 340 p.,1990.

[BAR02] Nicolas Barbey, Jean Guillemain, Géraldine Péoc'h, Patrice Ract La renaissance du livre ancien : bilan du projet DEBORA et perspectives d'avenir. Sous la direction de François Dupuigrenet Desroussilles Directeur de l'Ecole nationale supérieure des Sciences de l'Information et des Bibliothèques .

[BOZ97] Bozzi A., Calabretto S. « Digital Library and Computational Philology : the BAMBI (LIB -3114) project. Proceedings of the First European Conference on Research and Advanced Technology for Digital Libraries ». Lecture Notes in *Computer Science* N·1324 (Springer Verlag). Eds. C. Peters and C. Thanos. Pisa, Italie. September 1-3, 1997. pp. 269-285. ISBN 3-540-63554-8

[UTT05] S. Uttama, J Ogier, P Loonis, Top-down segmentation of ancient graphical drop caps : lettrines, GREC05, IAPR, pp 87-95, 2005

[RAM06]J.Y. Ramel and S. Busson and M.L. Demonet AGORA: the Interactive Document Image Analysis Tool of the BVH Project. DIAL , année 2006, pages = 145-155.

[JOU04] Dominique Jourdy Culture & Recherche n°100 février 04. Journal du ministère de la culture et de la communication. 2004.

[PAR05] R. Pareti, N. Vincent. Global discrimination of graphics styles, 6th International Workshop on Graphics Recognition (GREC 05), 25-26 août 2005, Hong-Kong, pp. 120-128

[COU03] B. Couasnon, J. Camillerapp : « Accès par le contenu aux documents manuscrits d'archives numérisées », Document numérique, vol. 7, 2003, éditions Lavoisier Hermès, p. 61-84.

[EGL98] V Eglin. Contribution à la structuration fonctionnelle des documents imprimés Phd Thesis LIRIS-université de Lyon. 1998.

[ALL04] B Allier, Contribution à la Numérisation des Collections : Apports des Contours Actifs LIRIS-université de Lyon. 2004.

[ZHU01] Y. Zhu and T. Tan and Y. Wang. Font Recognition Based on Global Texture Analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence, volume 23 n 10, pp 1192-1200. 2001