Analyse des Orientations pour la Caractérisation d'Images de Documents de la Renaissance

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

[1]L3I, Pôle Sciences et Technologie 17042 La Rochelle Cedex 1 - FRANCE njournet@univ-lr.fr

[2]LI, 64 Avenue Jean Portalis 37200 TOURS - FRANCE Jean-yves.ramel@univ-tours.fr

[3]LIRIS INSA de Lyon, Bâtiment Jules Verne Campus de la doua 69622 Villeurbanne cedex - FRANCE eglin@rfv.insa-lyon.fr

Résumé – Cet article présente une nouvelle méthode de caractérisation d'images de documents imprimés datant de la Renaissance. Notre approche se base sur une extraction des différentes orientations présentes sur la totalité de la surface de la page et qui sont caractéristiques de la présence de différentes entités textuelles, ou graphiques (incluant les enluminures, les ornements et bandeaux, les lettrines, ainsi que diverses illustrations). Cette caractérisation s'appuie sur le calcul et l'exploitation de la fonction d'autocorrélation qui a la particularité, lorsqu'elle est estimée sur une zone de texte ou de dessin, de générer une signature unique facilement identifiable. Ce choix nous permet de séparer le texte des dessins, tout en minimisant la quantité d'a priori relatif aux images traitées.

Abstract – This paper presents a method of a page layout characterization adapted to ancient printed documents. Our approach is based on the extraction and the quantification of the various orientations that are present in a document image. The documents are analysed using a normalized square windows in which we analyze significant orientations with a directional rose. Each kind of information (textual or graphical) is typically identified and marked by its orientation distribution. This choice of characterization allows us to separate textual regions from graphics by minimizing the a priori knowledge.

1. Introduction

Ces dernières années ont été marquées par la mise en place de nombreuses campagnes de numérisation. Au fil des années les historiens ont ainsi accumulé une grande quantité d'images numériques de documents anciens. Ces ouvrages proviennent de différents pays (France, Allemagne, Italie…) et recouvrent plus de trois siècles de notre histoire (du milieu du 15ème à la fin du 17ème) pour le patrimoine imprimé. Pendant de longues décennies le monde de l'imprimerie s'est naturellement inspiré des traditions héritées de l'écriture [1]. Les choix relatifs à la mise en page ou la typographie s'en trouvaient donc souvent liés aux seuls choix de l'imprimeur. Ainsi, d'un pays à un autre (ou d'un siècle à l'autre) les ouvrages produits sont caractérisés par une forte variabilité de mises en page et de règles d'édition. Du fait des contraintes techniques et historiques, on observe des structures visuellement très complexes basées sur le multi-colonnage, des corps de texte contenus dans des zones non rectangulaires, des localisations imprévisibles des dessins, ou encore des espacements entre zones graphiques ou textuelles non normalisés, voir figure 1.

FIG. 1 : Exemples de pages de documents anciens actuellement archivés au CESR.

Ce constat nous a poussés à aborder le problème de l'analyse d'images de documents imprimés anciens sous un autre angle. En effet, cette forte hétérogénéité du contenu comme de la mise en page, ne nous permet pas d'utiliser des méthodes classiques de segmentation ou de caractérisation de l'information (approches descendantes, ascendantes ou mixtes).

Dans cet article nous présentons donc une nouvelle approche de caractérisation d'images de documents imprimés extraits d'un vaste corpus de la Renaissance et actuellement numérisés au CESR de Tours. Cette méthode est non seulement robuste aux bruits fréquemment rencontrés dans ce genre d'images (détériorations de l'encre, défauts de numérisation…) mais elle se veut aussi complètement indépendante de la typographie employée, de la taille des caractères, de la présence de parties graphiques… Ce travail est une contribution à l'analyse de la mise en page de documents par une séparation texte/graphique basée sur une localisation et un marquage en zones d'intérêt. L'originalité de ce travail vient de sa grande généricité (applicabilité à des documents de structures très hétérogènes), de l'absence quasi-totale de connaissances a priori sur les propriétés des textes, des graphiques et de leur agencement et de sa grande robustesse aux variations locales de contrastes, aux changements de résolution, aux phénomènes fréquents d'inclusion de texte dans les graphiques.

2. Un constat sur les méthodes d'analyse d'ouvrages anciens

Dans le domaine du document ancien, il n'y a pas un modèle universel. Il est cependant possible de dégager certaines caractéristiques communes à tous ces livres, concernant notamment les mises en page complexes

(plusieurs colonnes de taille irrégulière), les fontes spécifiques ou inconnues, l'utilisation fréquente d'ornements (enluminures, lettrines, cadres…), le faible espacement entre les lignes, ou encore l'espacement non constant entre les caractères et les mots, la superposition de couches d'information (bruit, notes manuscrites…)

Les méthodes d'analyse de documents dédiées aux documents anciens sont quasiment inexistantes, la plupart sont majoritairement dédiées aux documents contemporains. Nous avons cherché à les évaluer sur les ouvrages du patrimoine et avons pu constater qu'elles n'étaient pas directement transposables : nous avons tout d'abord testé des méthodes procédant par agglomération ou découpage de blocs dans les pages. Que ce soit les approches ascendantes (RLSA, extraction de composantes connexes [2,3]) ou les approches descendantes ou mixtes (projections horizontales, analyse multirésolution… [4,5,6]), ces deux catégories de méthodes demandent une grande quantité de connaissances a priori à introduire. Les algorithmes mis en place étaient spécifiquement adaptés à chaque famille de document voir même à chaque type de pages.

Nous avons ensuite testé des outils de caractérisation de textures qui permettent de séparer le texte des dessins sans avoir besoin de connaissances sur le modèle des documents traités. Parmi les travaux les plus pertinents on peut citer ceux se basant sur l'extraction de la matrice de transition des niveaux de gris ou l'application de filtres de Gabor [7]. Ces méthodes permettent d'exprimer l'aspect fortement texturé du texte et donc d'extraire du texte quelque soit ses caractéristiques typographique. Nous avons finalement décidé de nous appuyer sur une fonction d'autocorrélation. Cette dernière nous permet de caractériser, de manière unique, l'anisotropie de grande zones de texte.

3. Caractérisation des zones d'intérêt par extraction des orientations

3.1 Principe de notre méthode

L'objectif de notre méthode est de mettre en place un système d'indexation robuste adapté à un corpus d'images au contenu hétérogène. La première étape de notre méthode consiste à extraire visuellement la mise en page du document en séparant grossièrement les zones de texte (caractères d'imprimerie) des zones de dessin (lettrines, enluminures…). L'image étudiée est tout d'abord redimensionnée pour traiter par la suite des images de taille fixe et connue (900X600). Le fond est ensuite identifié à l'aide d'un algorithme basé sur l'étude de l'homogénéité des niveaux de gris selon la méthode de Sauvola et al [8]. Le marquage des pixels est réalisé à l'aide d'une fenêtre d'analyse de taille fixe que l'on déplace sur l'image. A chaque déplacement, les orientations caractéristiques de la zone sont extraites et analysées afin de labelliser chaque pixel inclus dans la fenêtre comme étant un pixel appartenant à du texte ou un dessin. Cette technique de marquage implique que les pixels peuvent être marqués plusieurs fois, c'est pourquoi un algorithme permettant de choisir le label final de chaque pixel a été mis en place.

3.2 Extraction des orientations

L'extraction des orientations d'une zone de l'image se fait à l'aide d'une fonction d'autocorrélation (form.1).

$$C_{xx}(k,l) = \sum_{k'=-\infty}^{+\infty} \sum_{l'=-\infty}^{+\infty} x(k',l').x(k'+k,l'+l) \quad (1)$$

La rose des directions, outil proposé par Bres [9], permet d'effectuer une analyse de cette fonction d'autocorrélation. Cette rose donne, avec une grande précision, les orientations privilégiées présentes dans le bloc étudié. La rose est en fait un diagramme polaire. Soit (u,v) le point central de l'autocorrelation et θi l'orientation étudiée, on calcule alors la droite Di tel l'ensemble de ses points (a,b) respecte la relation suivante : angle formé par les deux droites $(a,b)(u,v) = \theta_i$. Pour chaque orientation θi on calcule ainsi la somme des différentes valeurs de la fonction d'autocorrélation (form. 2).

$$R(\theta_i) = \sum_{D_i} C_{xx}(a,b) \quad (2)$$

Ces valeurs sont ensuite normalisées (form. 3) pour ne garder qu'un aspect relatif de la contribution de chaque orientation.

$$R'(\theta_i) = \frac{R(\theta_i) - R_{\min}}{R_{\max} - R_{\min}} \quad (3)$$

La figure 2 montre quelques exemples de roses des directions pour 3 images différentes.

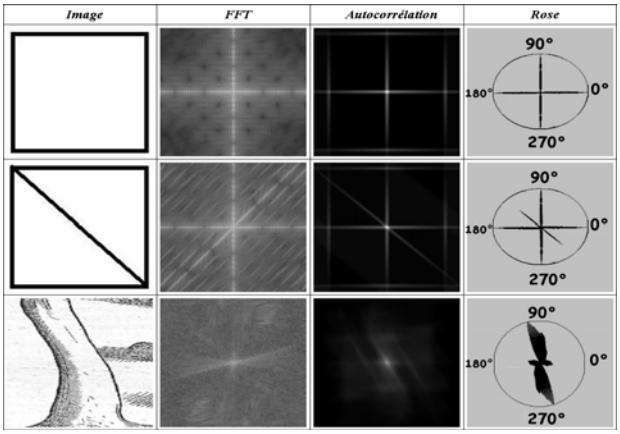

FIG. 2 : Exemples de roses des directions

3.3 Extraction de la mise en page

Compte tenu de sa définition mathématique, la rose possède des propriétés intéressantes. En effet, quand la zone étudiée possède une répartition homogène des directions, la rose produite est une boule parfaite (fig. 3.a). Si on ajoute une ligne à l'image précédente (fig 3.b), le fait que la rose soit normée génère deux caractéristiques invariantes : deux pics horizontaux de même intensité (pour 0° et 180°) et une même intensité pour toutes les autres directions (c'est ce qui donne cette impression de boule).

Une autre propriété intéressante est que la rose est robuste au changement de résolution. Nos tests ont confirmé que quelque soit la résolution à laquelle est calculée la rose, sa forme ne change pas (fig 4).

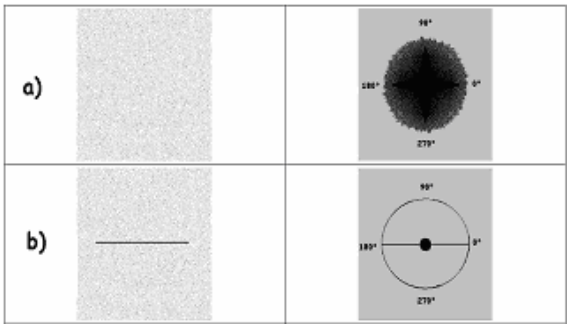

FIG 3 : Caractéristiques de la rose

Nos tests effectués sur le comportement de la rose face au contenu de notre corpus ont permis de déterminer 4 types caractéristiques de roses des directions.

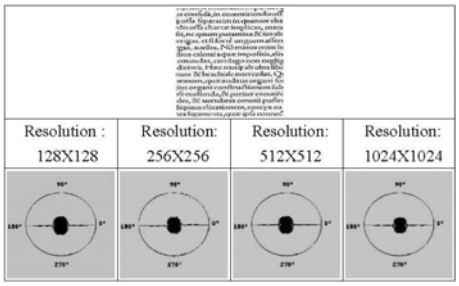

FIG. 4 : Comportement de la rose à différentes résolutions

Il y a deux types de roses qui permettent de caractériser du texte. Une zone de texte homogène (espace interligne constant, taille des caractères constants…) possède les mêmes propriétés que l'image b de la figure 3 : une répartition homogène des directions due aux formes des lettres et une forte corrélation horizontale due à l'orientation du texte (fig. 5.a). Il est donc possible de caractériser une zone homogène de texte par cette rose unique. Lorsque la zone n'est pas homogène (différentes taille de texte, espaces interligne variable…), la formule 3 implique que la taille de la « boule » peut diminuer voir même disparaître (fig. 5.b). Les deux autres formes de roses permettent d'identifier une zone de dessin. Ce type de rose est caractérisé soit par l'absence d'orientations horizontales (fig 5.c), soit par la présence d'une direction horizontale mais sans la présence de « pic » (fig 5.d). Toutes les autres formes de rose ne permettent pas de conclure sur la nature de la zone étudiée.

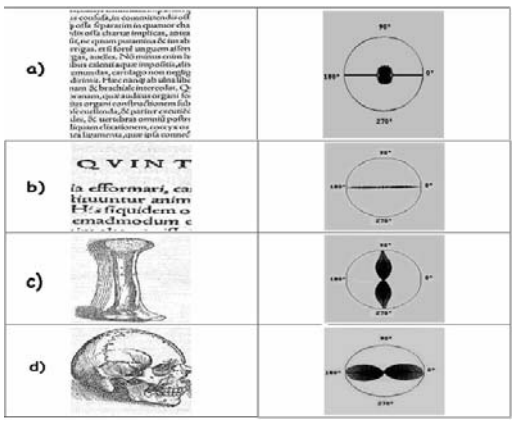

FIG. 5 : détermination de zones graphiques

Comme nous l'avons évoqué précédemment notre approche par fenêtre glissante implique qu'un pixel puisse être étiqueté plusieurs fois. La décision du label final de chaque pixel est prise une fois que le marquage est achevé. Pour décider du label final de chaque pixel de l'image, on comptabilise le nombre de fois ou il a été marqué comme texte par rapport au nombre de fois où il a été marqué comme graphique.

La taille de la fenêtre d'analyse est un paramètre important. Nous avons pour l'instant déterminé une taille optimale de la fenêtre pour analyser des images de documents ayant une résolution de 900X600. Expérimentalement nos tests ont montré que la taille de la fenêtre devait inclure un minimum de 5 à 7 lignes de texte pour permettre la détection d'une zone textuelle. Une alternative à ce choix consiste à laisser l'utilisateur, lors d'une phase d'initialisation, fixer lui-même la taille de cette fenêtre.

4. Résultats expérimentaux

4.1 Qualité de notre méthode

Le tableau 1 récapitule les résultats des tests effectués sur plus de 50 pages en provenant de 6 ouvrages différents de la Renaissance. Ces livres ont été choisis afin d'évaluer notre méthode sur des mises en page, des typographies et des résolutions variées. En ce qui concerne la qualité de la localisation des zones de texte, nous avons quantifié le pourcentage de lettres correctement identifiées, mal identifiées et enfin oubliées. En ce qui concerne la qualité de localisation de zones de dessin nous avons comptabilisé le nombre de pixels correctement identifiés, mal identifiés et enfin oubliés.

Tab. 1 Résultats (% de détection correcte / fausse / oubli)

Texte	Dessins
95/3/2	88/10/2

Quelques résultats sont présentés dans la figure 6.

FIG 6 : Quelques résultats (texte en bleu et dessin en rouge)

Dans la plupart des cas les différentes parties de l'image sont correctement détectées. Les erreurs principales de marquage proviennent de l'analyse de zones de transition entre textes et images mais aussi de titres contenant de gros caractères. Excepté ces deux cas, notre méthode donne de bons résultats et permet d'atteindre l'objectif principal qui est de mettre en évidence le contenu visuel du document sans forcément segmenter la page.

4.2 Utilisation de ce marquage

L'indexation de grosses quantités d'images au contenu hétérogène est un problème complexe. Nous pensons qu'il est possible, à l'aide de ce marquage, d'apporter plusieurs pistes de réponse à ce problème. Notre approche n'est pas une méthode de segmentation à proprement parler: elle a avant tout pour but de découper grossièrement l'image en zones labellisées. On peut trouver plusieurs intérêts à cette phase. Elle permet par exemple de mettre en évidence le contenu visuel du document et donc sa mise en page, elle permet également de simplifier l'application d'éventuels algorithmes de segmentation.

4.2.1 Catégorisation des pages d'un ouvrage

Le marquage obtenu (texte/graphique) peut servir de base pour comparer les pages d'un ouvrage entre elles afin de les regrouper en classes homogènes. La figure 7 démontre l'intérêt d'une telle approche en utilisant simplement comme critère de comparaison la quantité de pixels de dessin et de texte. Cette présentation a deux intérêts : d'une part elle permet d'illustrer la qualité de notre marquage et d'autre part, comme dans [10], nous voyons cette classification comme étant la première étape d'un travail d'indexation d'images de documents anciens. En analysant plus en profondeur les résultats de ce type de classification nous espérons pouvoir mettre en évidence une homogénéité intra-ouvrage (règles de mise en page, typographie spécifique…).

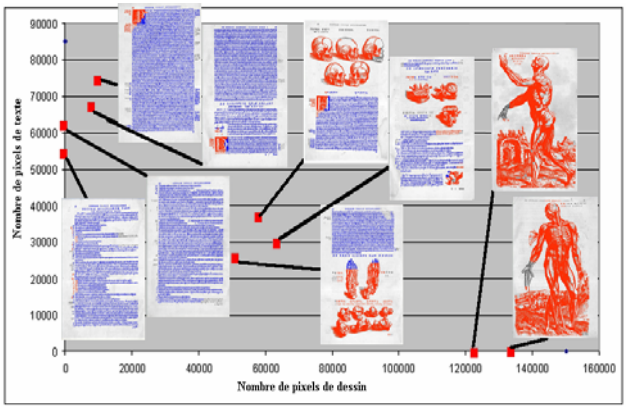

FIG 7 : Exemple d'une classification de pages d'un ouvrage

4.2.2 Segmentation fine des dessins

L'autre piste étudiée est celle d'une segmentation des illustrations basée sur les résultats du marquage. En effet, ce dernier fournis un a priori fort sur la localisation des dessins.

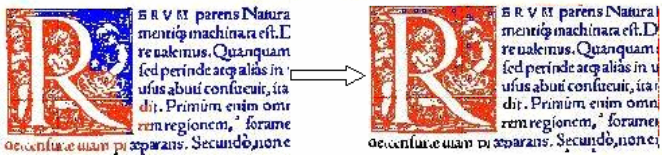

FIG. 8 : Segmentation basée sur l'étude du marquage

En se focalisant successivement sur toutes les régions rouges (graphiques) et en analysant le voisinage de chacune d'elles il est possible de localiser précisément les frontières des parties graphiques. Nos expérimentations montrent qu'il est possible, par exemple, de coupler l'information de notre marquage à un algorithme d'extraction de composantes connexes. Ainsi en étudiant l'information (dessin/texte) contenue dans chaque zone englobante, il est possible d'extraires de manière très fine les zones graphiques (fig. 8).

5. Conclusion

Dans cet article nous avons présenté une méthode de caractérisation d'images de documents imprimés anciens. L'originalité de notre approche tient à l'utilisation d'un outil qui permet de séparer, sans a priori, le texte des zones graphiques de différents types de documents. Le marquage obtenu constitue une représentation visuelle intéressante du contenu de chaque page et peut être utilisé par la suite aussi bien pour obtenir une segmentation physique fine que pour regrouper les pages d'un ouvrage en classes homogènes.

Références

[1] HJ Martin, La naissance du livre moderne, Editions du Cercle de la Librairie, 2000.

[2] Chew Lim Tan, Zheng Zhang Text block segmentation using pyramid structure. University of Singapore, 1998.

[3] L O'Gorman, The Document Spectrum for Page Analysis Layout, Trans IEEE One PAMI.15(11), 1993 P1162-1173.

[4] Jie Xi, J Hu, L Wu, Page segmentation of chinese newspaper Pattern recognition 2002 2695-2704

[5] D. Malerba, F Esposito Oronzo, Adaptive Layout Analysis of document. Università degli Studi di via Orabona Bari, ismis 2002.

[6] P. Duygulu, V. Atalay A Hierarchical Representation of Form Documents for Identification and Retrieval, International Journal on Document Analysis and Recognition IJDAR 5 2002 1, 17-27.

[7] P.B Pati,S Raju N Pali, AG Ramkrihnan. Gabor filters for document analysis in bilingual documents. DEE, Indian Institute of science. Bangalore 560 012 INDIA

[8] J. Sauvola, T. Seppänen, S. Haapakoski, and M. Pietikäinen. Adaptive Document Binarization. In ICDARecognition, volume 1, pages 147–152, 1997.

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

[10] G. Maderlechner, P. Suda, T. Bruckner, classification of documents by form and content, Siemens AG, Otto-Hahn-Ring, Munchen, Germany

Remerciements au CESR-CNRS (Tours – France) pour les images.http : www.cesr.univ-tours.fr