Université de La Rochelle - UFR Sciences

Ecole doctorale de La Rochelle

Analyse d'images de documents anciens : une approche texture

THÈSE

présentée et soutenue publiquement le 1er décembre 2006

pour l'obtention du

Doctorat de l'université de La Rochelle

(spécialité informatique)

par

Nicholas JOURNET

Directeur de thèse :	R Mullot	L3I, Professeur à l'université de La Rochelle
Rapporteurs :	R Ingold	Professeur à l'université de Fribourg
	JM Chassery	Directeur de Recherche - LIS-CNRS - Grenoble
Président du jury :		Université de
Examinateurs :	G Lorette	IRISA, Professeur à l'Université de Rennes 1
	JY Ramel	LI, MCF à l'université de Tours
	V Eglin	LIRIS, MCF à l'université de Lyon
Invités :	JM Ogier	L3I, Professeur à l'université de La Rochelle
	ML DEMONET	Professeur de Littérature française
		Directrice du CESR de Tours

Laboratoire Informatique Image Interaction — UPRES EA 2118

Table des matières

Table des figures **v**

Chapitre 1 Bibliothèques numériques et modalités d'accès au contenu **5**

1.1 Introduction . . . 5

1.2 Bibliothèques numériques indexées manuellement . . . 7

1.2.1 Quelques normes existantes . . . 7

1.2.2 Mise en place de systèmes d'indexation manuelle . . . 9

1.2.3 Conclusion . . . 12

1.3 Vers une indexation automatisée du contenu par analyse d'image . . . 13

1.3.1 Les OCR et leurs limites . . . 13

1.3.2 L'analyse de structure au travers des réalisations actuelles . . . 17

1.3.3 L'analyse des illustrations . . . 22

1.4 Conclusion . . . 23

Chapitre 2 Analyse de contenu d'images de documents : Etat de l'art **25**

2.1 Préambule . . . 25

2.1.1 Introduction . . . 25

2.1.2 Qu'est qu'une structure ? . . . 26

2.1.3 Notion de classe et de complexité de documents . . . 28

2.1.4 Notre corpus d'images de documents anciens . . . 30

2.2 Analyse du contenu d'images de documents : les approches classiques . . . 33

2.2.1 Quelques mots sur le problème de la binarisation des documents anciens . . . 34

2.2.2 Méthodes d'analyse de structures . . . 34

2.2.3 Conclusion . . . 46

2.3 Analyse de documents : les approches textures . . . 46

2.3.1 introduction . . . 46

2.3.2 Méthodes statistiques . . . 47

2.3.3 Méthodes géométriques . . . 50

2.3.4 Méthodes à base de modèles probabilistes . . . 55

2.3.5 Méthodes d'ordre fréquentielles . . . 56
2.3.6 Conclusion . . . 60
2.4 Caractérisation de blocs segmentés . . . 62
2.4.1 Caractérisation de fontes ou de style de textes . . . 62
2.4.2 Caractérisations d'illustrations . . . 65
2.4.3 Conclusion . . . 67
2.5 Comparaison de structure de documents . . . 67
2.5.1 Introduction . . . 68
2.5.2 Signature de mise en page par extraction de caractéristiques . . . 68
2.5.3 Comparaison par graphes . . . 70
2.6 Conclusion générale . . . 72

Chapitre 3 Notre approche texture pour la caractérisation du contenu 73
3.1 Objectifs . . . 73
3.1.1 Quels sont les enjeux ? . . . 73
3.1.2 Présentation de notre système de caractérisation de contenu . . . 74
3.2 Extraction d'indices textures dédiés à l'analyse d'images de documents . . . 76
3.2.1 Principe global . . . 76
3.2.2 Indices textures liés aux orientations . . . 85
3.2.3 Indices textures liés aux fréquences . . . 90
3.2.4 Conclusion . . . 92
3.3 Analyse et pertinence des données extraites pour la caractérisation des contenus . 93
3.3.1 Classification automatique du contenu . . . 94
3.3.2 Analyse factorielle des indices textures . . . 102
3.3.3 Discussions sur la pertinence des indices textures . . . 113
3.4 Conclusion générale . . . 114

Chapitre 4 Illustration de la pertinence des indices textures 117
4.1 Introduction . . . 117
4.1.1 Contexte de l'étude de la pertinence des indices de texture . . . 117
4.1.2 Principe de l'exploitation des indices de texture . . . 118
4.2 Indices de similarité proposées . . . 119
4.2.1 introduction . . . 119
4.2.2 Mesure de similarité post classification . . . 121
4.2.3 Mesure de similarité pour la comparaison pixel à pixel d'indices texture . 125
4.3 Exemples d'exploitation des indices textures proposés . . . 125
4.3.1 Comparaison de pages . . . 125

4.3.2 Comparaison d'images 130
4.3.3 Conclusion de nos expérimentations 133
4.4 Vers de véritables applications de recherche d'information par le contenu 134
4.4.1 Vers la recherche d'éléments de contenu 134
4.4.2 Vers une meilleur segmentation des images de documents 136
4.4.3 conclusion 140

Chapitre 5 Annexes **143**

Bibliographie **149**

Table des figures

1.1 Gallica : Recherche d'ouvrages par saisie de mots-clefs . . . 8
1.2 Exemple d'une illustration représentant le Christ et issue d'une requête textuelle sémantique . . . 9
1.3 Interface du logiciel développé lors du projet DEBORA [BELM00] . . . 10
1.4 Exemple de l'interface proposée par [Bén04] . . . 12
1.5 Limites des ocr lors de l'analyse de mots inconnus . . . 14
1.6 Exemple de résultats d'ocr [Bén04] . . . 15
1.7 Exemple d'une transcription de caractères d'une page avec DEBORA [BELM00] 16
1.8 Recherche d'information via Google Print [BELM00] . . . 17
1.9 Indexation de la structure de documents anciens avec le logiciel Fomuread [CR03] 18
1.10 Indexation de la structure de documents anciens avec Agora [RBD06] . . . 20
1.11 Indexation de la structure de documents anciens avec Docworks [Jou04] . . . 21
1.12 Indexation d'images de traits [UOL05] . . . 22
1.13 Indexation d'images de traits [UOL05] . . . 23

2.1 Exemples de documents structurés . . . 26
2.2 Les différentes structures d'un document . . . 28
2.3 Extrait de la variété des documents traités en analyse de documents . . . 30
2.4 Extrait de notre corpus d'images de documents anciens . . . 31
2.5 Synthèse des caractéristiques de notre corpus . . . 32
2.6 Exemple d'un scanner . . . 32
2.7 Exemple de corrections de défauts de numérisation avec[Tri03] . . . 33
2.8 Exemple de la difficulté de binariser une image de documents anciens . . . 34
2.9 Exemple de l'application d'un algorithme de composantes connexes sur des images de documents[Tri03] . . . 36
2.10 Problèmes couramment rencontrés avec l'utilisation des composantes connexes . . . 38
2.11 Limite de la segmentation avec des boîtes englobantes . . . 38
2.12 Exemple de l'application de l'algorithme RLSA sur des images de documents . . . 39
2.13 Exemple d'un découpage en XY-CUT (Pour chaque pixel, on calcule la moyenne des niveaux de gris en lignes et en colonnes) . . . 41
2.14 Exemple de construction d'un pavage de Voronoï . . . 42
2.15 Segmentation de documents par pavage de Voronoï [KIM99] . . . 43
2.16 Modification de la perception d'une image après changement de résolution . . . 44
2.17 Segmentation d'images de documents par approche psycho-visuelles [Egl98] . . . 45
2.18 Exemple d'un résultat de segmentation avec l'approche proposée par [Ros99] . . . 49
2.19 Application de projections sur des zones d'images de documents . . . 51

2.20 Segmentation texte/dessin par une méthode de projection horizontale (texte en bleu, dessin en rouge) . . . 54
2.21 Segmentation texte/dessin par une méthode de projection horizontale . . . 54
2.22 Détection de zones d'intérêt par l'utilisation d'une loi puissance [CCMV03] . . . 55
2.23 Segmentation d'images de documents manuscrits[NKPH06] . . . 56
2.24 Détection de points d'intérêts avec utilisation de coefficients d'ondelettes [Lou00] 58
2.25 Segmentation texte/dessin avec Gabor . . . 60
2.26 Segmentation texte/dessin avec Gabor . . . 61
2.27 Segmentation de documents anciens avec Gabor . . . 61
2.28 Segmentation de documents anciens avec Gabor . . . 62
2.29 [FWT98] Discrimination de différents types de textes à l'aide de l'étude de l'alignements des caractères . . . 63
2.30 [ZTW01] Discrimination de différents types de textes à l'aide de filtres de Gabor 64
2.31 Caractérisation de lettrine par la loi de Zipf [PVU+06] . . . 66
2.32 Caractérisation de lettrine par approche texture [PVU+06] . . . 67
2.33 Construction d'un graphe modélisant la structure d'un formulaire [DA02] . . . 70
2.34 Construction et réduction d'un graphe modélisant la structure d'une page [BMS03] 71

3.1 Présentation de notre système . . . 76
3.2 Exemples de la capacité de la fonction d'autocorrélation à faire ressortir les orientations principales . . . 78
3.3 Construction de la rose pour un angle θ_i . . . 79
3.4 Exemples de roses . . . 80
3.5 Exemples du comportement de la rose sur une image de document . . . 81
3.6 Exemple de roses des directions . . . 81
3.7 Exemples du comportement de la rose sur une image de document bruitée . . . 82
3.8 Exemples du comportement de la rose sur une image de document bruitée . . . 82
3.9 Difficulté de choisir la taille d'une fenêtre d'analyse . . . 83
3.10 Importance d'un calcul à différentes résolutions . . . 84
3.11 Dualité entre changement de taille de fenêtre ou de taille d'image pour un calcul multirésolution . . . 85
3.12 Détection de l'orientation principale d'une image à l'aide de la rose des directions 86
3.13 Analyse des orientations à faible échelle . . . 86
3.14 Roses calculées sur des illustrations . . . 87
3.15 Lien entre l'intensité de la réponse d'autocorrélation et l'isotropie/anisotropie des directions. Colonne de gauche : image d'origine, Colonne de droite : Autocorrélation de l'image . . . 88
3.16 Comportement de l'indice **Eq. 3.4** sur des textures orientées . . . 89
3.17 Comportement de l'indice **Eq. 3.5** sur des textures composées de zones isotropiques 89
3.18 Comportement de l'indice **Eq. 3.6** sur des textures composées . . . 90
3.19 Comportement de l'indice **Eq. 3.7** sur des textures composées de fréquences de transitions différentes . . . 92
3.20 Exemple de l'indice récursif **Eq. 3.8**, calculé sur deux images différentes. . . . 93
3.21 Classification de pixels avec différents algorithmes . . . 96
3.22 Classification à 3 classes de pixels avec Clara . . . 97
3.23 Classification de pixels de documents contemporains (images issues de [MD05, KIM99]) . . . 97
3.24 Classification de pixels d'un ouvrage avec Clara . . . 99

3.25 Logiciel de saisie terrain . . . 100
3.26 Classification de pixels d'images naturelles . . . 101
3.27 Résumé des résultats d'un ACP sur une image de document . . . 103
3.28 Extrait du corpus de test . . . 104
3.29 Comportement des données . . . 105
3.30 Exemple d'une mauvaise projection d'un espace n à p avec n>p. . . . 106
3.31 Décorrélation des indices multirésolution. . . . 107
3.32 Comparaison entre une classification page par page et une classification d'un ouvrage complet . . . 108
3.33 Comparaison entre une classification page par page et une classification d'un ouvrage complet . . . 109
3.34 Comportement des caractéristiques lorsqu'on classifie un ouvrage complet . . . 110
3.35 Illustration de la capacité à réduire les données après la réalisation d'une ACP sur un ouvrage complet . . . 111
3.36 Comportement des données selon la taille des images . . . 112
3.37 Comportement des données selon la taille des images (cas d'une grande image) . . . 112
3.38 Comportement des données selon la taille des images (cas d'une petite image) . . . 113
3.39 Comportement des données selon la taille de l'échantillon . . . 114

4.1 Principe de la comparaison de partitions . . . 122
4.2 Principe du calcul des indices a et b . . . 124
4.3 Exemples de résultats de requêtes utilisant l'indice de comparaison de partitions modifié (R') . . . 127
4.4 Exemples de résultats de requêtes utilisant l'indice de comparaison de partitions modifié (R') . . . 127
4.5 Comparaison des résultats obtenus de deux manières différentes . . . 128
4.6 Extraits des illustrations composant la base d'images testée . . . 131
4.7 Recherche dans un ouvrage . . . 132
4.8 Principe de la recherche d'éléments de contenu . . . 135
4.9 Segmentation combinant les informations composantes connexes et indices textures 137
4.10 Exemple de labellisations de composantes de dessins de traits . . . 139

5.1 Extrait des images de la base utilisée pour la comparaison de pages . . . 143
5.2 Extrait des images de la base utilisée pour la comparaison de pages . . . 144
5.3 Classification à 4 classes sur des images de documents contemporains . . . 145
5.4 Classification à 5 classes sur des images de documents contemporains . . . 146
5.5 Classification à 3 classes sur des images de documents anciens . . . 147
5.6 Classification à 4 classes sur des images de documents anciens . . . 147

Introduction générale

Inventée au milieu du XVème siècle par l'allemand Gutenberg, l'imprimerie moderne marque le début d'une nouvelle ère dans la production et la diffusion d'ouvrages. Jusque là, la réalisation de livres était l'oeuvre de moines copistes qui, armés de leur patience et de leur aptitude à la technique de la calligraphie, mettaient parfois plusieurs mois à recopier un livre. Avec l'invention de l'imprimerie, les éditeurs se sont mis à produire des livres en très grandes quantités. Six siècles plus tard, la masse de ces ouvrages stockés dans de nombreuses bibliothèques européennes représente une richesse culturelle et scientifique inestimable. Ces livres sont depuis de nombreuses années, le support de travaux de recherche en sciences humaines et sociales. Ainsi, l'étude de ces ouvrages permet la réalisation d'avancées scientifiques sur des sujets relatifs à l'évolution de la langue française, l'influence de la culture française dans les autres pays, l'évolution des techniques d'imprimeries...
Le passage à l'ère du numérique a marqué un tournant dans l'exploitation de ces documents. En effet, en dix ans à peine, l'établissement de nombreuses campagnes de numérisations de ces fonds patrimoniaux, combinées à la démocratisation de l'internet, ont permis de mettre en place de nombreuses bibliothèques numériques. Ces bibliothèques d'un nouveau genre garantissent un accès généralisé à des fonds, qui jusque là n'étaient accessibles qu'à un nombre réduit de personnes.
Cette "révolution" a très vite soulevé enthousiasme et préoccupation. Dans leurs articles respectifs, les auteurs de [LC06, LKPJ06] posent, entre autres, plusieurs questions ouvertes liées aux problèmes d'accessibilité et de visibilité inhérentes à la mise en place de ces bibliothèques numériques. Ainsi, les auteurs soulèvent des questions aussi diverses que variées, ayant trait au format des images, à la saisie et l'exploitation de métadonnées, aux attentes des chercheurs, à la propriété intellectuelle, à la mutualisation et la capitalisation des données...

La numérisation massive de milliers de pages de documents, mais surtout le désir de les rendre disponibles au plus grand nombre d'entre nous, nécessite la mise en place d'outils informatiques permettant un accès rapide et pertinent à l'information qui y est contenue. Plusieurs années de travaux scientifiques en analyse d'images de documents contemporains, ont d'ores et déjà permis la réalisation d'outils performants, permettant plusieurs formes d'indexation par analyse du contenu. On pense bien entendu aux logiciels d'OCR permettant d'accéder au sens des mots du texte. On peut également citer les outils de rétro-conversion permettant un accès à la structure ou encore ceux permettant d'indexer les illustrations et les photos qui composent les pages. Cependant, ce cadre nouveau que représente l'analyse d'images de documents anciens, limite la simple transposition de ces outils initialement dédiés aux documents contemporains. L'explication se trouve principalement dans la nature même du corpus d'images traitées. En effet, l'hétérogénéité des pages composant ce corpus, la taille des bases mises en place, la dégradation de certains documents sont quelques exemples reflétant la spécificité et les enjeux scientifiques à relever. En terme d'usages, la variété des attentes exprimées par les utilisateurs témoignent également de la nécessité que représente la mise en place d'une nouvelle réflexion et de la création de nouveaux outils de traitements d'images dédiés aux documents anciens.

Mes travaux de thèse, ont été menés en association avec l'action concertée initiative masse de données (ACI MADONNE[1]). Cette ACI, fruit de la collaboration de plusieurs laboratoires informatiques français, a eu entre 2003 et 2006 pour objectif de contribuer à l'étude et à la réalisation de systèmes d'indexation d'images de documents patrimoniaux. De ce travail commun,

[1] sources des informations :$http://l3iexp.univ-lr.fr/madonne/index.html$

sont nés différents systèmes d'indexation, répondant à un large spectre d'usages.
Ma contribution personnelle répond précisément à une problématique fondamentale de caractérisation de contenu des images de documents anciens et propose ,à travers une expérimentation, des pistes nouvelles pour la mise en place d'outils d'aide à l'indexation des données et à la navigation au sein d'un corpus d'images.

La motivation ces travaux de thèse prend son origine dans la nécessité à trouver une alternative aux méthodes d'analyse de documents contemporains basées principalement sur une segmentation des pages nécessitant la plupart du temps une interprétation de leur structure. En effet, le constat très récent que les approches développées jusqu'ici en analyse des documents s'avèrent insuffisantes lorsqu'elles sont étendues à des domaines d'application plus ouverts, tels que les corpus anciens de patrimoine, montre le besoin urgent de traiter et d'indexer. Tout l'enjeu de nos travaux est donc de montrer qu'il est possible de caractériser le contenu des images de documents patrimoniaux, tout en tenant compte de ses spécificités (forte hétérogénéité du contenu, bases de taille conséquente...) et sans passer par les étapes de segmentation des pages en blocs et l'interprétation de leurs structures. Ainsi, il faut repenser les méthodologies en relation avec l'arrivée rapide de nouveaux contenus très hétérogènes et très difficiles à catégoriser comme les collections imprimées de la Renaissance.

Pour apporter des éléments de réponses à certains de ces nouveaux besoins, ce mémoire est organisé autour de la rédaction de 4 chapitres :

1. Dans le premier chapitre, nous proposons une synthèse sur le fonctionnement et les usages relatifs aux bibliothèques numériques actuellement en ligne. Ce tour d'horizon permet de se faire une idée sur les outils disponibles et permettant un accès au contenu des images. Nous verrons d'une part, que le principal mode d'accès proposé actuellement se limite à rendre disponibles les images numérisées des pages, et d'autre part, que les avancées scientifiques traitent principalement de la production (manuelle ou non) de métadonnées textuelles et de la transcription du texte. En marge de ces thématiques de recherche, nous détaillerons également les attentes des usagers pour d'autres formes d'indexation de l'image et qui sont toutes aussi importantes. Il existe plusieurs catégories d'informations pour lesquelles la seule indexation par mots-clefs n'est pas suffisante. En effet, si un utilisateur souhaite, par exemple, comparer des illustrations (lettrines, enluminures...), indexer la structure (retrouver une information spécifique dans un formulaire, analyser un sommaire) ou encore comparer des mises en pages, seule une analyse fine de l'image peut permettre d'aboutir à un tel objectif. Ce premier chapitre met en évidence la nécessité de fournir des outils de traitement d'images résolument innovants pour caractériser le contenu de documents anciens dans les mises en pages ne sont pas toujours prévisibles et régulières.
2. Le deuxième chapitre propose un état de l'art sur les méthodes de caractérisation d'images de documents et oriente les discussions autour de l'analyse des structures (mises en pages) qui a toujours été considérée comme le point de départ incontournable des systèmes de reconnaissance des contenus et d'indexation. A travers l'étude de nombreuses références, nous mettrons en avant toutes les difficultés d'utilisation ou d'adaptation des méthodes classiques de la littérature sur des images de notre corpus. En nous appuyant sur des tests que nous avons réalisés, nous verrons notamment que certains outils (initialement dédiés aux documents contemporains) sont complexes d'utilisation sur des documents dont les contenus sont riches et variés. Les conclusions de ce chapitre mettent en avant l'intérêt de la mise en place d'outils de traitement d'images permettant une catégorisation robuste d'un tel corpus, et cela sans segmenter ou retrouver la structure du document.

3. Le troisième chapitre détaille notre contribution à l'analyse d'images de documents. Ce chapitre décrit comment, à l'aide du calcul de nouveaux indices textures dédiés aux documents anciens, il est possible de caractériser le contenu des images sans émettre d'hypothèses, ni sur la structure ni, sur les caractéristiques des images traitées (nature des contenus, origine des pages,...). Ces indices permettent d'exprimer toute la richesse des informations d'orientations et de fréquences des motifs présents dans les images. Ce chapitre se termine par une analyse des données obtenues grâce au calcul des indices textures. Cette analyse permet non seulement de valider (ou d'invalider) la pertinence des indices textures, mais également d'évaluer la robustesse de ces derniers selon le contenu et les caractéristiques des images analysées.
4. Dans le dernier chapitre, nous montrons comment notre contribution au domaine de la caractérisation de textures peut être exploitée à des fins d'indexation par le contenu. N'ayant pas de contraintes spécifiques de réalisations d'applications, nous avons préféré réaliser des expérimentations ayant pour principal intérêt de mettre en avant la pertinence de ces indices et les avancées qu'ils représentent en terme de caractérisation de contenu d'un corpus fortement hétérogène.

Ce mémoire propose donc une réflexion et une avancée sur un domaine encore peu traité dans la littérature qu'est la caractérisation d'images de documents anciens pour l'indexation et la recherche par le contenu.

Chapitre 1

Bibliothèques numériques et modalités d'accès au contenu

1.1 Introduction

L'idée de mener des campagnes de numérisation de documents a commencé à émerger dans les années 60. En effet, l'essor de "l'outil informatique" permettait d'entrevoir des perspectives attrayantes : sauvegarder les documents, constituer des bases d'images et de textes communes à plusieurs bibliothèques, concevoir des outils d'aide à la manipulation de données textuelles... Les premières campagnes de numérisation significatives, datent des années 70 (USA, France,...). Si l'objectif premier, était avant tout, de sauvegarder le patrimoine (dont une partie commençait à souffrir des outrages du temps) et de confectionner des catalogues d'images consultables, on était encore loin de la notion de bibliothèque numérique. Il faut attendre que les choses évoluent tant du côté des performances techniques, que du côté des décisions politiques pour voir, en 1993, la mise en place de la première bibliothèque numérique française. C'est dans un but de "développement et de promotion des supports numériques permettant la libre manipulation de l'information " que l'ABU (Association des Bibliophiles Universels) a mis en place cette bibliothèque d'un nouveau genre. L'ABU sera suivie 4 ans plus tard par la Bibliothèque Nationale de France, qui, via le projet Gallica est à ce jour la bibliothèque numérique française la plus importante avec plus de 80000 documents accessibles en ligne. Le succès de Gallica est la preuve de l'intérêt de ces campagnes de numérisations puisque quotidiennement plus de 20.000 connexions sont référencées et qu'en 2002 plus d'1,4 TO d'images a été téléchargé (source : [Jou04]).

La création de ces volumineux espaces numériques conduit très vite à la question suivante : à qui sont destinées ces bibliothèques et quels sont les besoins des différents acteurs visés par ces bibliothèques mises en ligne ?

D'après [Kal00], on peut distinguer 3 acteurs différents :

- Le premier est l'utilisateur général d'une bibliothèque qui souhaite examiner des sources manuscrites ou imprimés anciens.
- Les étudiant spécialisés en histoire des textes : philologues ou éditeurs critiques de travaux classiques ou médiévaux qui utilisent différents types de support : papier, papyrus, pierre. Ceci inclut, des étudiants en textes anciens comme les papyrologues (spécialistes dans l'étude des papyrus), les épigraphistes (spécialistes de l'étude scientifique des inscriptions - appelées Incipit - placées en tête d'un livre, d'un chapitre), les paléographes (spécialistes

en science des écritures anciennes), et les codicologues (spécialistes étudiant le support des manuscrits).
- Les chercheurs qui mènent des études de philologie (étude historique d'une langue par l'analyse critique des textes) ou d'histoire en général.

A ces 3 acteurs on pourrait également ajouter les "producteurs de données" c'est à dire les conservateurs, les experts, les informaticiens, soient tous ceux qui à un moment ou un autre viennent apporter leur connaissance pour l'enrichissement ou l'établissement de solutions informatisées qui permettent d'avoir accès à la version numérique des documents (production de métadonnées, structuration de l'information, création d'ontologies...).

Avec une telle variété d'acteurs potentiellement intéressés, on imagine vite à quel point les attentes des uns et des autres vont être diamétralement opposées. Les questions suivantes semblent donc légitimes : à quels besoins spécifiques doivent répondre les bibliothèques numériques ? Quels sont les outils nécessaires à leur création et leur maintenance ?

A posteriori, il est possible d'apporter des éléments de réponses en s'appuyant sur l'étude des services offerts par les bibliothèques actuellement accessibles sur internet. Globalement, ce qui est proposé est un accès au catalogue des images numérisées ; accès rendu possible grâce à la mise en place d'outils de recherche fonctionnant par mots-clefs (un peu à l'image des moteurs de recherche que l'on trouve sur internet). Nous le verrons dans la suite de ce chapitre, les solutions commerciales de digitalisation d'ouvrages anciens semblent avoir atteint une productivité et une qualité acceptable pour la consultation en ligne. De ce fait il est relativement simple de mettre en place une politique de numérisation permettant non seulement de préserver le patrimoine, mais aussi de rendre consultable une restitution fidèle des livres d'origine et de les mettre à disposition d'un large public (accès qui jusque là n'était réservé qu'aux experts du patrimoine ancien).

La mise en place et l'accès à ces bibliothèques numériques permet d'introduire deux notions importantes que sont l'indexation des images et la notion d'aide à la navigation. Ces deux concepts sont complémentaires. L'indexation consiste à trouver un moyen d'associer aux images une information pertinente (métadonnées sur l'ouvrage, index de mots, index d'illustrations,....). La notion d'aide à la navigation, touche au problème de l'accessibilité d'une information recherchée dans une masse de données de taille conséquente. La phase d'indexation doit permettre d'extraire une grosse quantité d'informations qu'il faut analyser et structurer, pour permettre in fine l'accès à ce que recherche un utilisateur.
Il existe deux manières d'appréhender l'indexation. La question est avant tout de savoir comment extraire de l'information des images ? Certaines bibliothèques numériques ont fait le choix d'indexer manuellement leurs images alors que d'autres ont fait le pari de se lancer dans des solutions incluant de l'indexation automatique. Cette deuxième solution reste néanmoins marginale.

La suite de ce chapitre s'articule sur la présentation de ces deux possibilités d'indexation. Au travers d'une présentation des bibliothèques numériques du Centre d'Etudes Supérieurs de la Renaissance de Tours et de la Bibliothèque Nationale de France, nous verrons à quoi peut ressembler un système d'information structuré dédié à l'archivage et l'accès à ce patrimoine littéraire unique. Nous aborderons aussi les difficultés et les limites induites par le choix d'une indexation manuelle de ces images.
Nous verrons, dans une deuxième partie, qu'il est possible d'extraire automatiquement une grande quantité d'informations d'une image. Lorsque la nature de l'information que l'on souhaite extraire n'est pas directement accessible via des descripteurs textuels, il est ainsi possible d'analyser l'image du document. Son extraction rend possible la mise en place de nouveaux type d'outils de recherche d'informations n'impliquant aucune hypothèse a priori sur les futures requêtes des

usagers. Ces outils n'ont pas pour objectif de remplacer ceux permettant la saisie d'annotations textuelles, ils apportent avant tout une réelle complémentarité aux outils existants.

1.2 Bibliothèques numériques indexées manuellement

1.2.1 Quelques normes existantes

Il est impossible de lister l'ensemble des bibliothèques numériques actuellement en ligne. Rien qu'en France, le ministère de la culture en recense plus de 300 en 2006 [2]. Après un rapide panorama des bibliothèques numériques référencées sur le site du ministère, il semble ressortir que la majorité des bibliothèques numériques en ligne ont opté pour une offre principalement axée sur la recherche par mots-clefs associée à une présentation des résultats sous la forme des images de pages numérisées. Ce type de fonctionnement est notamment celui adopté par la Bibliothèque Nationale de France via son portail Gallica[3]. La Figure **Fig. 1.1**, illustre ce mode de recherche par mots-clefs. Pour accéder aux images de documents que l'on recherche, il faut pour cela remplir un formulaire à l'aide d'un ou plusieurs mots-clefs saisis dans les champs appropriés (point A). Il faut également spécifier le type de documents que l'on recherche (point B). Enfin, le système d'indexation propose des images sensées correspondre à la requête (point C).

Ce type de mode de recherche est rendu possible par une phase préalable de productions manuelles de métadonnées permettant de décrire les ouvrages de la base. Ainsi, une grande partie des instituts qui mettent en place un tel accès opèrent via la saisie manuelle d'informations. Concrètement, lorsqu'un nouvel ouvrage est numérisé, une personne renseigne toute une liste de champs prédéfinis. Pour la production de métadonnées sur les ouvrages on retrouve généralement les champs suivants : auteurs, années d'impression, imprimeurs...
A l'heure actuelle, il n'existe pas d'ontologie du domaine concernant la classe des documents anciens qui soit adoptée par l'ensemble des bibliothèques. L'explication de ce vide de définitions communes se trouve probablement dans la diversité des documents accessibles (ouvrages littéraires, cartes, archives militaires...) mais aussi dans la diversité des usages. On retrouve néanmoins plusieurs initiatives relatives à ce problème complexe. Parmi celles-ci, on peut citer l'initiative pour les archives ouvertes (OAI) qui s'intéresse à l'ensemble des activités liées à l'archivage, l'échange et la valorisation d'archives numériques. Cette initiative est, en fait, un protocole de collecte de métadonnées au format XML respectant le format Dublin Core et visant à permettre l'inter-opérabilité entre ces archives ([OR03] détaille certaines de ces normes). L'objectif est de construire des métadonnées décrivant les ressources disponibles et de créer ainsi un lien entre toutes les bibliothèques numériques adoptant cette norme. Il existe une variété riche de documents déjà décrits avec l'OAI. La Bibliothèque Nationale de France a décidé de se lancer, elle aussi, dans l'utilisation de l'OAI. L'auteur de [PM05] retrace l'évolution de la constitution de Gallica et indique que *"la BNF s'achemine vers une solution OAI. Ce protocole permettra à terme à toute bibliothèque équipée de capturer des notices des ouvrages de Gallica qui l'intéressent pour la cohésion de son corpus sans doublonner la numérisation, et en renvoyant immédiatement le lecteur vers cette numérisation. Aujourd'hui, les premières notices de monographies simples de Gallica sont versées en Dublin Core afin que Gallica soit diffuseur de données"*.
Il est également important de citer la Text Encoding Initiative (TEI) qui est décrite dans [IV96] et qui a pour objectif *"la mise au point d'un ensemble de normes pour la préparation et l'échange*

[2] http ://www.culture.gouv.fr/
[3] http ://gallica.bnf.fr/

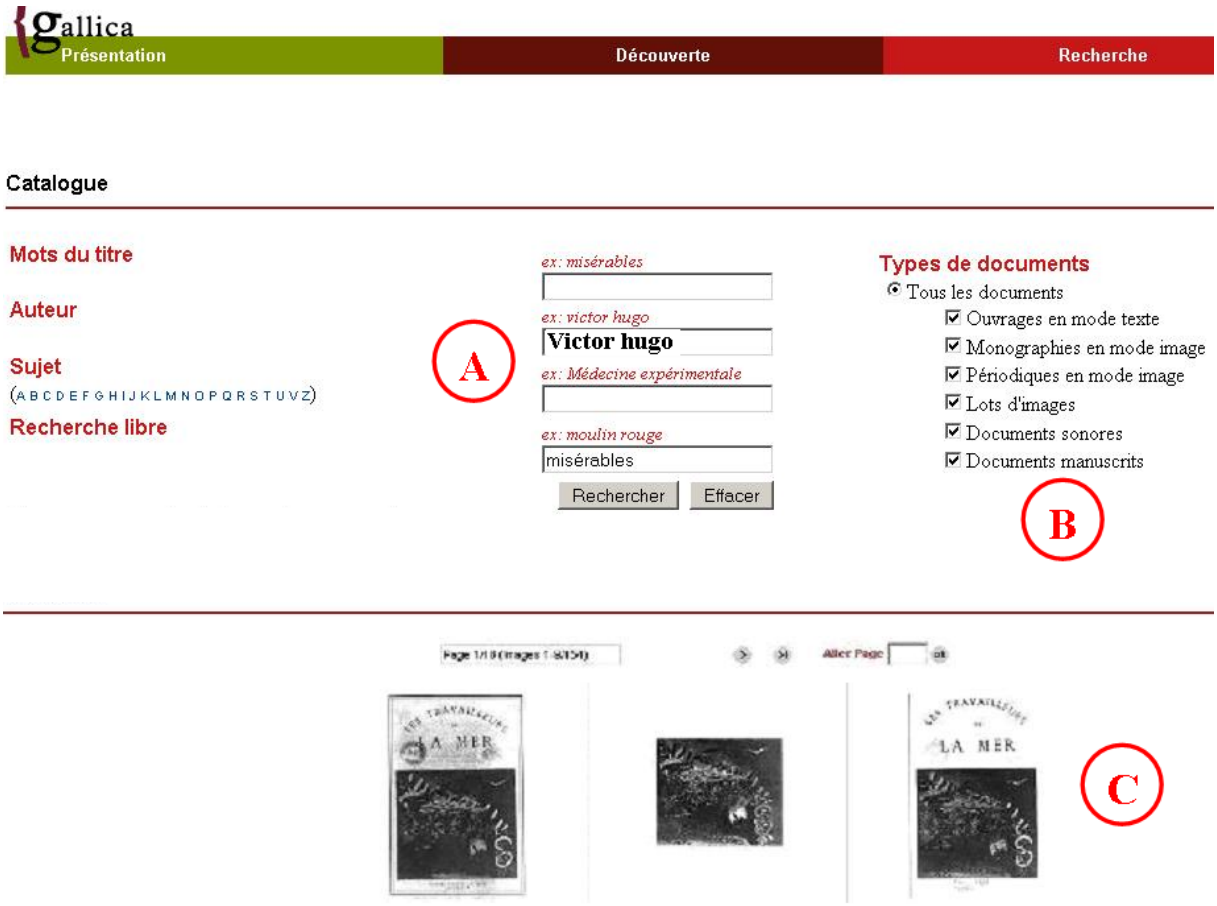


FIG. *1.1: Gallica : Recherche d'ouvrages par saisie de mots-clefs*

de textes électroniques". A l'origine développée en SGML, cette initiative a évolué vers le XML pour proposer un ensemble de balises permettant de structurer le texte, mais aussi les images afin de permettre une transcription de textes anciens qui soit la plus fidèle possible à l'originale. La TEI se présente sous la forme d'une DTD (permet de définir un modèle de document) dont l'utilisation de balises prédéfinies permet bien sûr de reproduire toutes les spécificités des textes anciens étudiés. Elle permet également de "décrire" des entités telles que les illustrations ou même de renseigner des informations qui serviront de métadonnées (auteurs, dates d'impression, informations relatives au fichier...). Dans la famille des langages de balisage orienté sémantique, on peut citer DockBook[4] et Encoded Archival Description Tag Library (EAD [Com98]) qui comme TEI sont des DTD permettant de structurer (et d'annoter) les ouvrages transcrits. Nous aurons l'occasion de revenir sur ces propositions puiqu'il se trouve que le travail de recherche relatif aux illustrations des documents anciens est un axe de recherche important. Afin de permettre une annotation de ces bases iconographiques et comme pour le texte dans l'optique de pouvoir effectuer des recherches sur ces illustrations, le projet Iconclass[5] propose une classification internationale de ce type d'images en recensant plus de 28000 définitions, et 40000 références à des

[4] http ://www.docbook.org/
[5] http ://www.iconclass.nl

livres du domaine iconographique. Iconclass a la particularité d'être un thésaurus arborescent. Ce choix de structuration, permet de chercher des mots-clefs de thésaurus bien spécifiques et aussi des images en fonction du contexte ce qui permet notamment d'éviter les problèmes liés aux homonymes et au contexte. Pour l'exemple, imaginons que l'on désire annoter une illustration représentant le Christ enfant (Figure **Fig. 1.2**). Après avoir parcouru successivement les rubriques Religion et magie/Religion chrétienne/le Christ/le Christ comme enfant ou jeune homme (en général)/l'enfant Jésus en compagnie d'autres, on a accès aux mots-clefs pouvant servir à décrire cette illustration : supra-naturelle,religion chrétienne, religion, Christ, enfant Jésus, groupe. Sans ce choix de modélisation, il serait extrêmement complexe d'aboutir aussi précisément aux images recherchées.



FIG. *1.2: Exemple d'une illustration représentant le Christ et issue d'une requête textuelle sémantique*

1.2.2 Mise en place de systèmes d'indexation manuelle

Les premiers besoins ont fait ressortir, de l'attente des usagers, la nécessité d'indexer manuellement les versions numériques des archives. D'après [BH01] la possibilité d'annoter les ouvrages arrive en deuxième position (avec 12,9%) des fonctionnalités désirées par les utilisateurs et cela juste derrière la possibilité de la recherche d'occurrence de mots (27,7%). En France, plusieurs projets d'envergure (BAMBI, DEBORA, Philectre, METAe, DMOS, Agora...) ont permi d'aboutir à la conception de plates-formes de navigation et d'aide à l'indexation. Parmi l'ensemble des fonctionnalités proposées, on retrouve à chaque fois un module de visualisation des ouvrages mettant à disposition toute une batterie d'outils tels que le zoom, le visionnage d'un ouvrage sous forme de vignettes, visualisation de la structure de l'ouvrage... En ce qui concerne la partie annotation, on retrouve des modules permettant de décrire les images traitées. Il est ainsi possible d'annoter un ouvrage (auteur, date de publication...). Ce sont ces types de logiciels qui permettent de mettre en place l'accès à une bibliothèque comme Gallica. En parallèle, il est possible de décrire les pages d'un ouvrage (poser des commentaires, annotation de zones précises de l'image...). Selon l'application, le choix des annotations peut être laissé au libre arbitre de la personne qui indexe les images ou bien alors peut obéir à un thésaurus ou une norme présentée précédemment. Ces annotations n'ont pas forcément pour objectif de servir à une indexation future du corpus. Elles permettent à un utilisateur quelconque de travailler sur la version numé-

rique un peu comme s'il travaillait sur un document papier (notion de prises de notes).
La Figure **Fig. 1.3** est un extrait de l'interface du logiciel développé lors du projet DEBORA [BELM00]. On y retrouve la plupart des fonctionnalités décrites précédemment. Le point A montre que l'on peut accéder à la structure de l'ouvrage de manière simplifiée. Le point B est un extrait du module de saisie d'annotations. Le point C illustre le module de recherche que propose le logiciel DEBORA.
Toutes les métadonnées extraites (automatiquement ou non) sont sérialisées sous un format propriétaire (format AdHoc). Ce dernier est, dans sa philosophie, proche du format XML. Il permet de mémoriser la structure de l'ouvrage et son contenu compressé en utilisant un système de balises qui répertorient les métadonnées.

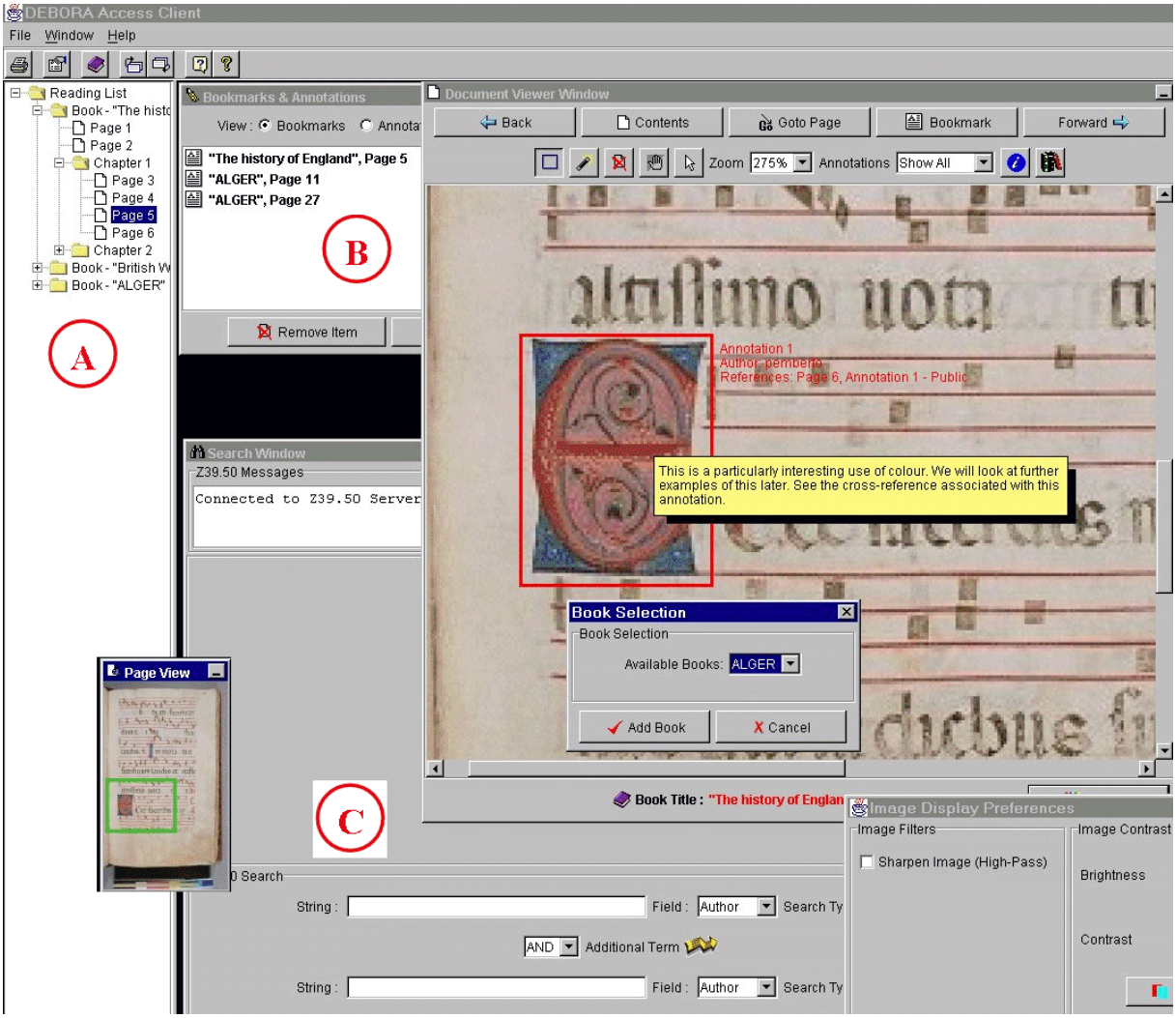


FIG. *1.3: Interface du logiciel développé lors du projet DEBORA [BELM00]*

Dans le cadre de son projet de mise en place des Bibliothèques Virtuelles Humanistes qui a pour objectif de diffuser son corpus d'ouvrages numérisés, le CESR de Tours a mis en place une interface basée sur un moteur de recherche par mots-clefs. Pour permettre ce type de recherches, le CESR a édité son propre thésaurus contrôlé par une base de données relationnelle. L'auteur

de [BUS06] décrit ce système qui s'articule autour de 4 tables renseignées soit manuellement soit automatiquement. Ces tables permettent d'effectuer des recherches sur les titres, auteurs, dates... Les requêtes peuvent aussi concerner le contenu des images, ce qui offre la possibilité d'une recherche plus fine et plus précise que lorsque les métadonnées décrivent uniquement les ouvrages. Il est par exemple possible d'indiquer le type d'ornements présents (lettrine, bandeau, fleuron...) ainsi que la sémantique associée (Lettre A, crâne, main...).

Même s'il n'est pas possible d'avoir accès aux structures des autres bibliothèques numériques disponibles sur la toile, une brève consultation de ces sites permet de supposer que la majorité des bibliothèques reposent sur des bases de données utilisant des requêtes très simples de recherche d'occurrences de mots. Cela se traduit par la seule possibilité d'effectuer des associations logiques de type "et/ou" entre mots-clefs.

Afin de permettre la mise en place de systèmes plus élaborés que les systèmes traditionnels, certains chercheurs ont proposé des solutions plus élaborées. On peut retrouver un état de l'art sur les logiciels de gestion et d'annotation de corpus dans [FB05].
La phrase de [SLYcC02] résume l'un des problèmes couramment rencontrés avec ce type de systèmes : *"le simple fait qu'un utilisateur ne puisse pas avoir l'ensemble des connaissances (historiques, scientifiques...) l'empêche de choisir les mots-clefs appropriés indispensables au bon fonctionnement du système"*. L'auteur prend l'exemple d'un utilisateur lambda cherchant l'image du chancelier de la dynastie Quin en histoire chinoise ancienne mais dont il ne se souvient pas du nom. Si le système est en full-text, cet utilisateur a de grandes chances de ne pas pouvoir retrouver l'illustration qu'il cherche. Sur ce constat, [SLYcC02] propose un système de requêtes par mots-clefs mais qui, cette fois-ci, se base sur l'analyse d'ontologies et de thésaurus partagés. Par exemple si l'utilisateur spécifie qu'il recherche "le général en armure de la dynastie Quin", la première étape va être de parser cette requête afin de permettre d'en ressortir l'information importante (par exemple dynastie Quin : Valeur, Porte : Propriété, Armure : Valeur, Général : sujet). La deuxième va être de comparer le résultat du parsing avec les annotations des images de la base ; annotations issues d'un ensemble d'ontologies et de thésaurus choisi à la mise en place du système.

Dans [Bén04, ABCI04, IBCH05] les auteurs expliquent toute la difficulté que représente la constitution d'un logiciel de gestion de corpus avec entre autres, l'étape de l'annotation et de la recherche d'information. En partant du principe qu'il est impossible de décrire objectivement une image, les auteurs proposent un modèle à base de points de vues au lieu du classique système à base de connaissances. Ainsi, plusieurs utilisateurs peuvent venir enrichir les descripteurs des images à l'aide de mots-clefs. Cette liberté peut engendrer des difficultés lorsque des images sont décrites de manières différentes voire même contradictoires (l'auteur prend l'exemple d'une mosaïque qu'une personne voit constituée de carreaux blancs sur un fond noir alors qu'une deuxième la voit faîte de carreaux noirs sur un fond blanc). Dans [Bén04], l'auteur montre qu'il est possible de construire des graphes orientés acycliques symbolisant les liens entre les points de vue des descripteurs sur une même image. L'originalité de ces travaux tient au fait que l'accession à l'information ne se fait pas à l'aide de requêtes mais en navigant dans les graphes créés. Pour gérer la subjectivité des experts, l'auteur propose un mécanisme de filtrage de graphes qui permet selon lui de "trouver des rapports entre descripteurs, non-dits au niveau des modèles, mais apparaissant dans leurs usages". Ce rapport se fait via une analyse des intersections entre les annotations de chaque utilisateur. La figure **Fig. 1.4** illustre un cas ou une image est annotée par 4 personnes. Les filtres permettent de faire apparaître les informations sensées aider un utilisateur à trouver l'information qu'il cherche.

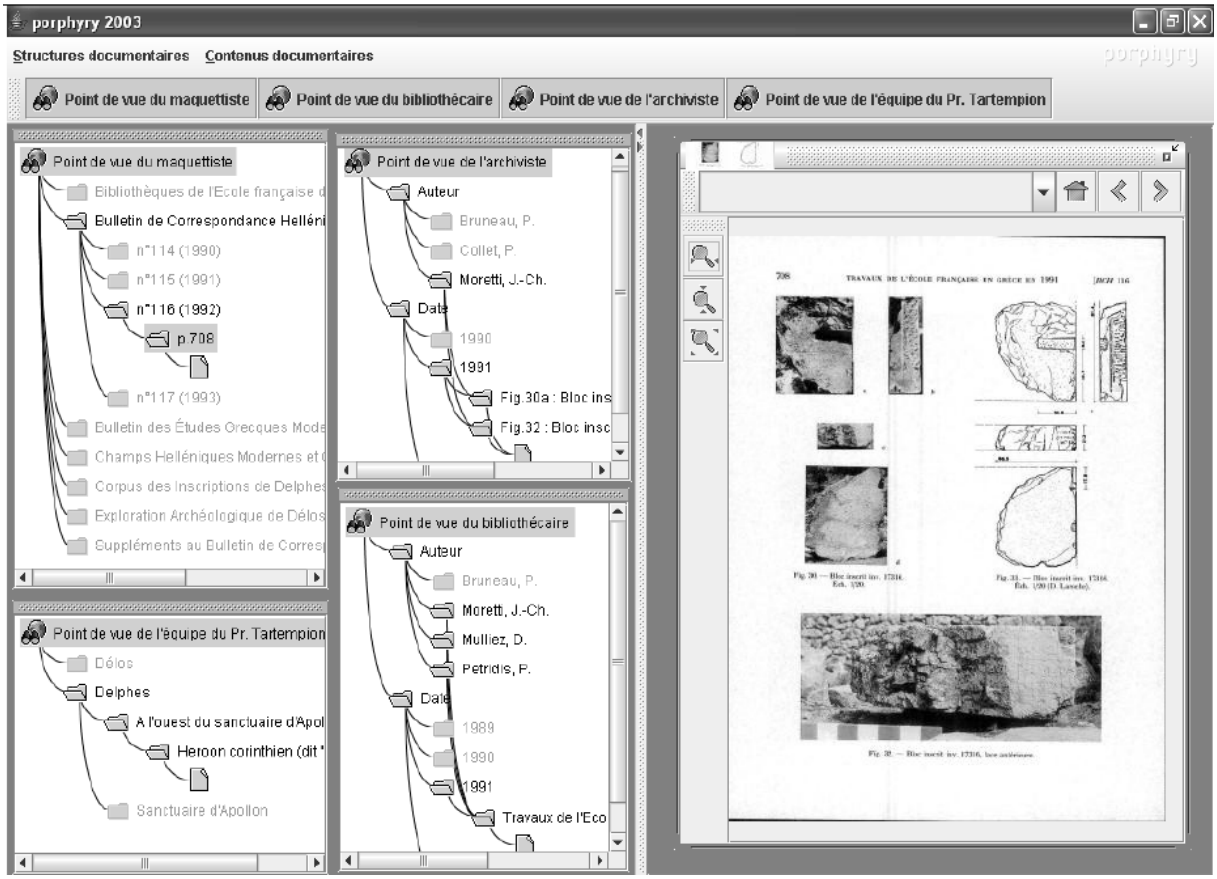


FIG. *1.4: Exemple de l'interface proposée par [Bén04]*

Le logiciel TreSy est une interface de recherche full texte développée par l'Ecole Normale Supérieure de Pise ([Bor01]) . Si cette interface n'est pas spécifiquement dédiée aux documents anciens, elle a la particularité d'être adaptée à la recherche d'informations dans un gros volume de données ; ce qui correspond parfaitement à la spécificité des corpus de documents anciens. Ainsi, dans le cadre du projet METAe, TreSy s'appuie sur les métadonnées extraites et leur stockage en XML pour proposer des outils puissants de recherche par mots-clefs, par caractère, en éditant des contraintes sur le contexte des mots...

1.2.3 Conclusion

Cette présentation des logiciels permettant la saisie de mots-clefs, a permis de mettre en évidence les avantages et les inconvénients inhérent au choix de l'indexation manuelle du contenu. Tout d'abord, il faut bien insister sur le fait que cette phase est indispensable. La raison principale est que certaines informations sont trop complexes à extraire automatiquement (nom des auteurs, mon du traducteur, année de publication...). Ce choix d'indexation possède également l'avantage de permettre l'acquisition de métadonnées sémantiquement de haut niveau. Enfin, le fait de décrire l'information par des mots-clefs, permet un stockage peu volumineux et un traitement systématique et automatisé de ces informations.
En ce qui concerne les limites de l'indexation manuelle, la subjectivité inhérente à la saisie de

mots-clefs (pour l'annotation ou la formulation de requêtes) est un problème clairement identifié par la communauté. La lourdeur et la masse de travail que représente la saisie manuelle de mots-clefs est également un frein important à l'indexation de très grandes bases d'images.

1.3 Vers une indexation automatisée du contenu par analyse d'image

Dans la section qui suit, nous allons détailler les différentes catégories de méthodes existantes permettant d'indexer automatiquement le texte, la structure ou encore les illustrations.

1.3.1 Les OCR et leurs limites

Nous l'avons évoqué précédemment, le souhait principal des utilisateurs de bibliohtèques numériques de pouvoir accéder au sens même des mots contenus dans les pages. Cette réalisation donnerait l'opportunité de pouvoir recherche rapidement des occurences de mots directement dans le texte. Cela permettrait également de pouvoir utiliser des outils de traitement du langage naturel (résumés automatiques, identification de noms propres...), ce qui dans ce cas pourrait donner plus de force au texte.
Etant donné que l'idée d'une transcription manuelle des millions de pages déjà numérisées à ce jour n'est pas envisageable, la communauté informatique et industrielle s'est lancée dans la conception de nouveaux outils dédiés à la transcription automatique de texte. Sur ce sujet, le principal effort des divers acteurs (recherches publique et privée) se situe essentiellement au niveau de la conception de logiciels d'OCR pour documents anciens. En effet, les OCRs sont traditionnellement utilisés pour la transcription automatique de documents contemporains et ne sont pas adaptés aux images de documents anciens.
Si les OCR actuels affichent un taux de reconnaissance de 99% (soit une erreur toutes les deux lignes) ils sont significativement inefficaces sur des documents anciens composés de termes, de rêgles de mise en page (césure, alignement...) et de polices disparues à ce jour. En effet, avec un taux moyen de reconnaissance de 70% à 75%, on peut conclure à une inefficacité des OCR sur les documents anciens.

Pour illustrer les problèmes des OCR, nous nous appuierons sur le logiciel FineReaderXIX[6] de la société ABBYY (leader des OCR sur le marché des documents contemporains). Il s'agit d'un logiciel d'OCR dédié aux documents latins, imprimés depuis le début du XVIIIeme siècle. Cet outil a été créé dans le cadre du projet européen METAe pour faire face aux lacunes des OCR classiques. Cette version de Finereader, s'appuie sur cinq dictionnaires (pour 5 langues) bâtis sur l'étude d'une centaine d'ouvrages anciens, ce qui a permis d'archiver près de 100.000 termes spécifiques issus de documents anciens tels que livres, journaux, magazines... Cet apprentissage a permis, selon l'offre commerciale, de lever les difficultés principales auxquelles les OCR traditionnels échouent systématiquement. La figure **Fig. 1.5**, qui est extraite d'un document français du XVIIIème, illustre une difficulté récurrente rencontrée par les OCR. Ici le premier challenge consiste à identifier la ligature entre le " f " et le " t ". La deuxième difficulté se situe au niveau de l'analyse du mot lui-même puisque le verbe " être ", ici conjugué au présent de l'indicatif, donne " eftes " et non " êtes ". En effet, la plupart des logiciels utilisent pour aider à transcrire les mots, des dictionnaires qui dans le cas des documents anciens ne sont pas adaptés.

La figure **Fig. 1.6**.a illustre le type de résultats que l'on obtient avec FineReaderXIX lorsqu'il

[6] http ://www.abbyy.com/

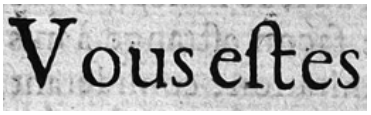


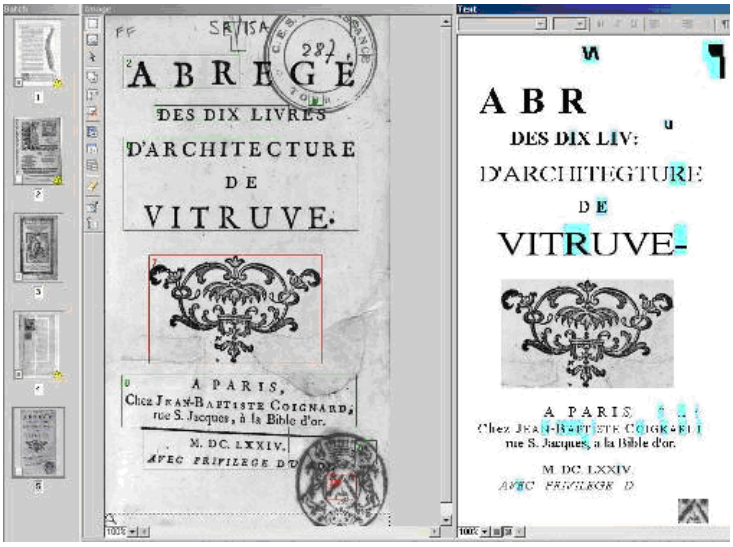
FIG. *1.5: Limites des ocr lors de l'analyse de mots inconnus*

est appliqué sur des documents anciens datant du 18ème siècle et la figure **Fig. 1.6**.b sur un document plus ancien. Nos conclusions corroborent celles avancées par [ELEL03, bET+01]. Sur des documents du 18ème, c'est à dire relativement proche de ce que l'on connaît actuellement en terme de mise en page et de polices utilisées, le taux de reconnaissance est proche de 80%. Pour les documents plus anciens les résultats ne sont pas évaluables. On remarquera sur les exemples donnés, que la structure de l'ouvrage est perdue après la phase de reconnaissance des caractères. Une autre difficulté provient des illustrations qui ne sont pas identifiées, voire même prises pour du texte (des exemples et des explications techniques seront donnés dans le chapitre suivant).

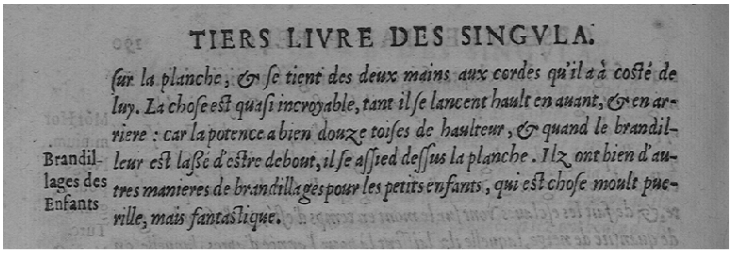
Il existe une alternative à l'OCR qui est la transcription, ou plutôt de l'aide à la transcription. Les Projets DEBORA et BAMBI proposent ce type d'outils.

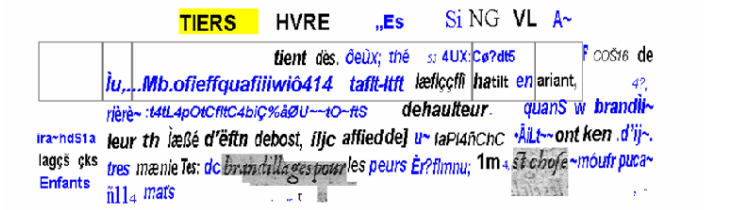
Projet européen lancé en 2000, DEBORA avait pour objectif de concevoir un ensemble d'outils permettant " l'accès distant et collaboratif à des livres numérisés du XVIème siècle " [BELM00]. Cet accès généralisé à des bibliothèques numériques a permis la création d'une plateforme prenant en charge plusieurs fonctionnalités afin d'assister les divers acteurs de la phase de numérisation à celle de l'exploitation et de l'archivage. La transcription, la structuration des données et la compression sont les modules " informatique " les plus intéressants développés dans le cadre de ce projet.
L'outil de transcription a pour objectif de palier aux lacunes des OCR traditionnels. En effet, ces derniers se basent sur des caractères dont les polices correspondent à des caractéristiques contemporaines et sont donc incapables de reconnaître les lettres utilisées. Les acteurs du projet DEBORA ont préféré aborder le problème d'une autre manière. Ainsi, la transcription est réalisée de manière semi-automatique. Dans un premier temps, chaque forme connexe de l'ouvrage est extraite et isolée automatiquement. Si nous n'avons aucune information sur la taille ou la police des caractères utilisés, on sait néanmoins qu'un caractère apparaîtra plusieurs fois sous la même forme dans tout un ouvrage. De ce fait, pour chaque caractère isolé, une comparaison de leur forme est réalisée et permet de construire un dictionnaire de formes référençant chaque occurrence (**Fig. 1.7**). L'utilisateur intervient ensuite pour associer un code ASCII à chaque classe de formes répertoriées dans le dictionnaire ; il va par exemple indiquer que la première ligne de formes correspond à la lettre "c". D'après [BGPR02], la saisie de 5% des caractères d'un livre de 200 pages contenant environ 2000 caractères par page permet sa transcription en 6 heures.
DEBORA offre donc une alternative au problème de la reconnaissance de caractères de documents anciens en incluant un utilisateur dans son processus de reconnaissance plutôt qu'un dictionnaire ou qu'une phase d'apprentissage.

Autre projet d'envergure , le projet BAMBI (Better Access to Manuscripts and Browsing of Images [BC97]) a pour objectif de créer une station de travail dédiée à la manipulation (par des experts) d'images de documents anciens. On y retrouve la majorité des fonctionnalités développées pour ce genre de stations de travail : visualiser, transcrire, annoter et indexer des images de manuscrits anciens. Dédiée à un public plus spécialiste que celui visé par le projet DEBORA, la plateforme développée pendant le projet BAMBI propose un outil de transcription basé sur



(a) document du 18ème siècle





(b) document du 16ème siècle

FIG. *1.6: Exemple de résultats d'ocr [Bén04]*

le modèle paléographique. Ainsi, après que le texte a été transcrit manuellement (à l'inverse de DEBORA), un algorithme calcule une similarité entre l'image et le texte transcrit. Il est ainsi possible d'établir un lien entre le texte et son image. En plus de cette transcription, BAMBI offre la possibilité de générer des index verborum et locorum qui permettent d'extraire ces informations sur le contenu du texte (occurrence des termes, position des mots dans le texte...).

Du côté des industriels, des projets sont également en développement. Le plus célèbre (ou le plus retentissant !) est certainement le projet Google Print (source [Sal05]). Initiés par Larry Page et Sergey Brin, les créateurs de Google ont mis en place un programme de numérisation et d'indexation du contenu de 15 millions d'ouvrages (l'équivalent de 4,5 milliards de pages). Au-delà des considérations politiques et des réticences nationales [Jea05], l'engagement pris par



FIG. *1.7: Exemple d'une transcription de caractères d'une page avec DEBORA [BELM00]*

la firme Google et plus récemment d'autres firmes leader du monde informatique (Microsoft, Yahoo, Amazon,...) et les diverses réactions suscitées témoignent des enjeux économiques mais aussi scientifiques qui en découlent. La figure **Fig. 1.8** montre le mode de fonctionnement de Google Print. Après avoir saisi un ou plusieurs mots-clefs, le résultat s'affiche sous la forme d'images extraites d'un ou plusieurs ouvrages. On remarquera que c'est le contenu de l'image qui est indexé. Le moteur de recherche a surligné les passages de la page qui font référence aux mots-clefs entrés. Les informations sur les méthodes utilisées pour l'indexation de ces bases d'images sont distillées au compte-gouttes et ne font en aucun cas foi de vérité. Il n'existe pas de sources officielles, mais il semblerait que cette indexation soit le résultat d'un traitement automatique d'OCR dédiés aux documents anciens. Google Print est donc l'archétype même d'une librairie numérique : on ne peut accéder qu'à des extraits d'images, l'accès à la totalité des données étant payant.

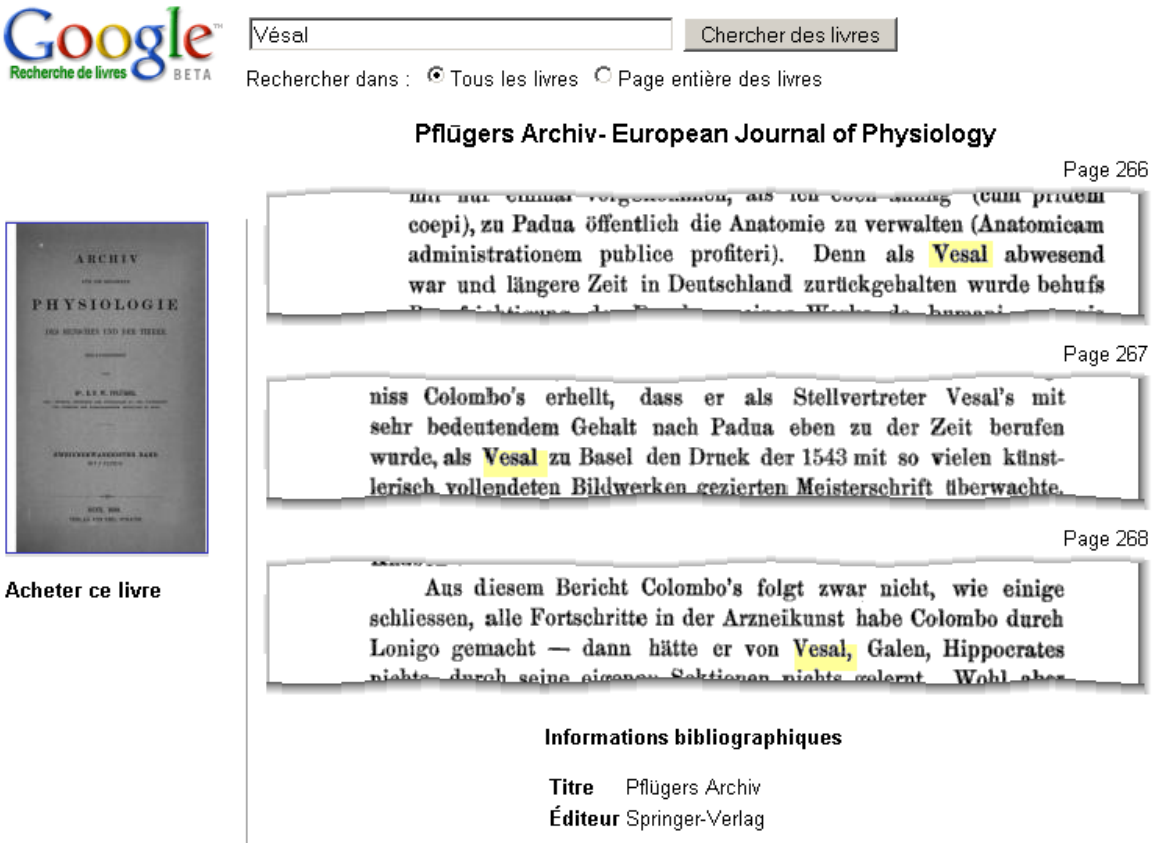


FIG. *1.8: Recherche d'information via Google Print [BELM00]*

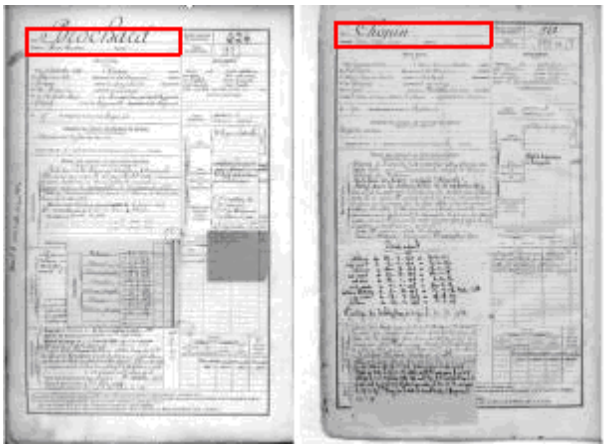
1.3.2 L'analyse de structure au travers des réalisations actuelles

La transcription où "l'océrisation" du texte ne sont pas les uniques outils susceptibles d'intéresser les utilisateurs. A l'instar de ce qui se fait sur les documents contemporains, certains besoins se font ressentir au tour de l'analyse de la structure des documents anciens. Il se trouve qu'au delà du texte, l'image de document contient d'autres informations (aspect visuel, répartition de l'information,...) qu'il est intéressant d'extraire pour permettre d'offrir de nouveaux types d'outils aux utilisateurs. Au travers de 4 projets, nous montrons les besoins spécifiques de certaines institutions et les premières propositions qui en découlent.

1.3.2.1 Le projet DMOS

Développé à L'IRISA de Rennes par l'équipe Imadoc, le projet DMOS (Description avec Modification de la Segmentation : [CR03]) est une plate-forme générique de reconnaissance de structures de documents. DMOS est en fait à la base de plusieurs plates-formes d'analyse de documents (ScoreRead pour la reconnaissance de partitions musicales, Matread pour la reconnaissance de formules mathématiques...). Les documents traités par DMOS sont bien spécifiques. Ces derniers doivent posséder une structure forte, stable et surtout descriptible par un ensemble de règles définies par un utilisateur expert. DMOS a été utilisé pour la création d'une plateforme d'extraction et d'indexation de formulaires d'incorporations militaires du XIXe siècle stockés

aux archives de la Mayenne et des Yvelines (**Fig. 1.9**.a). Dans le cas précis de FormuRead, l'indexation ou l'analyse de la structure se traduit, par exemple, par la localisation des cases où sont sensés se trouver les noms des incorporés. En s'appuyant sur des connaissances précises (les noms sont toujours en haut à gauche dans le premier rectangle), il est possible de produire un ensemble de règles visant à retrouver ce rectangle et à extraire le nom de l'incorporé qui y est écrit. On notera que pour chaque nouvelle famille de documents que l'on souhaite indexer, il faut produire de nouvelles règles adaptées. FormuRead est actuellement testé sur les formulaires d'incorporation militaire du XIXe siècle et ceci malgré leur dégradation. Ce logiciel a été testé sur 60 223 pages des Archives de la Mayenne. Certains postes d'accès aux formulaires d'incorporation militaires fonctionnent à l'aide d'une tablette graphique à retour visuel. Introduit dans "un poste du futur ", comme le dit [DeB04], DMOS trouve son utilité dans le fait qu'il permet " d'associer facilement la consultation d'archives papiers et numériques".



(a) Archives militaires dont il faut extraire le nom des incorporés

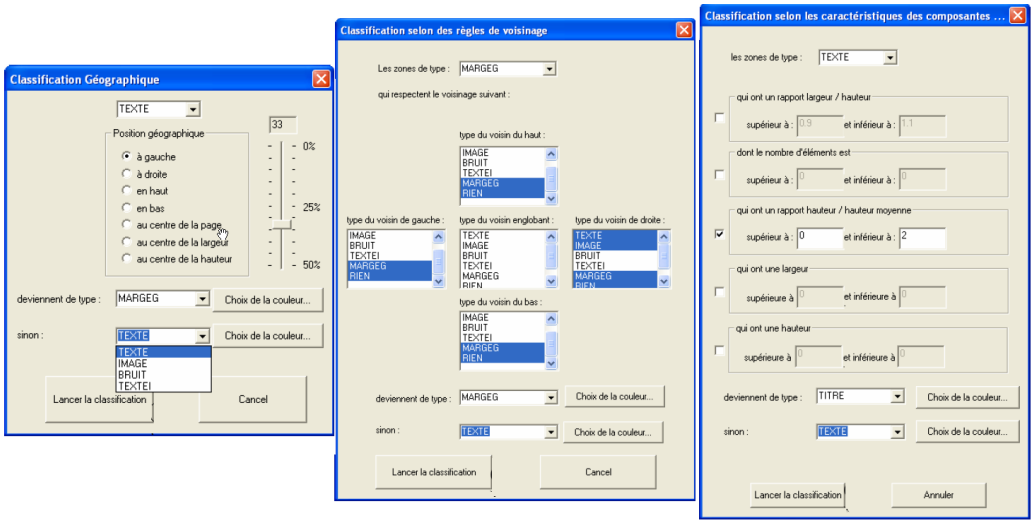


(b) Exemple d'une tablette graphique permettant à un utilisateur d'effectuer des recherches sur la structure

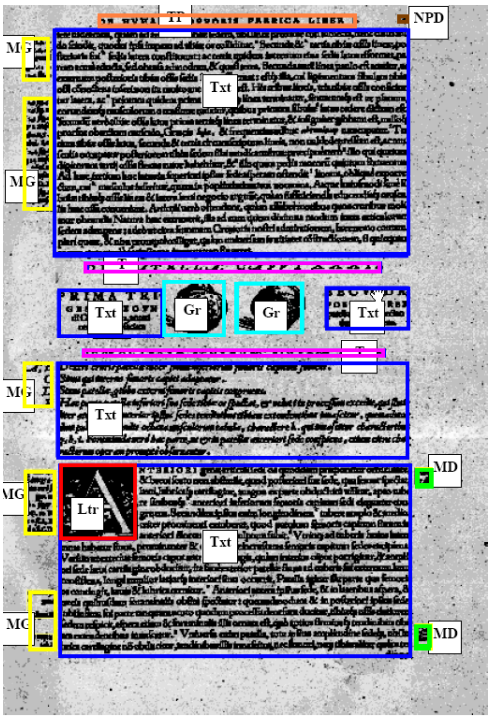
FIG. *1.9: Indexation de la structure de documents anciens avec le logiciel Fomuread [CR03]*

1.3.2.2 Le projet Agora

Le projet Agora a été développé en association entre le Laboratoire d'informatique de Tours et le Centre d'Etude Supérieur de la Renaissance. Le CESR désire donner accès à ses ouvrages de manière classique, via le Web, en proposant des liens vers les ouvrages. Il souhaite également offrir de nouveaux usages aux utilisateurs en leur donnant accès au contenu des images, par exemple les illustrations. La réalisation de cette fonctionnalité passe par réalisation d'outils informatiques permettant de localiser et d'extraire les illustrations présentes dans plusieurs milliers de pages de documents. La plate-forme qui découle de cette collaboration répond à cet objectif bien précis. En effet, il permet à un expert du domaine d'indexer le contenu d'un ouvrage ([RBD06]). A l'aide de descripteurs simples (position de l'entité, position par rapport aux entités voisines, forme et contenu de l'entité), l'utilisateur construit des scénarios d'indexation sur des images qu'il a choisies (un peu comme une macro). Une fois le scénario validé, il est appliqué sur le reste des images de l'ouvrage à indexer et permet d'extraire rapidement de grandes quantités d'illustrations alors que cela prendrait un temps incalculable de le faire manuellement. Par exemple, si un utilisateur souhaite extraire toutes les lettrines d'un (ou plusieurs) livre ancien afin de constituer une base de lettrines, il définira un scénario du type : " Cette entité est une lettrine, elle est toujours située dans les 20% gauche de l'image, son rapport largeur/hauteur est entre 0.75 et 1.25 et son plus proche voisin de droite est de type texte " (cf figure **Fig. 1.10**.a) Le résultat d'une identification de la structure après description de 6 entités par l'utilisateur est donné dans la figure **Fig. 1.10**.b. Après plus d'un an d'utilisation, Agora a déjà permis d'indexer plus de 17000 pages et de classer près de 5000 illustrations (lettrines, portraits...).



(a) Interface d'Agora permettant de définir des scénarios d'indexation



(b) Résultat d'une identification de la structure avec MG :marge gauche, MD :marge droite, Ltr :lettrine, NPD :Numéro de page, TP :Titre principal, T Titre

FIG. *1.10: Indexation de la structure de documents anciens avec Agora [RBD06]*

1.3.2.3 Projet METAe/DocWorks

Projet débuté en 2000 en partenariat avec le BNF, METAe ([Jou04]) avait pour objectif de développer un ensemble d'outils informatiques allant de la reconnaissance de la structure logique des documents à la reconnaissance de caractères en passant par la génération de métadonnées d'images de documents anciens du XIXème et XXème siècle. Le partenariat autour de ce projet a permis de faire émerger un cahier des charges précis où les spécifications de chacun des partenaires étaient clairement définies. Logiciel développé par la société allemande CSS-GmbH[7], docWorks propose une suite logicielle permettant d'assister ou de prendre en charge toute la chaîne de numérisation (de la capture au stockage XML). Leur association au projet METAe a permis de réaliser un outil d'analyse et de reconnaissance de structure. Le fait de traiter des documents bien spécifiques dans leur origine historique (XIXème et XXème siècle, ouvrages européens...) et dans leur contenu (ouvrages littéraires et articles) a permis de créer un modèle (au format EMTS) précis des documents sur lesquels va s'appliquer le logiciel. Ainsi, docWorks peut reconnaître des champs spécifiques comme les numéros de page, les titres, taille des polices, les notes de bas de page... et permettre de déterminer automatiquement la structure de la page et de l'ouvrage entier. Le projet étant financé et distribué par des fonds privés, il reste très difficile de trouver des informations sur les aspects scientifiques de ce projet. La figure **Fig. 1.11** montre un extrait du logiciel DocWorks après reconnaissance automatique de la structure d'un ouvrage. On voit notamment que cette analyse permet de construire automatiquement un arbre permettant d'accéder à la structure hiérarchique de l'ouvrage.

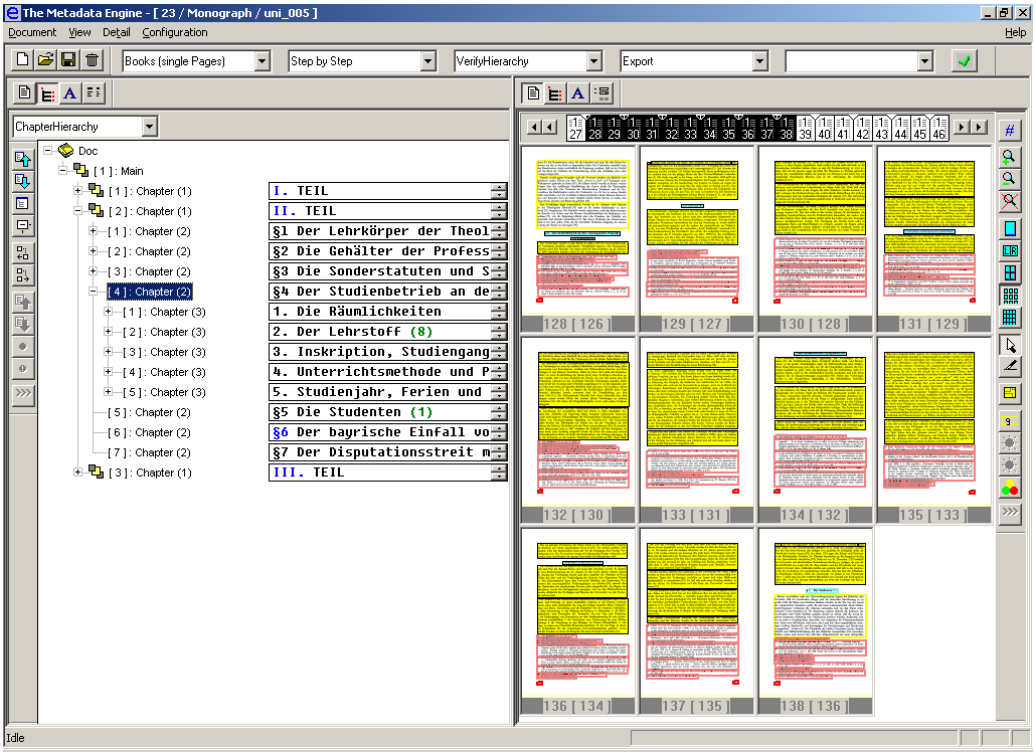


FIG. *1.11: Indexation de la structure de documents anciens avec Docworks [Jou04]*

[7] http ://www.ccs-gmbh.de

1.3.3 L'analyse des illustrations

La majorité des travaux relatifs à l'indexation de documents anciens a trait à l'analyse de la structure, à l'OCR, à la numérisation... Hors, peu d'entre eux sont consacrés à l'indexation des illustrations qui émaillent les pages des collections numérisées. La nature même de ces illustrations rend leur analyse extrêmement complexe. L'impression des lettrines, bandeaux et autres dessins était autrefois réalisée à l'aide de tampons (la plupart du temps en bois) façonnés à la main et que l'on imbibait d'encre pour finalement les apposer sur les pages blanches. Cette conception manuelle des tampons se caractérise par des illustrations constituées d'une multitude de petits traits disposés les uns à côté des autres (cf. figure **Fig. 1.12** pour des exemples d'images de traits). Comme nous aurons l'occasion de le développer dans ce mémoire, cette caractéristique rend ardue l'analyse automatique de ces illustrations. Face à cette difficulté, les industriels ou scientifiques ont fait le choix de l'indexation manuelle de leurs bases d'images. Or, on le sait, même avec un dictionnaire prédéfini, la subjectivité de la personne qui annote les bases (ou qui émet une requête) rend complexe, voire peu efficace la recherche par mots-clefs. D'autre part, cela nécessite une action manuelle sur chaque illustration. Il n'existe pas, à ce jour, de plates-formes opérationnelle qui incluent un outil d'indexation automatique d'illustrations. Des travaux sont néanmoins en cours ; ils relèvent de la mise au point d'outils logiciels permettant la comparaison d'images comme on en trouve en indexation d'images naturelles. Le principe de ces travaux consiste à fournir une image requête (par exemple une lettrine) au système qui recherche au sein de la base toutes les portions d'images semblables au sens d'un critère. Cet objectif s'applique à des besoins bien précis exprimés par la communauté des spécialistes travaillant sur des ouvrages anciens. Les cas d'utilisation sont bien concrets : reconstruire l'alphabet de lettrines utilisées par un éditeur, étudier l'usure des tampons, étudier les dessins des lettrines, étudier l'évolution des techniques utilisées, identifier les faussaires qui imitaient certaines illustrations d'éditeurs...

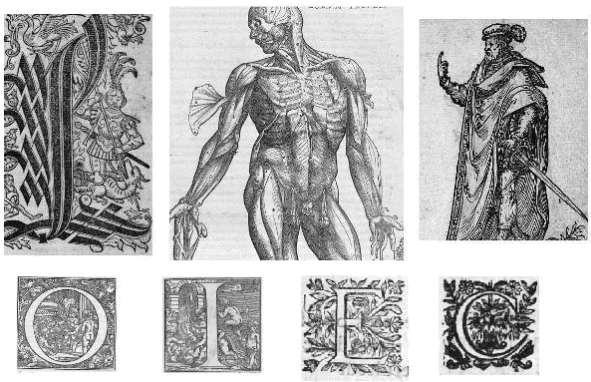


FIG. *1.12: Indexation d'images de traits [UOL05]*

Au niveau des bases numériques d'ornements existantes, il est impossible d'en faire une liste exhaustive. En voici néanmoins quelques une référencées par [BBM96] :

1. La BCU de Lausanne : base d'illustrations du XVIIIe dont l'objectif est de permettre la comparaison du matériel utilisé à l'époque pour illustrer les ouvrages.
2. Le projet Fleuron : Un peu comme pour la BCU de Lausanne, l'objectif est de mettre à disposition une base d'images permettant de comparer les illustrations entre elles et d'observer, par exemple, la récurrence des motifs d'un éditeur à l'autre...

3. Projet Môriane : Permet d'avoir accès aux ornements contrefaits utilisés par de imprimeurs liégeois du XVIIIe.
4. Projet Athens : Catalogue d'ornements.

Parmi les travaux déjà réalisés ou toujours en cours de développement, le projet TODAI (Typographical Ornaments Database and Indexation [BBM96]), et les travaux réalisés par [PVU+06] et [UOL05] sont ceux qui semblent être les plus aboutis. Leur principe est simple : offrir la possibilité d'une recherche automatique d'illustrations de documents anciens sans description par mots-clefs ; seul les critères visuels sont pris en compte. Dans la pratique, il suffit de soumettre une image en exemple et le système se charge du calcul de toute une quantité d'indices qui lui permetta de comparer l'image requête, avec celles stockées dans la base. La figure **Fig. 1.13** donne un exemple de requête réalisée sur une lettrine et dont la similarité est calculée comme l'indique [UOL05].

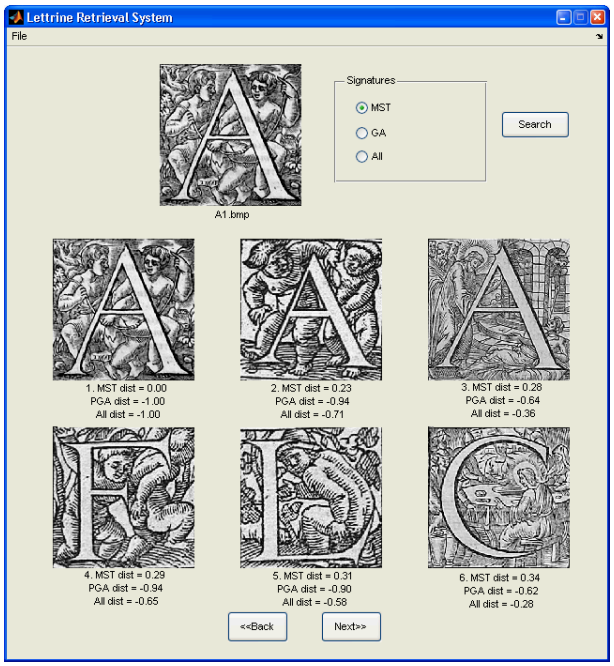


FIG. *1.13: Indexation d'images de traits [UOL05]*

1.4 Conclusion

Lorsque le 14 juillet 1988, le Président de la République annonçait *"la construction et l'aménagement de l'une ou de la plus grande et la plus moderne bibliothèque du monde....(qui) devra couvrir tous les champs de la connaissance, être à la disposition de tous, utiliser les technologies les plus modernes de transmission de données, pouvoir être consultée à distance et entrer en relation avec d'autres bibliothèques européennes"*, il ne se doutait peut être pas à quel point cette déclaration serait le détonateur de 20 ans de recherches (toutes disciplines réunies) d'un enjeu majeur et visant à sauvegarder, diffuser, communiquer et transmettre notre patrimoine historique.
Ce premier chapitre a permis de faire ressortir les principales attentes des usagers en matière

de recherche d'informations dans les bibliothèques numériques et de relever les propositions innovantes des scientifiques et chercheurs pour l'établissement de solutions d'aide à l'indexation des contenus. Dans ce contexte, on peut discerner deux catégories d'approches venant compléter celles citées précédemment. D'une part, l'indexation manuelle, qui a l'avantage d'être simple à réaliser, avec par exemple dans sa mise en oeuvre, la mise à disposition de moteurs de recherches par mots-clefs. D'autre part, l'indexation automatique au travers des premières offres logicielles qui permettent d'entrevoir tous les bénéfices qu'un traitement automatisé pourrait apporter, d'autant plus quand on travaille sur une grande masse de données. On peut ainsi remarquer que dans le domaine de la recherche, les premiers travaux réalisés commencent à porter leurs fruits. Du côté de la recherche industrielle, les solutions proposées répondent à des problèmes bien spécifiques, par exemple, la suite logicielle proposée par ABBY est indiquée comme fonctionnant pour des ouvrages du XIXème avec des fontes bien précises. Du côté de la recherche universitaire, une problématique plus large a permis l'élaboration d'outils non encore testés à grande échelle, mais dont les premiers résultats permettent d'espérer la mise en place de systèmes d'indexation performants.

Chapitre 2

Analyse de contenu d'images de documents : Etat de l'art

2.1 Préambule

2.1.1 Introduction

Selon [Mic00] : *"Tout document textuel est construit selon une structure, qui est reconnue par les lecteurs humains grâce à des marques typographiques, des conventions de mise en page, des connaissances pragmatiques, culturelles, relatives aux informations génériques qu'est susceptible de contenir ou bien que doit contenir tout document particulier, appartenant à une certaine catégorie."* . Cette définition illustre bien le fait que sans sa structure un document textuel, peut perdre tout son sens et que la structure d'un document est elle-même une information. La structure d'un document, même si elle est implicite, véhicule une information indispensable à la compréhension du texte. La lecture d'une page de journal (**Fig. 2.1**.a) illustre bien ce principe. L'organisation et la hiérarchie du texte et des images, qui dans cet exemple, se présentent sous la forme de titres/colonnes/paragraphes/légendes... rendent l'interprétation possible grâce à la diversité des graisses, polices, tailles de textes, positions des caractères... utilisés pour mettre en page le texte. A l'aide de ces "règles" ou "consensus visuels" on peut par exemple accéder ou prendre connaissance de l'information principale sans avoir à lire l'ensemble du texte. Mais les journaux ne sont pas les seuls exemples que l'on peut citer : Que serait une partition de musique sans la structure que représente une portée ? Que serait un calligramme de Guillaume Apollinaire (**Fig. 2.1**.b) sans la structure qu'il arbore ? Que serait enfin un annuaire téléphonique si les numéros étaient écrits les uns derrières les autres et non plus en colonne avec le nom sur la ligne correspondante ?

La reconnaissance de structures qui se place classiquement dans un dispositif de rétro-conversion, a été très rapidement l'objet de nombreuses recherches appliquées. Identifier la structure d'une enveloppe pour localiser et analyser le code postal et trier automatiquement le courrier ([KS92, Yac96]), localiser où se situe le montant d'un chèque pour permettre le traitement des montants ([LLC93]), analyser ou trier des documents papiers dont le format est standardisé (CAF, impôts, banques, INSSEE...) afin de permettre un traitement accéléré des réponses ([HDD+05]), séparer les zones de textes de dessins pour améliorer les performances des OCR [Bre02] sont quelques exemples parmi l'ensemble de ceux que l'on peut trouver dans la littérature. Ce chapitre propose une synthèse des travaux réalisés en analyse de documents impri-



(a) *Une d'un journal*



(b) *Un calligramme*

FIG. *2.1: Exemples de documents structurés*

més (livres, articles, journaux, publicités...). Les méthodes présentées portent principalement sur l'analyse de la mise en page, qui à ce jour demeure encore l'étape fondamentale à tout système de reconnaissance et d'indexation. Après la présentation et la définition de termes que l'on retrouve de manière récurrentes dans la littérature, nous exposons les différentes approches utilisées pour reconnaître les structures de documents. La rédaction de cet état de l'art est motivé par le fait que la problématique de l'indexation d'images de documents (texte, illustrations, mise en page...) est étroitement liée à la reconnaissance de structure. En effet, nous allons voir dans la suite de ce chapitre, que la majorité des méthodes s'appuie sur cette structuration de l'information pour proposer des outils d'indexation du contenu. Nous accompagnons la présentation de cet état de l'art, par des discussions et des comparaisons d'approches basées sur de nombreux résultats expérimentaux que nous avons menés en développant certaines de ces méthodes.

Il est à noter qu'il existe déjà un bon nombre d'états de l'art dans ce domaine. Parmi ces derniers, on peut citer les plus connus et les plus récents comme : [Doe98b, Nag00, Tru05, CCMM98, TLS96, TCL+99, MRK03]. Tous relèvent la nécessité de structurer l'information avant de la reconnaître en engageant des techniques de rétroconversion. Nous avons choisi une démarche alternative qui consiste à ne pas segmenter les données contenues sur les pages des ouvrages, mais à les enrichir toutes d'une description relative à leur proche voisinage. C'est sur la base de cette description que reposera notre système de recherche d'informations.

2.1.2 Qu'est qu'une structure ?

Selon [Doe98b] la structure d'un document se présente sur 3 niveaux : la structure physique, la structure fonctionnelle intermédiaire et enfin la structure logique. Cette définition trouve son sens dans le processus même de création d'un document que l'on peut considérer comme fonctionnant en 3 étapes. Prenons l'exemple de la conception d'une première page d'un journal. Tout d'abord, il faut concevoir la structure logique du document : celle-ci correspond au sens que l'on souhaite donner à cette page et se traduit par un choix de titres, de photos et de mise

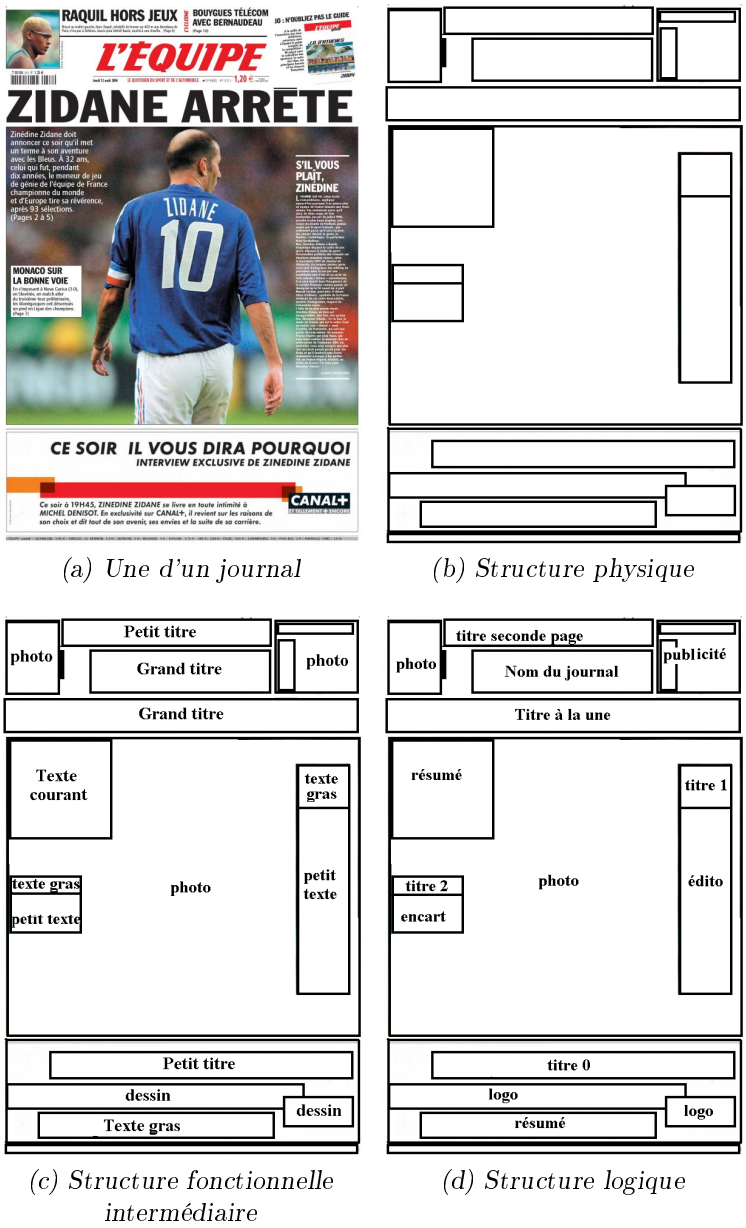
en page. La deuxième étape consiste à affecter une caractérisation visuelle concrète et définir l'organisation spatiale des éléments à mettre en page (eg : un titre sera une zone de texte de gros caractères en haut de la page, un édito est un texte en italique sur deux colonnes et toujours à gauche). Finalement, la structure physique est le résultat du processus de réalisation intégrant les contraintes logiques, fonctionnelles et enfin la taille de la page. La rétroconversion de ce document est le processus inverse de celui qu'on vient d'exposer. Partant du document papier, une acquisition et une segmentation permet d'accéder à la structure physique du document (**Fig. 2.2**.b). La reconnaissance des éléments segmentés (texte en gras, image couleur, dessins, petits caractères...) donne accès à ce qu'on appelle la structure fonctionnelle intermédiaire (**Fig. 2.2**.c), pour finalement permettre de retrouver la structure logique du document (**Fig. 2.2**).c via l'interprétation de la structure fonctionnelle intermédiaire (l'image est composée d'un titre, d'un édito, d'une grande photo...)

Les techniques mises en place pour la rétroconversion de document sont souvent le fruit de la combinaison de plusieurs techniques de reconnaissances. Usuellement les états de l'art les séparent en deux familles :

- Les approches mélangeant connaissances a priori et analyse d'images. On a coutume des les présenter sous trois formes :

  1. Approches ascendantes : elles sont guidées par les données et n'incluent donc pas (ou peu) de connaissances sur le modèle. Elles se basent sur l'extraction de données bas niveaux (couleur, position...) relatives aux pixels. Au fur et à mesure des traitements une interprétation des résultats est rendue possible, pour finalement remonter vers la forme logique du document et d'une comparaison avec le modèle
  2. Approches descendantes : elles sont guidées par le modèle du document. Souvent utilisées pour des documents à structure bien définie et invariante les approches ascendantes s'appuient sur cette forte connaissance a priori pour guider la segmentation et la reconnaissance.
  3. Approches mixtes : elles regroupent souvent des méthodes non séquentielles qui embarquent en partie des approches descendantes et également des approches ascendantes. La frontière entre ces trois familles de méthodes n'est pas toujours clairement définie. Si quelques outils (ou algorithmes) classiques peuvent être catalogués comme appartenant à l'une ou à l'autre, les systèmes mis en place ont plutôt tendance à emprunter aux trois.

- les approches textures : traditionnellement issus de l'analyse d'images naturelles, les outils de caractérisation de textures se trouvent être particulièrement bien adaptés à l'analyse de structure de documents. En effet, les aspects fortement texturés du texte, des photos et des dessins rendent possible la description très fine de fontes, de polices mais aussi des éléments non textuels. Ces techniques peuvent être utilisées dans les approches ascendantes mais, on les retrouve également dans une autre classe de traitements s'appuyant non plus sur l'interprétation de l'ensemble de la page, mais plus sur le parcours des documents pour la recherche (le butinage). Dans ce cas, la structure n'est pas à interpréter, on parle de caractérisation du contenu.



(a) Une d'un journal (b) Structure physique

(c) Structure fonctionnelle intermédiaire (d) Structure logique

FIG. 2.2: Les différentes structures d'un document

2.1.3 Notion de classe et de complexité de documents

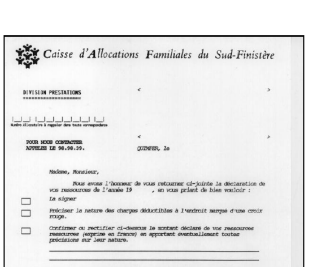
Il n'existe pas de méthode générique permettant de traiter tous les types de documents existants. Les traitements sont plutôt adaptés à une famille de documents.
La diversité des documents traités dans la littérature est si vaste qu'on aurait même l'impression qu'elle en est inépuisable. A chaque nouvelle famille, on ne recrée pas nécessairement de nouveaux outils de traitement d'images. On a souvent une adaptation d'une part des algorithmes disponibles, et d'autre part une mise en place de nouveaux séquencements adaptés des algorithmes. Bref, le processus est différent mais les algorithmes le sont rarement. Les états de l'art

ont l'habitude de présenter ces méthodes en fonction de la famille de traitements utilisée. Mais il ne faut pas perdre de vue que c'est la nature même des documents qui, la majeure partie du temps, oriente les méthodes et les procédures de reconnaissances choisies, et non l'inverse.
Dans [**?**] l'auteur aborde cette question en introduisant une notion de « complexité » de documents qui serait liée à la variabilité des objets qui les composent. Cette notion de complexité des documents est également fondamentalement liée au degré et à la variabilité de la mise en page.

Prenons quelques exemples de familles de documents traités dans la littérature et qui illustrent cette variété de complexité :

- Les formulaires ([DA02, HER01]) : Ces documents ont une partie de leur structure qui est connue a priori et dont on sait qu'elle restera stable. Selon [DA02], un formulaire est décomposable en trois couches : le squelette (lignes, cadres...), les données pré-imprimées (intitulées du genre : nom, prénom, adresse,...) et la dernière qui est celle relative aux données ajoutées (manuscrites, imprimées...).
- Les journaux ou documents publicitaires : Leur spécificité est qu'il est très difficile d'émettre des hypothèses sur les éléments les composant ou sur des règles de mise en page étant donnée la variété des journaux existants. La plupart du temps, les solutions proposées traitent un type bien précis de journaux dont la mise en page et les règles d'édition changent peu. On peut citer [ROB01a] qui propose un système de reconnaissance de structure de journaux du Los Angeles Times.
- Les documents anciens : Composés principalement de texte et parfois de dessins de traits, les documents anciens ont pour particularité d'être des témoins de leur histoire. En effet, plusieurs siècles ont été nécessaires avant d'obtenir une relative stabilité des techniques d'imprimerie et de mise en page. Ceci se traduit par des ouvrages aussi divers que variés, dont il est impossible de ressortir un modèle générique. Les solutions proposées à ce jour ([RBD06, CR03]) se basent essentiellement sur des systèmes incluant un utilisateur expert qui indique lui même le modèle logique et physique des documents anciens qu'il souhaite indexer. Sans aller jusqu'à retrouver la structure logique des documents anciens, certains se sont intéressés à la séparation des couches de dessin et de texte sur la base d'approches ascendantes, descendantes ou texture [BELM00, JMRE05].

La figure **Fig. 2.3** montre quelques uns des documents cités précédemment.



(a) formulaire de la CAF



(b) Journal



(c) Documents anciens

FIG. *2.3: Extrait de la variété des documents traités en analyse de documents*

2.1.4 Notre corpus d'images de documents anciens

Comme nous l'avons montré au paragraphe précédent, il convient donc de définir le corpus de travail afin de construire l'analyse. Dans le cadre d'une collaboration avec le CESR de Tours, nous avons eu accès à plus d'une centaine d'ouvrages déjà numérisés. Parmi cette masse d'information disponible, nous avons choisi 17 ouvrages, ce qui représente près de 700 images. Cet échantillon d'ouvrages a été réalisé dans l'optique de proposer une forme de variabilité sans prétention de généricité. En effet, la caractéristique des documents anciens porte avant tout sur une hétérogénéité forte des ouvrages disponibles. 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.

Ainsi notre corpus, avec près de 700 pages, recouvre 3 siècles d'imprimerie et d'histoire. Ces ouvrages ont pour spécificité leur mise en page complexe et variée (plusieurs colonnes de taille irrégulière), l'utilisation de fontes spécifiques (plus utilisées de nos jours), 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...). La figure **Fig. 2.4** illustre cette diversité des documents traités.

D'un point de vue plus technique la figure **Fig. 2.5** synthétise les caractéristiques des images que nous allons être amener à traiter. Pour information, cette base est constituée pour la moitié d'images au format JPEG et de l'autre moitié d'images au format TIFF (compressé). L'écrasante majorité des ouvrages a été numérisée à 300 points par pouces (ppp) dont une minorité en couleurs (quelques pages de garde possèdent des illustrations en couleurs).



FIG. *2.4: Extrait de notre corpus d'images de documents anciens*

Ces images ont été numérisées selon un protocole bien précis et avec un matériel adapté. En ce qui concerne la numérisation des ouvrages anciens, l'équation posée est simple : comment numériser des quantités phénoménales d'ouvrages tout en préservant leur intégrité physique et en assurant une qualité de numérisation permettant leur exploitation future ? En 2000, [Kal00] a proposé un bilan des méthodes de numérisation. Ce qui en ressort, entre autres, c'est qu'au fur et à mesure des années, les progrès liés aux évolutions du matériel ont permis d'améliorer la qualité et l'efficacité du processus de digitalisation. Un bref tour d'horizon de l'offre commerciale montre que les performances techniques permettent, à la fois une numérisation de bonne qualité (de 250 à 800 ppp), pour une qualité de compression devenue tout à fait acceptable. A ce propos, les auteurs de [BELM00] attirent l'attention des lecteurs sur le fait que les formats de compressions classiques (JPG, JPEG2000) dégradent considérablement les images de traits dans un processus de compression par blocs. Face à ce constat, certains chercheurs ont proposé des algorithmes de compression dédiés aux documents (anciens comme contemporains). Les deux plus connus sont ceux développés lors du projet Débora [BELM00, LET03] et le format de compression DjVU ([BHL+00]). Leur principe est relativement similaire, l'image est séparée selon un certain nombre de plans (dans le cas de [LET03] : le plan textuel, le plan graphique, l'arrière plan et le plan compensatoire). Etant donné que chacun de ces plans n'a pas la même importance

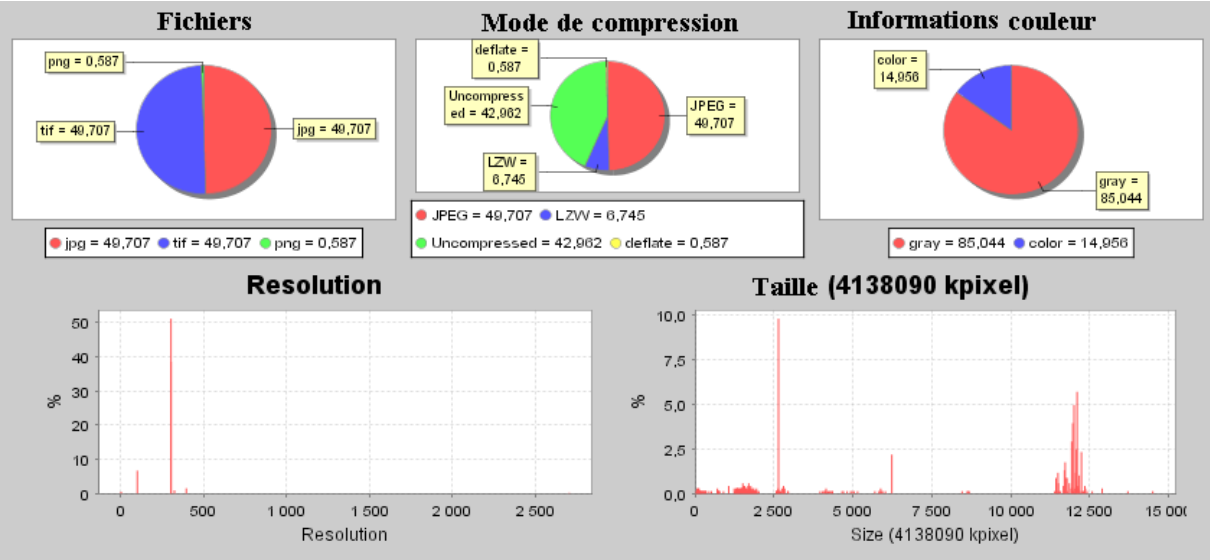


FIG. *2.5: Synthèse des caractéristiques de notre corpus*

visuelle, chacun d'entre eux va être compressé de manière différente. Ainsi, l'arrière plan va être considérablement compressé alors que le plan graphique va être compressé sans perte. Dans les deux approches, le plan textuel est compressé sur le principe de la redondance des formes entre les divers caractères. Ainsi, à l'aide d'un dictionnaire de formes, ne sera stockée dans l'image compressée que l'adresse vers le caractère et pas le caractère lui-même (c'est à dire la forme pixelaire). Le matériel utilisé permet également de préserver l'intégrité physique des ouvrages (pas de contact entre une vitre et le livre, éclairage non agressif...). Un exemple de scanner fabriqué par i2S[8] et qui est utilisé pour la numérisation de documents anciens est présenté sur la figure **Fig. 2.6**.Ce système de numérisation est composé d'une caméra linéaire associée à un dispositif de compensation de la forme de l'ouvrage (problème lié à la reliure).



FIG. *2.6: Exemple d'un scanner*

En parallèle , l'offre logicielle s'est elle aussi améliorée en proposant des outils performants dédiés en particulier à la restauration d'images de documents anciens. En effet, même avec le

[8] http ://www.i2s-bookscanner.com/fr/default.asp

meilleur matériel de numérisation disponible, il arrive que les fichiers images soient difficilement exploitables. Dans [Tri03], l'auteur décrit les défauts les plus couramment rencontrés. Ainsi, on distingue deux " familles " de défauts :

1. Ceux inhérents aux ouvrages : détérioration du papier, transparence du verso due à l'acidité de l'encre...
2. Ceux liés à la phase de numérisation : défauts d'éclairage, problèmes de courbure et d'inclinaison de l'image...

De nombreuses solutions ont été proposées. [Tri03] en recense plus d'une dizaine de méthodes (essentiellement développées par des industriels) et propose, lui aussi, diverses solutions pour venir à bout des problèmes évoqués précédemment. Dans la figure **Fig. 2.7**, on peut remarquer le type de corrections réalisées sur l'image après l'utilisation d'un logiciel de restauration de documents anciens. On peut notamment observer que les taches ont disparu et que les défauts de courbure ont été atténués. L'objectif est donc principalement d'améliorer le rendu visuel de la version numérique de ces documents. Ces outils déforment ou corrigent l'image afin de rendre un meilleur rendu. Cependant, ces traitements "endommagent" l'image initiale, ce qui n'est pas sans conséquences sur l'exploitation du contenu.



(a) Image numérisée en couleur *(b) Image en niveau de gris redressée* *(c) Image restaurée*

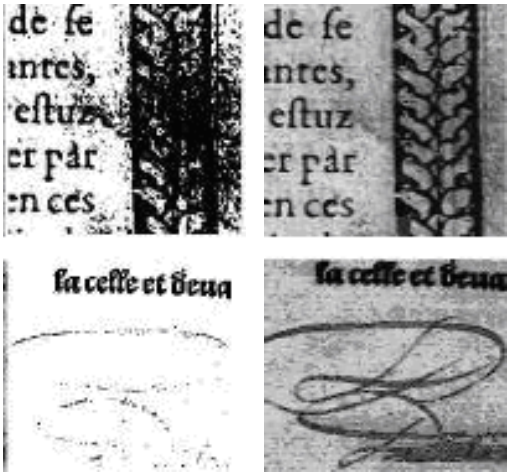
FIG. *2.7: Exemple de corrections de défauts de numérisation avec[Tri03]*

2.2 Analyse du contenu d'images de documents : les approches classiques

Si on s'intéresse au contenu des documents, il existe des approches classiques permettant d'extraire de l'information des images. Ces approches se basent dans la plupart des cas, d'une part sur un traitement de l'image afin de rendre l'information binaire, et d'autre part sur une segmentation des pages sur la base de familles de méthodes basées sur l'étude de la structure. La suite de ce chapitre présente chacun de ces points, illustré par des exemples.

2.2.1 Quelques mots sur le problème de la binarisation des documents anciens

La majorité des outils de traitement de contenu est utilisée sur des images binaires. En règle générale, l'étape de binarisation ne pose pas de problèmes spécifiques lorsque l'on utilise des algorithmes adaptatifs comme ceux proposés par Sauvola ou Otsu. Ils sont largement utilisés dans la littérature et produisent une binarisation de bonne qualité. Cette phase n'est malheureusement pas aussi triviale dès lors qu'on applique ces algorithmes sur des documents anciens. Le nombre d'articles récemment publiés et qui sont dédiés au processus de binarisation dédiés aux documents anciens témoigne de l'importance mais aussi de la complexité de cette étape. Parmi ceux-ci on peut citer [HBS05, KS06, HDDK05]. La figure 2.8, illustre les problèmes courants que nous avons observé après application de différents algorithmes de binarisation. Le premier problème principalement observé est, qu'après binarisation certaines parties de l'image peuvent disparaître. En effet, l'usure du temps fait disparaître l'encre de son support (ou tout du moins la rend moins visible) ce qui complique la tâche d'algorithmes usuellement habitués à disposer des caractères correctement imprimés sans équivoque entre le fond et l'encre. Le deuxième problème est celui lié à la présence très fréquente de taches d'encre. Elles ont des propriétés de couleurs et de formes très proches de celles des caractères d'imprimerie (ou de certaines illustrations) ce qui encore une fois complique la tâche des algorithmes de binarisation.



(a) Image en niveaux de gris *(b) Image binarisée*

FIG. *2.8: Exemple de la difficulté de binariser une image de documents anciens*

En règle générale, la démarche admise dans les travaux récents en analyse d'images de documents, consiste à repousser le plus tard possible l'étape de binarisation. Ce choix permet d'éviter la perte précoce d'une grande quantité d'informations, en simplifiant le contenu à une simple version binaire. Une voie raisonnable consiste à conserver les deux informations et de maquer des cycles de retour entre ces supports en cas de doute.

2.2.2 Méthodes d'analyse de structures

D'une manière générale, les méthodes d'analyse de structures utilisent de forts a priori sur la répétitivité supposée de la structure des documents dans un corpus. Elles se basent sur l'étude

de caractéristiques "physiques" via, par exemple, des techniques de regroupement ou fusion de pixels noirs, de découpage de zones d'images, d'étude d'alignements de pixels blancs... Il existe énormément de méthodes se classant dans cette catégorie et de très bons états de l'art les recensent déjà.

Plus qu'un listing des méthodes existantes, nous avons souhaité faire un rapide panorama des outils les plus utilisés (quelque soit leur type d'analyses : ascendantes, descendantes ou mixtes). Nous avons également porté une attention toute particulière à l'évaluation de certaines d'entre elles sur les images de notre corpus afin d'illustrer leur pertinence. Des travaux de comparaison de méthodes de segmentation d'images de documents sont proposés dans [SKB06]. Ils ont fait apparaître que, selon les images testées, certaines de ces méthodes sont sensibles au bruit, sont peu robustes aux mises en pages complexes, qu'il est parfois ardu de fixer les paramètres nécessaires... C'est sur la base de ces mêmes critères que nous allons évaluer ces méthodes.

Dans l'introduction de ce chapitre, nous avons défini la notion de complexité de documents. De manière un peu schématique, on retrouve, d'un côté les documents à structure stable et de l'autre, ceux à structure variable. Là où les premiers ont des caractéristiques de mise en page connues au préalable (notion de modèle de document), le deuxième type de documents se caractérise par une prévisibilité de la structure réduite (texte multi-orienté, espaces inter-lignes irréguliers, présence aléatoire de différents types de dessins...).

2.2.2.1 Documents à structure stable

a- Les composantes connexes

Très largement utilisé dans la littérature, cet algorithme est calculé après binarisation de l'image traitée. Les auteurs de [JS95] le définissent selon la notion de voisinage entre pixels : *"une composante 4-connexe (resp. 8 connexe) est telle qu'entre tout couple de pixels de la composante, il existe un chemin 4-connexe (resp. 8 connexe)"*. Il existe différentes versions d'algorithmes permettant d'extraire ces composantes, [JS95, CM91b] en recensent plusieurs. Leur popularité vient du fait que cet algorithme est applicable sur tout type de documents binaires, sans aucune connaissance a priori et qu'il s'adapte parfaitement au traitement d'images de documents. En effet, un caractère est composé d'un ensemble de pixels tous connectés les uns aux autres (exception faite de sa diacritique), et certaines photos ou illustrations le sont aussi parfois. De ce fait, il est relativement simple d'extraire les caractères (resp. les contours d'une photo) afin, par exemple, d'en étudier la taille ou encore la disposition. La figure **Fig. 2.9** illustre divers résultats obtenus sur des images de documents. Une exploitation d'informations tels que les alignements, les espaces entre composantes peut dans certains cas constituer une information relative aux lignes, aux paragraphes, aux colonnes...

Si l'information "composantes connexes" est populaire, ces dernières restent tout de même complexes à utiliser ou plutôt à analyser. En effet, leur extraction génère une grande quantité de composantes qui ont pour particularité de se chevaucher, d'être inclues les unes dans les autres, d'être sensibles aux bruits (**Fig. 2.9**)... Pour arriver à extraire la structure physique ,il faut être capable de fusionner ces composantes et c'est là qu'est toute la difficulté. Voici quelques exemples d'utilisation des composantes connexes :

- [MY01] étudie des caractéristiques bien spécifiques pour permettre une segmentation et une classification des composantes en 9 classes (texte, titres, texte vertical, photos, dessins...). Les règles d'affectation sont prédéfinies et se basent sur l'étude des tailles, de la densité de niveaux de gris, du nombre de pixels noirs successifs sur une ligne. Pour mener à bien cette



FIG. *2.9: Exemple de l'application d'un algorithme de composantes connexes sur des images de documents[Tri03]*

classification, cette méthode nécessite pas loin de 10 seuils à paramétrer manuellement. La fusion des composantes (en paragraphes pour le texte) est réalisée à l'aide d'une extraction des lignes de texte et de l'étude des distances moyennes entre composantes d'un même label.

- [AZ03] propose une segmentation texte/dessin en deux étapes. Dans un premier temps, un certain nombre d'images sert pour l'apprentissage. Pour cela, une personne segmente manuellement les différentes zones de l'image et affecte un label à chacune d'entre elles (texte ou non). Ensuite, un algorithme de segmentation de type XY-CUT (cf. section suivante) "sursegmente" l'image. Pour chaque zone segmentée, les auteurs extraient des informations des composantes connexes afin d'alimenter un perceptron multi-couches. Les caractéristiques servant à l'apprentissage et à la classification sont au nombre de 22. Parmi celles-ci, on retrouve des caractéristiques du type : aire d'une composante, aire commune entre deux composantes qui se touchent, proportion de composantes de même taille que celle analysée...
- [Bre02, Bre03] décrit une solution pour segmenter des documents dont la mise en page est relativement complexe (avec, entre autres, des textes sur plusieurs colonnes). L'extraction des composantes connexes est à la base de sa proposition. La première étape consiste à trouver les séparateurs de zones de textes, matérialisés par de grands espaces blancs (colonnes, marges...). Les auteurs proposent un algorithme basé sur la formation du plus grand rectangle blanc possible (horizontalement pour identifier les inter-lignes et verticalement pour identifier les séparateurs de colonnes). Une fois ces "obstacles" (comme les appelle l'auteur) détectés, une fonction à maximiser, calculée en fonction des composantes atteignables sans traverser ou sans toucher d'obstacles, permet de détecter des lignes de texte quelque soit leur orientation.
- [RBD06] se base sur l'expertise de l'utilisateur qui fournit au système des informations sur les tailles des composantes, pour permettre de dissocier des composantes de texte/dessin/bruit. La fusion des composantes est réalisée à l'aide de l'étude des niveaux de gris séparant les

composantes.

Discussion :
Si les résultats annoncés dans les différentes références trouvées démontrent bien la pertinence d'un tel outil, il faut néanmoins garder à l'esprit que selon les images analysées, il existe quelques limites à l'utilisation des composantes connexes. Ainsi, une binarisation doit être effectuée au préalable nécessitant des documents d'une bonne qualité (image bien numérisée et lisible). D'une manière générale, le principal défaut des approches ascendantes est que le traitement de l'information générée reste compliqué à traiter. L'extraction des composantes connexes en est l'exemple parfait. En effet, cet outil génère un grand nombre de composantes dont le tri et l'analyse sont généralement laissés à un algorithme qui analyse la taille des composantes, les espaces entre elles, leurs positions... La figure **Fig. 2.10** montre à quel point, sans l'intervention d'un expert ou sans modèle prédéfini, il est impossible d'effectuer un traitement automatique des images. Quelques essais effectués sur notre corpus ont fait remonter 3 particularités physiques de nos documents qui rendent complexe l'utilisation "tout automatique" de ce type d'approches :

1. Les espacements entre composantes sont irréguliers. Un exemple est donné par le point 1 de la figure **Fig. 2.10** où l'on voit que la distance qui sépare le corps de texte de la marge est égale à la distance séparant deux mots. Cette caractéristique récurrente de notre corpus rend compliquée l'utilisation de règles basées uniquement sur les espaces entre composantes.
2. Le bruit, très fréquent dans nos documents, se caractérise par un grand nombre de petites taches noires ou par l'apparition des caractères du verso de la page. La binarisation, nécéssaire pour l'extraction des composantes, peut amener à confondre des caractères et ces taches d'encre (idem pour les caractères du verso) étant donné que leur taille est relativement semblable (cf. point 2 de la figure **Fig. 2.10**)
3. Les illustrations qui émaillent les documents anciens sont des images de traits. Ces dernières ont pour caractéristiques d'être composées d'une multitude de petits segments pas toujours connectés entre eux (cf. point 3 figure **Fig. 2.10**). De ce fait, on se heurte à un double problème puisque non seulement le nombre de composantes va être important dès lors que l'illustration sera de grande taille, mais en plus cela va rendre l'analyse plus compliquée car il va falloir trouver des règles de décision permettant de discerner les différents cas d'inclusion de composantes existant (texte dans dessin, dessin dans texte, dessin dans dessin, ...)

On notera que [RBD06] a proposé une solution combinant une analyse des composantes et l'introduction de connaissances d'un expert, rendant ainsi possible la segmentation et l'analyse d'images de documents anciens par approche ascendante.

La seconde difficulté dans l'utilisation des composantes connexes, porte sur son inadaptation à traiter des mises en page complexes. La figure **Fig. 2.11** en donne deux exemples. Dans le premier (**Fig. 2.11**.a), la structure non rectangulaire du dessin fait que le rectangle englobant chevauche des composantes de texte qui du coup ne sont pas prises en compte (une règle de gestion des composantes était de supprimer les petites composantes inclues dans une grosse). De plus, les informations relatives à la composante de dessin (taille de la composante, position...) ne seront pas fidèles à la forme de cette dernière. Le deuxième exemple illustre (**Fig. 2.11**.b) une mise en page complexe où les relations d'inclusion entre les composantes de texte et de dessin deviennent complexes à traiter.

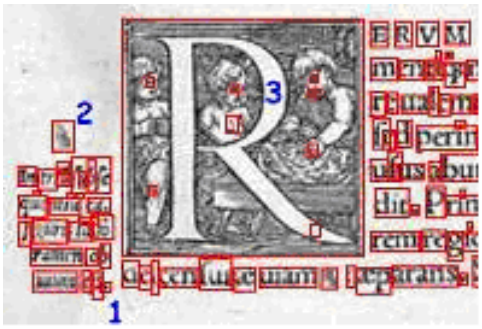


FIG. *2.10: Problèmes couramment rencontrés avec l'utilisation des composantes connexes*

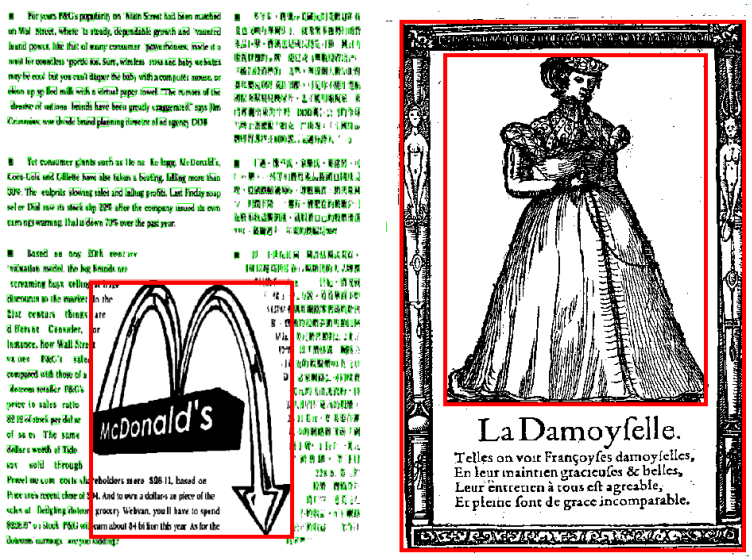
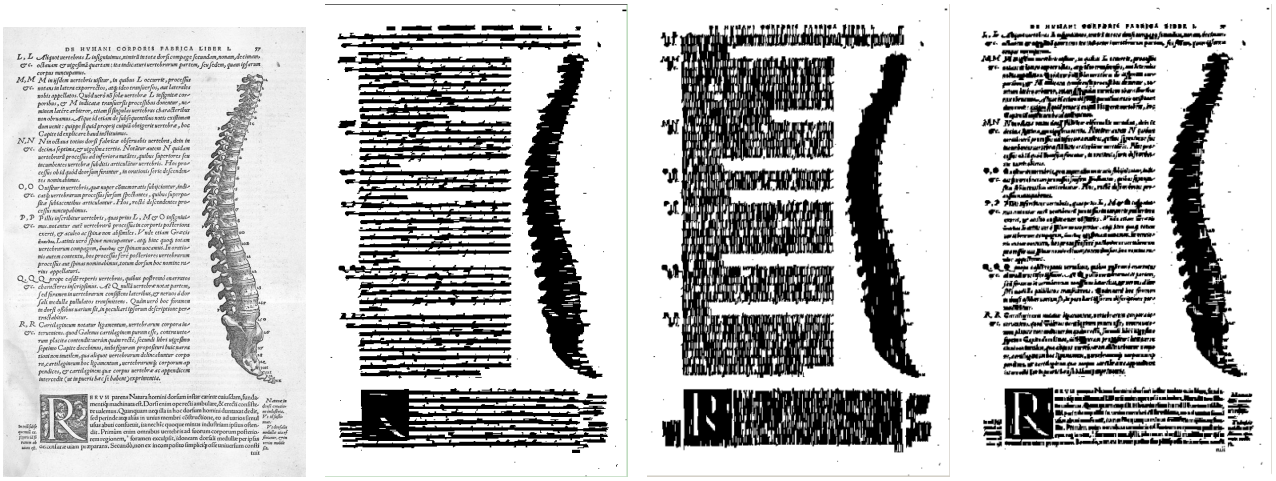


FIG. *2.11: Limite de la segmentation avec des boîtes englobantes*

b- L'algorithme RLSA

Proposé par [WCW82], l'algorithme RLSA (Run Length Smearing Algorithm) est à la base de nombreux travaux de séparation texte/dessin, de segmentation en lignes/paragraphes des caractères, d'OFR (Optical Font Recognition)... Il est très simple à mettre en place puisqu'il suffit d'analyser les séquences horizontales de pixels noirs/blancs. Plus précisément, si le nombre de pixels blancs qui séparent deux pixels noirs est inférieur à un certain seuil, alors tous ces pixels blancs deviennent noirs. Cet algorithme est applicable horizontalement (**Fig. 2.12**.a) et verticalement (**Fig. 2.12**.b). Si l'on effectue un *ET* logique entre les deux versions de RLSA, il est possible d'obtenir une séparation du texte en lettres ou mots ou lignes ou paragraphes... le tout dépendant de la valeur du seuil. La figure (**Fig. 2.12**.c) montre qu'il est possible de segmenter le texte en "mots".

En règle générale, l'algorithme RLSA fait partie d'une succession de traitements réalisés sur le document, il est rarement utilisé seul pour segmenter le texte. Dans [**?**] les auteurs, en complément d'une extraction des composantes connexes, extraient les lignes à l'aide de RLSA.



(a) image d'origine (b) RLSA horizontal (c) RLSA vertical (d) Et logique entre les deux RLSA

FIG. 2.12: Exemple de l'application de l'algorithme RLSA sur des images de documents

Dans un premier temps, c'est la taille des composantes qui permet d'étiqueter chacune d'entre elles en texte ou dessin. RLSA permet ensuite de fusionner les lettres en lignes. Le paramètre étant choisi pour fusionner des caractères de corps de texte, l'auteur fait face à un double problème récurrent avec cet outil : la non fusion des caractères de titres (espacement supérieur à celui du corps de texte) et la non prise en compte de la diacritique (les accents, points... ne sont pas détectés). Pour y remédier, l'auteur propose l'utilisation de 3 nouveaux seuils basés sur les distances supposées entre pixels dans ces cas de figure. Une fois les lignes obtenues la fusion en paragraphes est rendue possible par l'étude des espaces inter-lignes (une étude statistique classique, soit moyenne et écart-type).
Dans [XHW02], les auteurs combinent également RLSA et extraction de composantes connexes, mais cette fois-ci sur des documents asiatiques. Après un RLSA (seuil approximé à l'aide de projections horizontales), les composantes connexes sont extraites. La fusion s'opère en deux temps. Tout d'abord, les composantes sont étudiées pour voir si elles appartiennent à du texte horizontal ou vertical. C'est la taille des composantes qui va permettre ce classement. Fixé de manière heuristique, un premier seuil (en pixels) permet de déterminer si la longueur de la composante correspond à du texte horizontal ou vertical. Un deuxième seuil (basé sur la distance entre les composantes) permet d'identifier les composantes de type titre, ou image. Dans un deuxième temps, les auteurs cherchent à fusionner les composantes pour arriver à une segmentation en paragraphes. Pour ce faire un graphe est généré. Dans ce graphe, chaque composante est un noeud et chaque arc représente la distance entre deux composantes. Un calcul de graphe complet de taille minimal, associé à un algorithme visant à retirer les arcs qui ne remplissent pas certaines conditions, permet de définir les composantes à fusionner.

D'un point de vue général, cet outil s'adresse à des images de documents bien spécifiques où le contenu et la mise en page varient peu. Dans notre cas, la nature de nos images rend impossible l'utilisation de cet outil.

c- Analyse des espaces blancs

Des méthodes de segmentation par analyse des espaces blancs ont été proposées par [Ant98,

PZ91]. Cette technique de segmentation est particulièrement bien adaptée aux documents dont les zones sont clairement délimités et rectangulaires (type journaux ou documents scientifiques). Elle se base sur la détection et l'analyse des grandes zones d'espaces blancs. Un peu à la manière de ce qui se fait avec les composantes connexes, les auteurs se basent sur des propriétés physiques supposées (le blanc des marges, le blanc des inter-lignes, le blanc des espaces inter-lettres...) pour regrouper les plages blanches extraites.

Dans [CWS03], l'auteur propose une méthode de segmentation d'images de documents contemporains et une classification des blocs extraits (Texte/dessin puis Texte latin/texte asiatique). Pour segmenter en régions, c'est-à-dire une séparation en paragraphes et zones de dessins, l'auteur se base sur l'algorithme proposé par Kise dans [KYT96] et qui s'appuie sur une étude des plages blanches de l'image. Un premier algorithme génère une première version d'une segmentation donnant une multitude de zones homogènes répondant à un critère prédéfini (eg : pixels noirs qui ont 8 voisins blancs). [CWS03] propose une amélioration de cet algorithme afin de régler les problèmes liés aux zones inclues les unes dans les autres. Les décisions de classification se font à l'aide d'un réseau de neurones alimenté par trois types d'informations. Ainsi, pour chaque zone l'auteur calcule un indice lié à la distribution du gradient de l'image (pour 24 angles différents), un indice lié à un calcul d'une autocorrélation (pour deux angles), et un dernier indice basé sur le calcul de la complexité de Kolmogorov (étudie la répétitivité d'un motif dans une zone).

d- XY-CUT

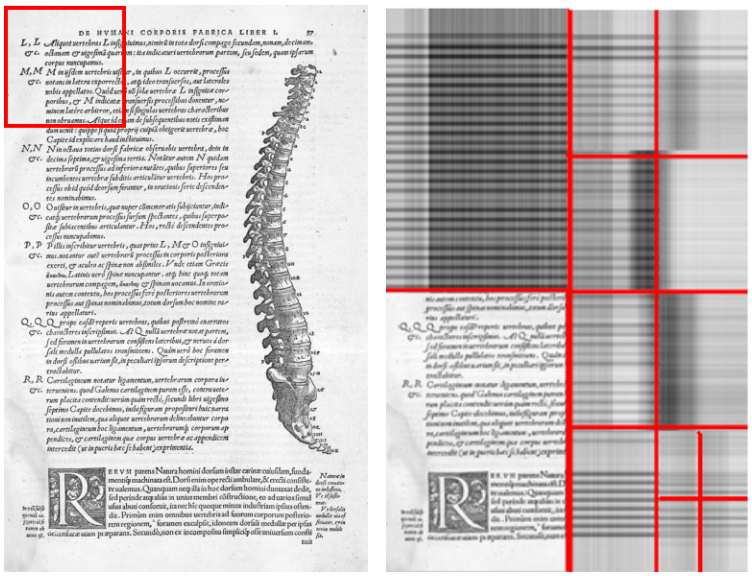
Le XY-cut fait partie des algorithmes par approche région. Le principe est d'appliquer récursivement le même algorithme sur une zone, tant qu'une condition n'est pas satisfaite. Le découpage peut, par exemple, être une division en quatre parties égales. Appliquer sur des images de documents, le XY-cut possède un double avantage :

1. Il est plutôt bien adapté à des images de type imprimés (formulaires, journaux, ouvrages, ...) qui sont composées en majorité de lignes de texte horizontales organisées en paragraphes et d'illustrations aux formes bien délimitées et séparées du texte. La figure **Fig. 2.13** montre le résultat de trois découpages récursifs. Pour chaque itération, on a calculé la moyenne des niveaux de gris en ligne et en colonne. Ainsi, plus une ligne est "foncée" plus cela signifie que la présence d'encre est prononcée. Dans cette figure, chaque quart de l'image correspond à un niveau de découpage.
2. Le partitionnement en quatre zones permet de modéliser de manière très simple la structure logique d'un document

A l'origine de plusieurs innovations en analyse de documents, Nagy proposait dès 1988 dans [NKK+88], une méthode de segmentation de caractères dans laquelle le découpage est relatif aux nombres de pixels successifs blancs (ou noirs) que l'on trouve en coupant verticalement (ou horizontalement). Ensuite la définition d'une grammaire qui analyse les caractéristiques du découpage récursif permet de segmenter et labelliser les documents scientifiques analysés.

l'auteur de [WS89] a effectué une comparaison des deux approches que sont RLSA et XY-CUT. A la suite de résultats expérimentaux, les auteurs annoncent que si l'objectif est d'extraire les lignes, alors dans ce cas RLSA est mieux adapté puisqu'il analyse des lignes de pixels (marche très bien pour le texte isolé). Si L'objectif est de segmenter en paragraphes alors, selon les auteurs, c'est l'algorithme XY-CUT (associé à une fusion des zones) qui est le mieux adapté. Ceci se justifie encore plus si la mise en page est complexe (plusieurs colonnes, illustrations non rectangulaires...). Toujours selon les auteurs, si les images ne sont pas parfaitement droites ces deux algorithmes deviennent déficients.

Nos tests ont fait ressortir deux informations. La première est qu'à l'instar de RLSA les



(a) images d'origine (b) 3 niveaux de récursivité

FIG. 2.13: *Exemple d'un découpage en XY-CUT (Pour chaque pixel, on calcule la moyenne des niveaux de gris en lignes et en colonnes)*

critères de découpage (et éventuellement de fusion) restent difficile à déterminer. L'autre information qui est ressortie, est que le XY-CUT revient à découper une image en un ensemble de rectangles, ce qui, comme on l'a évoqué pour les composantes connexes, n'est pas adapté aux documents de mise en page complexe. L'atout principal du XY-CUT est la simplicité qu'il offre pour modéliser la mise en page. En effet, la structure rectangulaire récursive permet de modéliser des relations de type "contenu/contenant" ou "à coté de" beaucoup plus simplement qu'avec des composantes connexes ou autre outils de segmentation.

2.2.2.2 Documents à structure variable

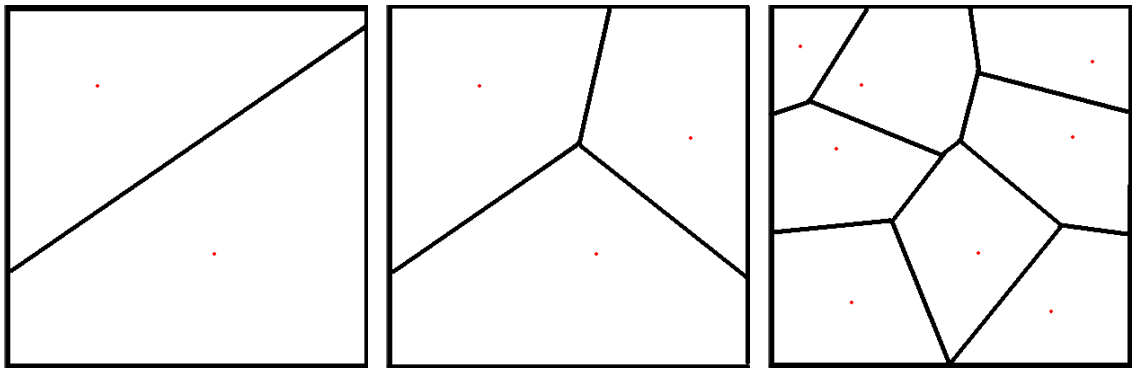
Certains corpus d'images ont la particularité de présenter de fortes variations en terme de structure. La difficulté de leur analyse tient donc principalement au fait de pouvoir utiliser des algorithmes qui soient suffisamment génériques. La section suivante propose de revenir sur certaines d'entre elles.

a- Diagramme de Voronoï

Le diagramme de Voronoï (mathématicien du XIXème siècle) est "une décomposition particulière d'un espace métrique déterminée par les distances à un ensemble discret d'objets de l'espace" (on trouvera une définition mathématique complète ici[9]). D'un point de vue géométrique cela se traduit par un partitionnement de l'espace en régions à partir de points caractéristiques disposés sur un plan. Cette opération revient à tracer les bissectrices de tous les segments formés par deux points. En rajoutant un point, une partition supplémentaire est créée. Le résultat final est obtenu par intersection des demi-plans créés (cf. **Fig. 2.14** pour des exemples).

Appliqué aux documents, le partitionnement par pavage de Voronoï permet de découper une

[9] http ://fr.wikipedia.org/wiki/Voronoi_diagram



(a) Bissectrice entre deux points *(b) Ajout d'un point* *(c) Voronoi pour 8 points*

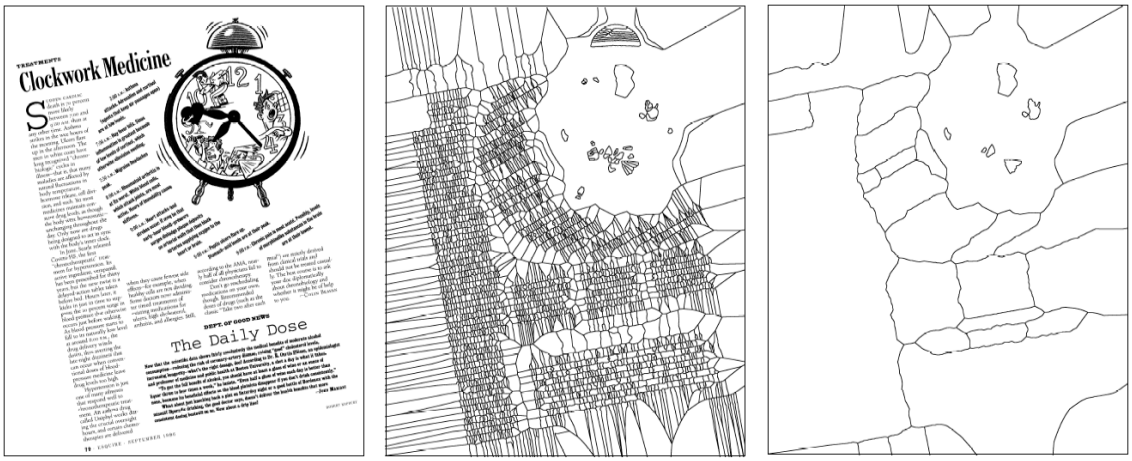
FIG. *2.14: Exemple de construction d'un pavage de Voronoï*

page de manière plus fine que ne le permet le XY-CUT : au lieu d'un découpage rectangulaire il est possible de découper les zones selon leurs contours et ainsi permettre la segmentation de documents dont la mise en page se trouve être complexe. Le problème est ici de définir les points participant au partitionnement.

Dans leurs articles, Kise et al. ([KIM99, KSM97, KSI98]) présentent une méthode de segmentation texte/dessin (avec les textes séparés en paragraphes). Pour cela, l'auteur extrait les composantes connexes. Ensuite, le choix d'un échantillon dans les points des contours des composantes, permet de calculer le pavage de Voronoï. Le graphe étant très dense, certains segments du pavage sont supprimés (une frontière du pavage totalement inclue dans une composante est supprimée). Il reste à fusionner l'ensemble des pavés. Pour cela un graphe de voisinage est généré. Dans ce graphe, un noeud correspond à une composante connexe et un arc entre deux noeuds est créé s'il existe une frontière commune entre ces deux composantes dans le pavage de Voronoï. Le principal intérêt de ce graphe, est qu'il permet de réduire les temps d'analyse des composantes en les limitant aux voisines de chaque composante.
L'objectif des auteurs est d'extraire les lignes de texte. Pour ce faire, une étude statistique des distances entre composantes et de la densité de chacune est réalisée. Cette étude permet d'émettre des hypothèses sur les espaces inter-letrres, inter-mots, inter-lignes et du fusionner les pavés en conséquence.

La figure **Fig. 2.15** illustre deux étapes du processus de segmentation.

Dans leur article [LWT04], les auteurs proposent uns solution très proche de celle de Kise et al. En effet, après avoir extrait les composantes connexes de l'image, les auteurs construisent un pavage de Voronoï et génèrent un graphe quasiment semblable à celui expliqué précédemment. La différence principale se situe au niveau des règles de fusion des différents pavés. Ici chaque composante est analysée séquentiellement. Pour chacune d'elles, on étudie les deux pavés les plus proches reliés par un arc et on cherche à déterminer si ces deux pavés correspondent à deux lettres d'un même mot ou pas. Les deux premiers indices sont des proportions de la distance de la composante étudiée avec respectivement la première et la deuxième plus proche. Le troisième indice est une distance moyenne séparant la première plus proche à la deuxième. Le dernier indice est une proportion relative à l'aire des deux composantes les plus proches de celle étudiée. Une étude sur une centaine de pages permet à l'auteur de fixer 4 seuils. Plus précisément, les



(a) Imiage d'origine

(b) Pavage de Voronoï à partir des composantes connexes

(c) Pavage de Voronoï après suppression de certains segments

FIG. 2.15: Segmentation de documents par pavage de Voronoï [KIM99]

auteurs (pour chaque indice) construisent un diagramme avec en abscisse la valeur observée et en ordonnée le nombre de fois que cette valeur est observée. Cela permet de fixer manuellement 4 seuils auxquels seront comparés les 4 indices extraits via 4 règles prédéfinies et qui finalement permettront de décider s'il faut fusionner deux pavés ou non.

Discussion : La construction non rectangulaire du pavage de Voronoï, a pour avantage de permettre une segmentation d'images dont la mise en page complexe est impossible avec des pavés rectangulaires. Cependant, la tâche consistant à déterminer les pavés à fusionner entre eux reste complexe. Il faut pour cela avoir une connaissance précise des polices et des styles utilisés, de la taille des images, de la qualité de numérisation. En effet, le diagramme de Voronoï reste avant tout un outil de découpage, dont l'utilisation se trouve être souvent combinée à celle des composantes connexes, et des défauts qui vont avec. Une autre utilisation consisterait à ne plus s'appuyer sur les composantes connexes. A notre connaissance, cette approche n'a pas été exploitée dans la littérature.

b- La multirésolution dans le document

Comme le dit l'auteur de [TZ00] : "la beauté de la multi-résolution, c'est qu'elle a tendance à grouper, au fur et à mesure, les caractères en mots les mots en lignes et les lignes en paragraphes". En effet, la segmentation par changement d'échelle (ou multirésolution) permet de percevoir et d'accéder à des structures de tailles différentes. Ce processus est comparable à celui mis en jeu par notre cerveau selon la distance à laquelle on regarderait une image. En baissant par exemple la résolution d'une image, on obtient une vision plus approximative de cette dernière ; on ne discerne pas l'information telle qu'elle existe mais plutôt une vision grossière (position des blocs , proportion des couleurs, certaines formes...). Appliqué aux images de documents, le changement d'échelle permet d'obtenir plusieurs versions d'une même image (cf.**Fig. 2.16**). Certains éléments très fins de la structure sont discernables à une résolution élevée dans la première image (notes de marges, polices, ...), mais ne le sont plus dans les versions suivantes où l'information disponible est celle de la présence de texte et de dessins et leurs positions relatives.

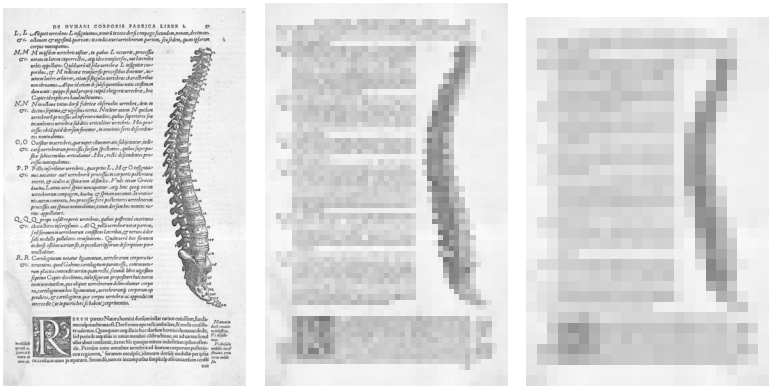


FIG. *2.16: Modification de la perception d'une image après changement de résolution*

De nombreux auteurs intègrent cet aspect multirésolution dans leurs approches de segmentations ascendantes ou descendantes. Dans [CLM98], les auteurs utilisent une pyramide de 4 niveaux de résolution pour permettre la reconnaissance de la structure physique de documents. Pour chaque résolution de l'image, les auteurs créent 4 cartes. Une fenêtre de taille 16X16 est déplacée sur les 4 images. A chaque itération, le pixel du centre de la fenêtre est remplacé par la moyenne des niveaux de gris des 15 voisins (carte 1), la variance (carte 2), la médiane (carte 3), le nombre de voisins dont le niveaux de gris est supérieur à un seuil prédéfini (carte 4). Ces 4 cartes sont à la base de la détection des entités d'une page. Le label de chaque zone est le fruit d'une analyse des cartes à différentes résolutions. Après le calcul des composantes connexes , les auteurs recherchent des caractéristiques de la distribution des niveaux de gris de chacune d'entre elles. Par exemple, une zone de fond est sensée avoir pour caractéristique une moyenne de niveaux de gris élevée (avec 255=Blanc) et une variance faible (zones homogènes). Une zone de texte est composée, à faible résolution, d'un grand nombre de pixels dont les niveaux de gris sont homogènes. Pour ce qui est des illustrations, se sont des composantes de grande tailles avec des caractéristiques précises de densité de niveau de gris. La dernière étape consiste en une fusion des données obtenues aux différentes résolutions. Les illustrations et le texte sont identifiés après extraction des composantes connexes pour deux résolutions de la pyramide. Si pour un label, le même nombre de composantes est détecté à deux résolutions différentes, alors l'analyse est finie, sinon la même comparaison est faite pour deux autres résolutions.

Les auteurs de [TZ00] utilisent également les avantages de la multirésolution. Ainsi, après avoir enlevé les grosses composantes connexes de l'image initiale, 9 niveaux de résolutions différents permettent de segmenter le texte en paragraphes et titres. La taille de l'image est divisée par deux à chaque fois, l'originale est donc l'image de plus grande taille. Une résolution de départ pour lancer le processus de segmentation est choisie de manière expérimentale (étude de la densité des niveaux de gris d'une résolution à une autre). Ensuite, l'application de 5 règles, basées sur une étude de la taille des composantes de leurs positions les unes par rapport aux autres et de l'évolution de ces caractéristiques d'une résolution à l'autre, permet d'identifier les paragraphes et les titres.

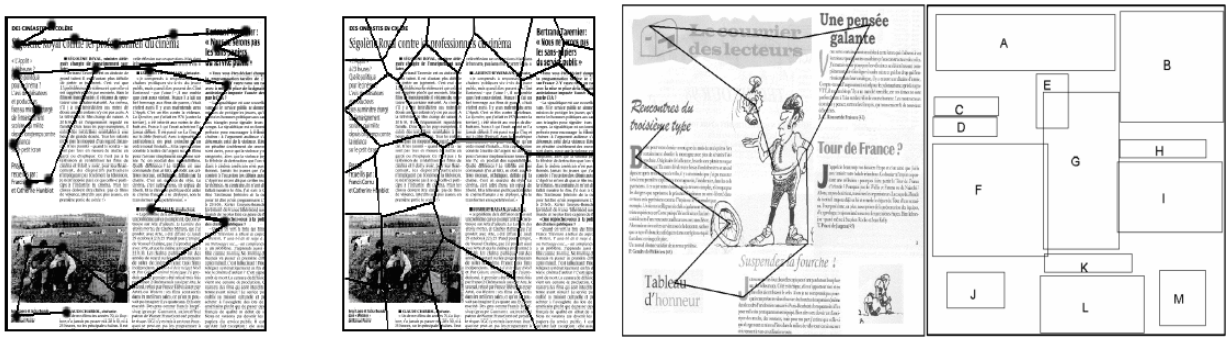
Dans [SG05, SG04] les auteurs combinent eux aussi une extraction des composantes connexes avec une analyse multirésolution. On retrouve le même type de raisonnement avec des décisions de segmentation/fusion prises en fonction de la distance entre composantes à différentes résolutions.

Discussion : En analyse de documents, la multirésolution apporte une information riche. En plus de la diversité des informations qui apparaissent selon la résolution à laquelle on regarde l'image, l'intérêt du multi-échelle se situe également dans l'alternative qu'elle apporte face aux systèmes d'analyse mono-résolution. En effet, chaque résolution apporte un point de vue différent de l'image à traiter.

c- Méthodes "hybrides"

On peut regrouper ici les approches généralement rangées dans les approches mixtes. En effet, elles combinent à la fois des connaissances de haut niveau et l'extraction de primitive de niveau pixel.

Parmi celles-ci on retrouve les approches psycho-visuelles. Dans [Egl98], l'auteur segmente des images de journaux en simulant le parcours visuel humain de lecture d'un document. A partir d'un point pris au hasard, l'auteur met en place un système permettant d'élire un point correspondant au second endroit de la page où un lecteur focaliserait son attention visuelle. En reproduisant plusieurs fois ce processus l'auteur simule un parcours visuel composé de plusieurs points d'intérêts. Le choix du prochain point se base sur le calcul de certains critères à fort pouvoir attractif pour l'oeil humain. Ainsi l'auteur met en jeu des indices relatifs aux contours, aux orientations, aux niveaux de gris... présents aux alentours du point étudié pour déterminer son successeur. Plusieurs techniques de segmentation utilisant ces points d'intérêts sont présentés par l'auteur. On peut citer par exemple la construction d'un diagramme de Voronoï **Fig. 2.17**.a. Cette séparation en région est assimilable à une vision préattentive. Il est également possible d'extraire les composantes connexes relatives aux points trouvés **Fig. 2.17**.ab



(a) Utilisation de Voronoï *(b) Découpage en zones*

FIG. *2.17: Segmentation d'images de documents par approche psycho-visuelles [Egl98]*

Dans [O'G93], est présentée une approche de segmentation originale basée sur l'analyse du "spectre" du document (quelque soit l'inclinaison). Le spectre est construit de la manière suivante : pour chaque composante connexe de l'image on répertorie sur un diagramme polaire, les distances et les angles relatifs au k plus proches voisins de la composante étudiée. Ce spectre permet d'identifier quatre nuages de points symétriques deux à deux. Le premier couple correspond aux espaces inter-mots et des distances inter-caractères. Le deuxième couple de nuages correspond aux espaces inter-lignes. A partir de ces données, on peut reformer la structure des mots puis celle des lignes de texte. Finalement les blocs sont fusionnées selon leur proximité.

Dans [FHA90], les auteurs utilisent une combinaison d'outils empruntés aux approches ascendantes et descendantes. En partant de la supposition que les espaces blancs entre les mots et

entre les lettres restent constants, ils parviennent à estimer les paramètres d'un RLSA vertical et horizontal. Dans un premier temps les auteurs estiment l'espace inter-ligne à l'aide de projections horizontales. Une analyse du fond par pavage permet ensuite de déterminer la distance inter-mot.

On peut également citer les auteurs de [ABG+03, DCES01, Egl98] qui adaptent le principe de filtrage par gradient sur des images binaires de documents (notamment en privilégiant le sens horizontal du filtrage). Appliqué sur du texte, le gradient va être fort pour des transitions encre/fond sur cette zone alors que sur des photos l'homogénéité des plages engendre un gradient plus faible et donc caractéristique (ce sont les contours des formes qui possèdent un fort gradient). Dans [YUA01] les auteurs utilisent, entre autre, le filtre de Canny et un gradient gaussien pour détecter les frontières des formes que représentent le texte et les photos. La fusion des frontières est réalisée à l'aide d'un algorithme combinant un filtrage bidirectionnel (horizontal et vertical) et des règles à respecter (eg. deux segments doivent se recouvrir pour être fusionnés, la distance entre deux segments doit être comprise entre deux seuils...).

2.2.3 Conclusion

Après comparaison de 6 méthodes de segmentation (XY-CUT, RLSA, recherche d'espaces blancs [KIM99], Voronoï, le docstrum d'Ogorman [O'G93] et l'algorithme de [Bre02]), les conclusions tirées par les auteurs de [SKB06] sont équivalentes aux nôtres. Ainsi, le XY-CUT et RLSA sont sensibles aux bruits et peu robustes aux textes inclinés. Les algorithmes de recherche d'espace blancs possèdent des critères complexes à paramétrer. D'un point de vue général, il se pose la question de la variation des tailles et des styles de polices d'un corpus. Il semble complexe d'éditer des règles de classification ou d'effectuer un apprentissage si l'on est pas certain de la stabilité des caractéristiques d'une image à l'autre. Dans le cadre des travaux sur des bases d'images de documents anciens, des données essentielles ne sont pas (ou difficilement) disponibles. Les informations liées aux propriétés de l'image (qualité de la numérisation, résolution des images..), mais aussi des caractéristiques plus haut niveau (mise en page, police utilisée, information sur le modèle...) en sont quelques exemples.

2.3 Analyse de documents : les approches textures

La section qui suit est organisée de la manière suivante : après avoir donné quelques définitions, nous détaillerons les outils les plus utilisés en analyse de texture d'images. Pour une majorité de ces outils, nous nous efforcerons de voir comment ils ont été (parfois) adaptés à l'analyse d'images de documents. Lorsque cela est possible, et que cela a du sens, nous testons certains de ces outils sur des images de notre corpus d'images de documents anciens. En effet, face aux lacunes des méthodes citées dans la section précédente, il est impératif de trouver de nouveaux outils qui permettent de segmenter, mais surtout de caractériser de manière générique des documents complexes comme ceux que nous traitons.

2.3.1 introduction

Véritable alternative aux méthodes décrites dans la section précédente, les approches textures permettent de palier à certains problèmes évoqués précédemment. Les outils d'extraction texture sont des outils de bas niveau : ils permettent d'extraire tout un ensemble d'informations sans

aucune connaissance nécessaire relative au contexte, à la sémantique ou aux caractéristiques physiques de l'image étudiée.

Il n'existe pas une mais plusieurs définitions de ce qu'est une texture. D'un point de vue formel, [All04] recense deux façons d'aborder cette problématique :

1. Approches déterministe : dans ce cas de figure, la texture est vue comme la répétition spatiale d'un symbole (notion de texton)
2. Approches probabilistes : la texture est vue comme un ensemble de microtextures présentant des primitives distribuées de manières précises.

Il est bien entendu impossible de lister l'ensemble des outils permettant de caractériser des textures d'images naturelles. Il existe un grand nombre d'états de l'art sur ce sujet ([Ros99, Lou00, BUR91, TJ98, JS95, TFMB04]).

Dans [TJ98] l'auteur présente 4 "familles" d'outils de caractérisation de texture. On distingue parmi elles :

- les approches spatiales que sont les méthodes statistiques, géométriques et à base de modèles.
- Les méthodes issues du traitement du signal sont, quant à elles, liées au domaine des fréquences. Dans son écrasante majorité, ces outils sont utilisés pour la segmentation ou la classification d'images naturelles qui ont la particularité d'être fortement texturées (images satellites, imagerie médicale...). Nous le verrons, certains auteurs ont utilisé ces outils pour la segmentation ou la caractérisation d'images de documents. Ce choix semble pertinent, puisque sous certaines conditions, une image de document possède elle aussi de fortes propriétés de texture.

A l'instar des approches ascendantes/descendantes/mixtes détaillées précédemment, l'objectif de la segmentation par approche texture consiste à séparer les différentes zones homogènes les unes des autres. Ces méthodes permettent, par exemple, de pouvoir séparer le texte des illustrations voire même de caractériser différents styles de textes ou de fontes. On trouvera de très bons états de l'art sur ce sujet spécifique de la segmentation d'images de documents par approche texture dans [OP00, All04]. Les techniques utilisées pour la localisation de texte dans des vidéos peuvent aussi fournir des informations intéressantes. Certes, les auteurs s'appuient généralement sur les mouvements et les couleurs, mais ils utilisent également des techniques très proches de ce qui se fait en analyse d'image. On trouvera un état de l'art relativement récent dans [JKJ04].

2.3.2 Méthodes statistiques

Parmi les grands classiques, il est impossible de ne pas citer les travaux d'Haralick et de Laws. La Grey Level Co-occurrence Matrix (GLCM) a été proposée par Haralick dans [HSD73]. La GLMC (P_d) est une matrice qui indique, dans une image I le nombre d'apparitions de couples de pixels ayant des niveaux de gris (i ,j) selon une direction et un déplacement donné ($d = (d_x, d_y)$) (**Eq. 2.1**).

$$P_d(i,j) = Card\{((r,s),(t,v)) : I(r,s) = i, I(t,v) = j\} \tag{2.1}$$

Avec (r,s),(t,v)$\in$ hauteur,largeur de I, $(t,v) = (r + d_x, s + d_y)$.

La GLMC ne permet pas, à elle seule, de caractériser une texture. Ce sont les indices que l'on va calculer sur cette matrice qui vont permettre de "signer" une texture. Il existe plus d'une dizaine d'attributs que [HB00] regroupe en 3 familles :

1. Les mesures de contrastes.
2. Les mesures d'énergie qui viennent renseigner sur la régularité des textures.
3. Les mesures statistiques qui donnent une idée de la répétitivité de la texture.

Dans [TJ98] les auteurs références les 5 mesures les plus couramment utilisées.

Caractéristique texture étudiée	*Formule*
Energie	$\sum_i \sum_j P_d^2(i,j)$
Entropie	$-\sum_i \sum_j P_d(i,j) \log P_d(i,j)$
Contraste	$\sum_i \sum_j (i-j)^2 P_d(i,j)$
Homogénéité	$\sum_i \sum_j \frac{P_d(i,j)}{1+\|i-j\|}$
Corrélation	$\frac{\sum_i \sum_j (i-\mu_x)(j-\mu_y) P_d(i,j)}{\sigma_x \sigma_y}$ Avec μ_x μ_y la moyenne et σ_x σ_y l'écart type de $P_d(x)$ et $P_d(y)$ et $P_d(x) = \sum_j P_d(x,j)$ et $P_d(y) = \sum_i P_d(i,y)$

Une autre méthode de caractérisation de texture basée sur le calcul de caractéristiques est celle de [Law80]. Cette méthode consiste à construire 25 versions d'une image texturée à l'aide de convolutions spatiales dont les filtres sont prédéterminés. Chacune de ces versions fait ressortir une caractéristique précise de la texture (présence de lignes horizontales, verticales,...). Les masques de convolution 2D sont générés à partir des masques 1D suivants :

L5=[1 4 6 4 1]
E5=[-1 -2 0 2 1]
S5=[-1 0 2 0 -1]
W5=[1 2 0 -2 1]
R5=[1 -4 6 -4 1]

A partir de ces noyaux on peut générer 25 filtres bi-dimensionnels en multipliant entre eux ces noyaux. L'extraction des caractéristiques texture s'effectue en 4 étapes :

- On applique ces 25 filtres sur l'image. Chaque pixel est donc décrit par 25 valeurs correspondant à chaque convolution effectuée. Soit $F_k(m,n)$ avec $k = 0, ..., 25$
- Pour chaque résultat de convolution, on calcule une énergie de texture avec une fenêtre d'analyse de taille 15x15 : $E_k(m,n) = \sum_{j=n-7}^{n+7} \sum_{i=m-7}^{i=m+7} |F_k(i,j)|$.
- Chaque matrice d'énergie est normalisée par le produit des deux maques $L5 * L5^t$
- On peut finalement calculer des caractéristiques textures ayant un sens bien précis avec, par exemple, le rapport de l'énergie des segments horizontaux sur celle des segments verticaux : $L5E5/E5L5$

Pour compléter ce court listing, on peut citer la matrice des longueur de plages qui se construit en recherchant des successions (plages) de pixels selon un niveau de gris et un angle précis. Un peu à la manière de la GLMC, on peut calculer des attributs sur cette matrice (importance des plages courtes, répartition des plages...).

La fonction d'autocorrélation est un autre outil couramment utilisé en analyse de texture. Elle permet d'obtenir des informations sur les caractéristiques de la texture (**Eq. 2.2**).

$$C_{xx}(k,l) = \sum_{k'=-\infty}^{+\infty} \sum_{l'=-\infty}^{+\infty} x(k',l').x(k'+k,l'+l) \tag{2.2}$$

Ainsi, si la texture est "grossière" (motifs larges) alors la fonction baisse lentement en augmentant la distance d'analyse (ici k' et l'). Au contraire, si la texture est plus fine (petits motifs peu espacés) alors la fonction décroît rapidement. Nous détaillerons plus finement cet outil dans le chapitre suivant.

Dans ses travaux, [Ros99] utilise certains de ces indices d'ordre statistique, pour mettre en place une méthode de segmentation d'images naturelles. Dans un premier temps l'auteur explique qu'il analyse l'image de manière globale afin de savoir si l'image est fortement texturée ou non. Pour cela il calcule, pour différentes résolutions de l'image, un indice d'homogénéité sur la GLMC. Dans un deuxième temps, l'auteur analyse les zones texturées via le calcul de 107 descripteurs (qu'il réduit à 9 après analyse de données). Au final, chaque pixel est décrit par un vecteur de caractéristiques qui va permettre de classer (donc segmenter) chaque pixel de l'image dans autant de classes qu'il y a de textures différentes (**Fig. 2.18**).

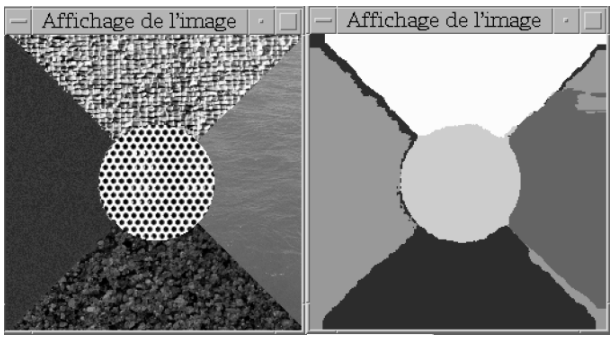


FIG. *2.18: Exemple d'un résultat de segmentation avec l'approche proposée par [Ros99]*

Dans [HOP+95] l'auteur propose une méthode d'analyse de textures basée, entre autres, sur le calcul de l'autocorrélation. A l'aide d'un masque de taille fixe, l'auteur calcule 4 mesures de textures différentes (autocorrélation linéaire, en fonction du rang des pixels du masque, moyenne et variance des niveaux de gris du masque) qui lui servent de descripteur de textures.

Dans [PA02] les auteurs proposent une méthode de signature d'image par approche texture combinant une transformation de l'image et l'utilisation des attributs d'Haralick.

Dans [ZTB04], les auteurs proposent une méthode d'indexation d'une base de photos de stèles. La première phase de l'indexation passe par une segmentation en régions qui se trouve être réalisée à l'aide des convolutions de Laws. Après combinaison des images énergies, les auteurs décrivent chaque pixel par 14 valeurs. Chaque pixel est donc décrit par un vecteur caractéristique. Les auteurs utilisent un kmeans à 8 classes (déterminé de manière heuristique) qui leur permet de segmenter leurs images avec une précision annoncée de 80%.

Discussion :

Si d'après [HB00], les approches d'ordre statistiques ont l'avantage d'être relativement simples à mettre en place et leur efficacité n'est plus à démontrer ; il n'en reste pas moins qu'elles possèdent des défauts non négligeables. En effet, d'une part le nombre de réglages manuels reste complexe à définir (taille des fenêtres d'analyse, direction de la matrice de cooccurrence, choix des caractéristiques à calculer...). D'autre part, la complexité des algorithmes mis en jeu rendent d'autant plus longs les temps de calcul quand le nombre de niveaux de gris différents augmente ou que l'on essaie de faire des calculs à haute résolution.

On peut également ajouter que ces outils basés sur l'étude statistique des niveaux de gris, semblent peu appropriés aux images de documents anciens. En effet, les techniques d'imprimerie de la Renaissance donnent un rendu d'image très proche d'une image binaire. Les seules variations de niveaux de gris sont dues à la numérisation ou à la dégradation du papier et de l'encre. On est donc très loin du panel de variation de niveaux de gris qu'on retrouve dans les images naturelles. De ce fait, les attributs d'Haralick ou de Laws ne semblent pas appropriés à la segmentation ou la caractérisation d'images de documents anciens.

2.3.3 Méthodes géométriques

Les méthodes géométriques correspondent particulièrement bien à la définition déterministe de ce qu'est une texture (formes + relations spatiales). Les références présentées ici cherchent donc à retrouver, exprimer, caractériser la notion de texton. Déjà présenté dans la section précédente, le maillage de Voronoï est très utile pour caractériser à la fois des régions de formes complexes et comment elles sont disposées entre elles. Les moments géométriques sont également utilisés pour permettre la description des motifs recherchés.

Dans [CM91a], les auteurs détaillent comment il est possible de segmenter une image texturée à l'aide du pavage de Voronoï. Dans un premier temps un petit nombre de points est choisi aléatoirement sur l'image pour calculer une première partition de l'image. Ensuite, pour chaque polygone est calculé une mesure d'homogénéité qui va permettre de décider, si oui ou non il faut rajouter des points dans ce polygone. Plusieurs critères sont proposés. Ils se basent sur l'étude des niveaux de gris des pixels (variance, écart type, valeur min/max des pixels...). Une fois les nouveaux points déterminés le diagramme est remis à jour. Enfin, un critère de convergence (tout est homogène ou la taille des polygones restant est trop petite) permet d'arrêter la construction du diagramme et une phase de fusion va décider quels polygones font parties de la même zone (au sens texture). La décision de fusionner deux polygones est le fruit de trois critères : les deux polygones sont homogènes, la différence entre les deux polygones doit être faible et enfin la somme de longueur de frontière commune doit être supérieure à un seuil prédéterminé.

Dans [TJ90], les auteurs proposent également une méthode de segmentation basée sur le calcul du maillage de Voronoï. Pour réaliser ce maillage, les auteurs recherchent des tokens (formes précises) pour ensuite y extraire des informations de forme des polygones générés. Ces caractéristiques sont utilisées pour labelliser chaque pixel et ainsi déterminer les frontières entre les régions.

Dans [Tuc94], les mêmes auteurs montrent qu'il est possible de segmenter des textures à l'aide du calcul des moments géométriques. Pour ce faire, une fenêtre est convoluée et déplacée sur l'image analysée. A chaque itération 6 moments calculés, permettent de générer 6 images résultats. Selon les auteurs, les moments ne sont pas suffisants pour discriminer tous les types de texture. C'est pour cela que les auteurs appliquent une transformation non linéaire de ces images. Au final, chaque pixel est décrit par un vecteur de dimension égale au nombre de moments calculés. Les auteurs utilisent un algorithme de classification avec échantillon (pour palier au problème de la taille des données) et où le nombre de clusters est un paramètre d'entrée.

Dans le monde du document, Voronoï n'est pas exactement une approche texture. C'est plutôt un outil qui permet un découpage plus fin que celui proposé, par exemple, par un XY-CUT. De ce fait, nous avons préféré tester un algorithme de segmentation à base de projections de pixels. Cette technique revient à caratériser l'image binarisée en projetant tous les pixels sur un axe et à rechercher un motif précis. En règle générale la projection est effectuée selon un axe horizontal ou

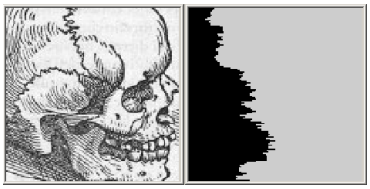
vertical ce qui permet d'identifier de manière très simple les grandes plages noires ou blanches. Sur des images que l'on suppose composées de dessins et de textes, on peut ainsi obtenir des motifs caractéristiques. Pour une zone de texte, le motif correspondant à une projection horizontale sera une alternance de pics noirs correspondant aux lignes de texte et aux espaces inter lignes **Fig. 2.19**.a. Le motif d'une projection verticale sera lui aussi une alternance de pics mais qui, cette fois seront caractéristiques des marges **Fig. 2.19**.b. Sur une zone de dessin une projection horizontale comme verticale donnera lieu à un motif difforme et irrégulier **Fig. 2.19**.c.



(a) *projections horizontales*



(b) *projections verticales*



(c) *projections horizontales sur une illustration*

FIG. *2.19: Application de projections sur des zones d'images de documents*

Nous allons voir, au travers de quelques exemples comment segmenter des documents à l'aide de projections. Dans certains cas, c'est un motif de projection que l'on recherche afin d'étiqueter chaque partie d'une image. Dans d'autres cas, la projection est un outil qui permet d'obtenir certaines informations qui seront utilisées par d'autres algorithmes (taille des lignes, taille espace inter-lignes...).

Les auteurs de [KRSG03] proposent une méthode de séparation texte/dessin de documents hébreux basée sur la construction d'histogrammes horizontaux. L'algorithme fonctionne en 7 étapes :

1. Binarisation de l'image.
2. Génération de l'histogramme horizontal de l'image.
3. Recherche des minima de l'histogramme (correspond aux lignes blanches)
4. Classification des blocs dans deux classes (reguliers/irreguliers). Après projection, les auteurs recherchent des caractéristiques précises pour classer les zones. Ces caractéristiques sont celles montrées dans la figure **Fig. 2.19**.a. Ce sont des informations comme la hauteur des minima, la largeur des pics ou encore la mesure des intervales entre les pics qui permettent de classer le blocs.
5. Les blocs réguliers sont identifiés en tant que texte et sont redécoupés avec une projection verticale
6. Pour les blocs identifiés comme irréguliers (qui peuvent contenir une illustration avec du texte sur les côtés), les auteurs appliquent un algorithme récursif basé sur des projections verticales pour séparer l'image du texte.

7. Une étude des positions relatives des blocs entre eux permet finalement de retrouver l'ordre de lecture de la page.

Sans évaluer quantitativement leur approche, les auteurs semblent satisfaits des résultats de segmentation obtenus sur des images de documents qui n'avaient pas encore été traités dans la littérature (Documents Devanagari). Cependant, ils notent toutefois certaines limites. Par exemple, les textes dont la mise en page est non rectangulaire ne peuvent pas être traités, une ligne noire horizontale peut être assimilée à une ligne de texte, les petits dessins ne sont pas identifiés...

Dans [CLKH96], l'auteur analyse des blocs prédécoupés dans le but de les classer, soit en tant que dessin soit en tant que texte. Les critères texture extraits sont issus d'une analyse des résultats de projections de pixels selon différents angles. Ce diagramme polaire est transformé en diagramme cartésien afin d'en extraire 6 critères (nombre de pics, intensité...). La classification des images analysées, est réalisée par un arbre de décision binaire après une phase d'apprentissage. Sa construction est proche du principe des treillis de gallois dont la théorie est expliquée dans [SBK05]. Ainsi, on part d'un noeud initial dans lequel tous les éléments servant à l'apprentissage sont labellisés. A chaque itération, un critère d'entropie permet de choisir la meilleure décision de séparation en deux classes. Ce processus est répété sur les noeuds fils jusqu'à ce qu'un critère d'arrêt prédéfini soit atteint.

Une approche originale est proposée par l'auteur de [Var04]. Il compare une segmentation texte/fond/dessin d'images postales basée sur les transformées en ondelettes à sa propre approche basée sur le calcul de 6 caractéristiques textures. Après application d'un pavage sur l'image, l'auteur calcule pour chaque pavé les caractéristiques suivantes :

1. La moyenne (μ) et écart-type (σ) de l'histogramme de niveaux de gris
2. Le nombre de pixels de l'histogramme inférieurs au seuil $\mu - k.\sigma$ avec k une constante fixée manuellement
3. La somme des coefficients du filtre de Savitzky-golay (horizontaux et verticaux). Ce filtre a normalement pour but de lisser une image, il est ici utilisé dans sa version 2D avec dérivée d'ordre 2 afin de détecter les segments horizontaux ou verticaux.
4. Des mesures de gradients (pour 4 orientation). L'auteur calcule des différences de niveaux de gris entre pixels selon un angle déterminé et propose une fonction d'énergie (somme au carré) pour signer le pavé étudié.

Dans ses conclusions, l'auteur montre qu'avec un arbre décisionnel dont les critères de chaque noeud sont basés sur les valeurs des 6 caractéristiques, il est possible de classer chaque pavé (fond/texte/dessin) avec un meilleur taux de reconnaissance que s'il utilisait les histogrammes des coefficients d'ondelettes de Haar. La signature par ondelettes se base sur le principe qu'appliqués sur un dessin les coefficients résultants ont une distribution de type laplacienne alors qu'ils sont concentrés sur certaines valeurs quant on les calcule sur du texte.

Comme l'énonce [KRSG03], les méthodes à base de projections ont une efficacité limitée. Un texte aux formes non rectangulaires, du texte mutli-orienté, des illustrations composées de peu de pixels sont autant de cas de figures où ce type d'approches textures peut devenir défaillante. Dans son article, [MEA02] effectue quasiment le même constat et propose donc, pour y remédier, une méthode de segmentation faisant appel à l'utilisateur pour corriger et donner des indications afin d'améliorer les résultats. Après une première analyse globale de l'image, via le calcul d'histogrammes, une première proposition de découpage est fournie à l'utilisateur. Ce dernier a alors la possibilité d'effectuer 3 corrections : scinder un bloc en deux (découpage horizontal ou

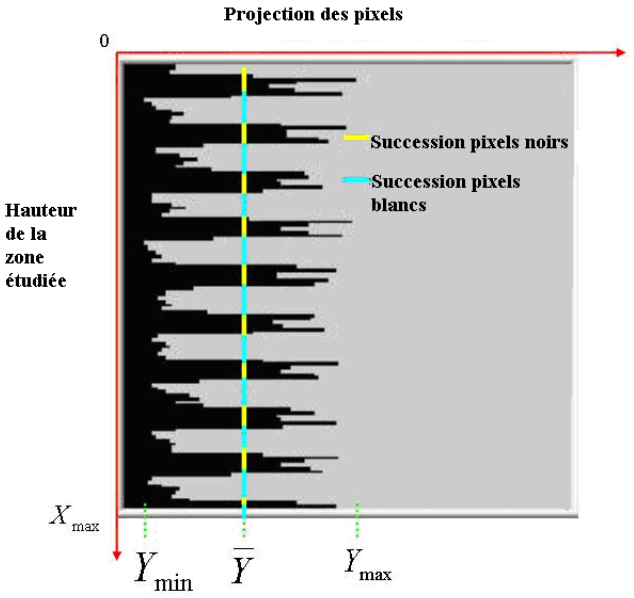
vertical), la dernière possibilité étant le regroupement de deux zones. La subtilité, est que l'opération indiquée par l'utilisateur permet, en fait, d'obtenir des informations sur le type d'erreurs obtenues lors de la segmentation (sur-segmentation ou sous-segmentation). L'information liée au type d'erreur et surtout la localisation du problème permet de décider où segmenter ou bien encore les blocs à fusionner. Cette phase de proposition/correction peut être vue comme une phase d'apprentissage. Le but est d'éditer des règles de segmentation qui seront appliquées sur le corpus entier. Cela permet ainsi d'arriver à des règles du type : "il faut séparer un bloc en deux verticalement quand sa longueur est entre 540 et 567 pixels et que sa largeur est entre 848 et 875 pixels". Testée sur des articles scientifiques, cette méthode interactive de segmentation permet, selon les auteurs, de segmenter plusieurs pages avec seulement 1% des opérations mal choisies automatiquement.

Discussion :

Nous avons testé une approche par projection inspirée de [KRSG03]. L'objectif consiste à chercher un motif significatif de la présence de texte : une alternance de pics et de creux (cf. **Fig. 2.19**.a). Pour cela, il est nécessaire d'analyser les histogrammes générés par la projection. Entre une approche XY-cut et une approche par fenêtre glissante, nous avons choisi la deuxième solution. A chaque itération la zone de l'image recouverte par la fenêtre est analysée. Tous les pixels dela fenêtre sont étiquetés "texte" ou "dessin" selon l'analyse de l'histogramme, puis la fenêtre est déplacée d'un pixel. Ce choix fait qu'un pixel est étiqueté plusieurs fois (pas forcément avec le même label). En fin de parcours un vote majoritaire permet d'attribuer le label final de chaque pixel.

L'algorithme d'analyse d'histogramme utilisé pour reconnaître le motif de projection recherché est basé sur l'étude des transitions noir/blanc des pics générés. Si ces transitions sont régulières alors le motif est celui d'un texte. La partie ci-dessous illustre l'algorithme implanté qui se déroule en 4 étapes :

1. On détermine Y_{max} qui est la hauteur maximale d'un pic
2. On détermine Y_{min} qui est le premier Y où la présence de pixels blancs est détectée
3. $\bar{Y} = \frac{Y_{max} - Ymin}{2}$
4. La zone étudiée est de type texte si la variance des plages noires est inférieure à un seuil α et que la variance des plages blanches est inférieure à un seuil β.



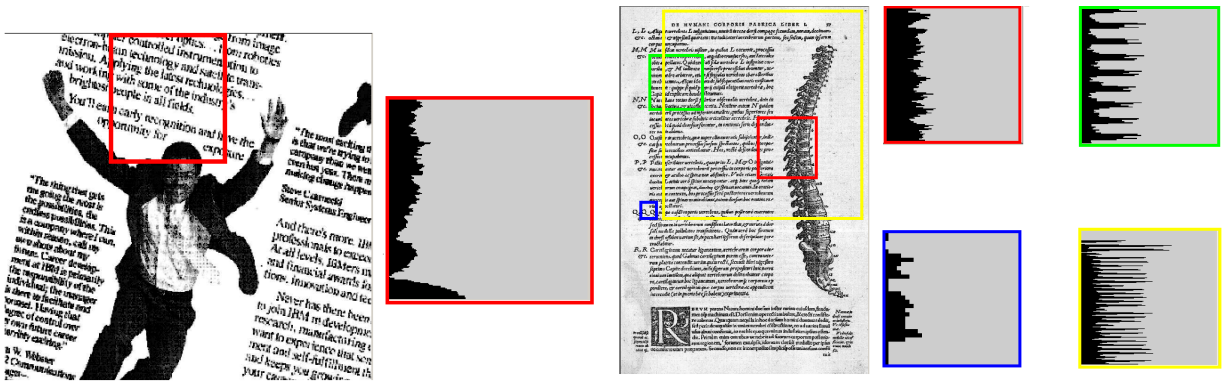
Dès lors que les paramètres sont correctement déterminés (α et β), on obtient de bons résultats avec quelques erreurs d'identification au niveau des zones de transitions (cf. **Fig. 2.20**).

Ces bons résultats sont malheureusement trompeurs, puisque cette méthode se heurte à deux



FIG. *2.20: Segmentation texte/dessin par une méthode de projection horizontale (texte en bleu, dessin en rouge)*

problèmes de taille. Le premier est que la segmentation par projection n'est absolument pas robuste lorsque l'image présente une légère rotation. La figure **Fig. 2.21**.a montre que même avec une faible inclinaison l'allure de l'histogramme n'est plus caractéristique. L'autre problème est lié au choix de la taille de la fenêtre. La figure **Fig. 2.21**.b montre à quel point la taille de la fenêtre d'analyse est un paramètre crucial. il se trouve être lié à la taille de l'image, à la taille des caractères, à la mise en page... Cette contrainte oblige à devoir remanier ce paramètre dès que l'une de ces caractéristiques change.



(a) Problèmes liés aux légères rotations

(b) Importance du bon choix de la taille de la fenêtre

FIG. *2.21: Segmentation texte/dessin par une méthode de projection horizontale*

2.3.4 Méthodes à base de modèles probabilistes

Dans son état de l'art, [TJ98] définit les méthodes de segmentation texture à base de modèles comme étant *"celles se basant sur la construction d'un modèle d'image permettant non seulement de décrire une texture mais aussi d'en générer"*. Les Champs de Markov et les fractales sont les deux outils de cette catégorie les plus utilisés. Lorsqu'ils sont utilisés, les champs de Markov reviennent à considérer l'image comme étant une réalisation d'un champ aléatoire dans un voisinage de pixels. On considère que l'intensité d'un pixel est directement liée à celle de ces voisins (l'hypothèse d'une distribution gaussienne est celle qui est la plus utilisée). Il est possible d'évaluer la probabilité d'un niveau de gris x_t d'un pixel en fonction des voisins R_t(on étudie généralement les 4 ou 8 voisins d'ordre 1 ou 2). On a donc la probabilité de l'état suivant $P(x_t|R_t) = \frac{1}{Z_t} e^{-w(x_t|R_t)^T \theta}$, avec Z_t une constante de normalisation relative à l'ensemble des probabilités des niveaux de gris de l'image. Le calcul du vecteur $w(x_t, R_t)$ peut être réalisé de différentes façons. [TJ98] en présente 2 : le modèle de Derin-Elliott qui utilise une fonction indicatrice et le modèle auto-binomial qui étudie tout simplement les niveaux de gris des voisins du pixel. Le vecteur θ est un paramètre qui sert à définir les propriétés texture de l'image. Plusieurs méthodes d'attribution de label sont décrites dans [Ros99].

La dimension fractale est, quant à elle, utilisée pour mesurer la rugosité d'une texture et la répétitivité (spatiale ou à différentes résolution) d'un motif.

Voici quelques travaux se basant sur des méthodes à base de modèles :

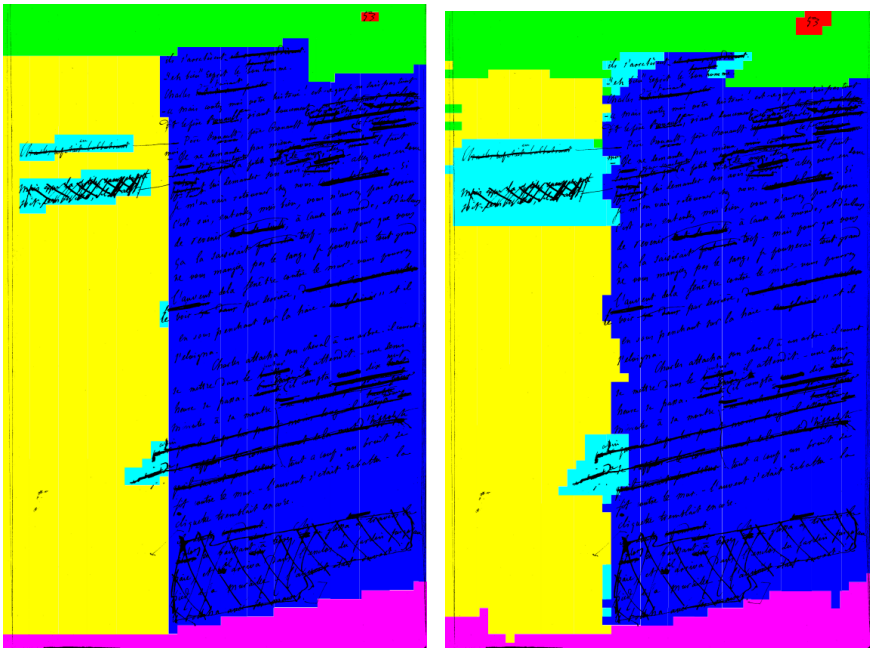
Dans [CCMV03], les auteurs utilisent la loi de Zipf pour identifier des zones d'intérêts dans une image naturelle. Très couramment utilisée pour l'analyse statistique de texte, la loi de Zipf est ici appliquée au domaine de l'analyse d'image. Pour cela, les auteurs associent à chaque bloc de taille 3X3 pixels, 9 lettres qui vont former un "mot" dont on va étudier la fréquence d'apparition avec la loi de Zipf. Pour limiter le nombre de combinaisons les auteurs partitionnent l'échelle des niveaux de gris en 9 (soit 9^9 motifs possible). En règle générale, les motifs les plus fréquents correspondent à des zones homogènes et les motifs appartenant aux contours sont moins fréquents. Sur la base de ce constat, les auteurs partitionnent l'image en un ensemble d'imagettes qui vont être analysées avec cette loi. Trois méthodes de détections sont ensuite comparées (étude des motifs fréquents, étude des motifs peu fréquents, étude de l'entropie de l'information) dans le but de déterminer les imagettes faisant partie d'une zone d'intérêt. La figure **Fig. 2.22** montre le type de résultats obtenus.



FIG. *2.22: Détection de zones d'intérêt par l'utilisation d'une loi puissance [CCMV03]*

Dans [NKPH06], les auteurs utilisent des HMM pour segmenter des images de documents manuscrits en zones d'intérêts labellisées (lignes de texte, rayures, notes de marge...). L'image est

découpée en une multitude de zones, afin de leur affecter le label le plus probable. Ce problème est non trivial dans le sens où cela revient à trouver une solution optimale dans une situation pour laquelle une recherche itérative est impossible. Le choix des auteurs s'est porté sur un modèle gaussien qui permet (à l'aide d'un apprentissage) de fixer les paramètres nécessaires à l'établissement du modèle. Les informations bas niveau extraites pour chaque pixel étudié, sont relatives aux densités de niveaux de gris du voisinage du pixel étudié. Ces choix permettent aux auteurs de segmenter les notes manuscrites de Flaubert qui ont la particularité de contenir de nombreuses hachures et ratures rendant les approches classiques peu performantes. La figure **Fig. 2.23** illustre les résultats obtenus par les auteurs.



(a) Vérité terrain *(b) Résultat de segmentation*

FIG. *2.23: Segmentation d'images de documents manuscrits[NKPH06]*

A notre connaissance, il existe très peu de travaux dans la littérature s'appuyant sur ces méthodes en vue d'une segmentation ou d'une caractérisation des pages. Les temps de calcul nécessaires aux traitements, les difficultés liées à la phase d'apprentissage, une faible diversité des indices bas niveaux calculables, sont selon nous quelques exemples des raisons qui rendent les approches par modèle probabiliste complexes à mettre en place.

2.3.5 Méthodes d'ordre fréquentielles

Les méthodes basées sur l'utilisation de primitives issues du traitement du signal sont idéales pour permettre la caractérisation de textures. En effet, ces outils permettent de détecter des caractéristiques de fréquences et d'orientations ; caractéristiques qui se trouvent être l'essence même de la définition d'une texture.

Ces outils fonctionnent dans le domaine fréquentiel. Les transformées de Fourier, Gabor ou ondelettes sont largement utilisées dans la littérature de l'indexation et segmentation d'images

naturelles. Dans la suite de ce chapitre, nous avons plus particulièrement porté notre attention sur les filtres de Gabor. Ainsi, nous présenterons tout d'abord la théorie relative à ce type de filtres. Ensuite, nous détaillerons plusieurs méthodes utilisant des méthodes fréquentielles et nous décrirons comment nous nous en sommes inspirés pour la réalisation de tests de segmentation sur notre coprus.

Un filtre de Gabor est un filtre linéaire qui est défini par le produit entre un filtre Gaussien et une sinusoïdale orientée **Eq. 2.3**

$$h(x,y) = g(x',y') \exp j2\pi Ux \tag{2.3}$$

Les variables x' et y' sont les coordonnées, après rotation d'un angle θ, du pixel (x,y). U est la fréquence en cycle/image.

$$g(x,y) = \frac{1}{2\pi\sigma_x\sigma_x} exp(-\frac{1}{2}[(\frac{x}{\sigma_x})^2 + (\frac{y}{\sigma_x})^2]) \tag{2.4}$$

Avec σ_x et σ_y les écarts types le long des axes de la Gaussienne.

La réponse E d'un filtre de Gabor pour un pixel (i,j) pour une fréquence U_l et un angle θ_k est : $E(U_l,\theta_k) = \sqrt{e(U_l,\theta_k)^2 + o(U_l,\theta_k)^2}$ ou e et o sont les sorties cosinus et sinus du filtre.

Dans [SS01], les auteurs combinent la transformée en ondelette et le spectre de Fourier pour permettre l'indexation de bases d'images. Le principe reste globalement le même dans la plupart des références citées dans cette section. Après avoir transformé l'image, des coefficients sont utilisés pour "résumer" ou "signer" l'image (coefficients de la transformée cosinus, de la transformée en ondelette...). Les auteurs construisent un vecteur caractéristique issu de l'utilisation d'une transformée de Fourier et d'une transformation en ondelette. Chaque image possède le sien. Les auteurs utilisent ensuite trois métriques différentes pour comparer les images.

Plutôt que de calculer la transformée sur l'image entière, [Lou00] propose une méthode basée sur la détection de points saillants. Ainsi, l'image ne sera décrite que par les caractéristiques calculées aux points correspondant aux parties les plus discriminantes de l'image. Les auteurs proposent une extraction de points basée sur l'étude des coefficients d'ondelettes. Pour cela, les auteurs partent du principe *"qu'un coefficient d'ondelette important à une échelle globale correspond à une région avec des variations globales"*. Les points sont trouvés en étudiant l'évolution des coefficients à des résolutions plus fines. Un exemple est donné dans la figure **Fig. 2.24**.
Une extraction de caractéristiques est effectuée en chacun de ces points. En plus des informations extraites de l'histogramme des couleurs et d'indices relatifs aux formes, les auteurs utilisent des indices de texture. Ainsi, un banc de filtre de Gabor (3 fréquences et 4 orientations) est calculé dans le voisinage du pixel étudié. Selon les paramètres du filtre, la réponse générée donnera des indications sur les orientations présentes à une fréquence donnée. L'ensemble de ces informations permet de signer l'image avec un vecteur caractéristique de dimension 9.

Dans [MM96b, MM96a] les auteurs utilisent également des filtres de Gabor.

Les auteurs calculent les réponses au filtre de Gabor pour plusieurs résolutions. Après chaque transformation, la moyenne et l'écart type des coefficients calculés sont extraits. le banc de filtres est calculé pour 4 résolutions, 6 orientations. La mesure de similarité entre deux vecteurs, est la somme totale de la différence terme à terme des moyennes et écarts types du vecteur. Sur une base constituée de 116 classes de textures différentes (soit 1800 images), la qualité du classement (retrouver les images appartenant à la classe de l'image requête) est de l'ordre de 74% pour un top 15 et de 92% pour un top 100.



FIG. *2.24: Détection de points d'intérêts avec utilisation de coefficients d'ondelettes [Lou00]*

Certaines méthodes de segmentation d'images de documents, s'appuient également sur des outils issus du monde du traitement du signal. Ainsi, les chercheurs exploitent généralement le fait que les zones de texte, du fait du grand nombre de transition encre/papier, sont caractérisées par de hautes fréquences, alors que les images sont composées de grandes zones homogènes et donc associées à des fréquences faibles. Parmi celles-ci on retrouve le binôme Gabor/Ondelettes. Ils permettent de fournir des indices qui vont différer selon des paramètres liées aux fréquences et aux orientations. Le choix de ces paramètres est relativement complexe car il est lié au contenu de l'image, à la résolution, à l'orientation des caractères, à la dimension de l'image... Plutôt que de devoir définir ces paramètres pour chaque image, les auteurs utilisent la plupart du temps des bancs de filtres qui permettent de faire varier les fréquences et les orientations avec pour chaque combinaison le calcul d'une énergie (relative à la réponse du filtre) et qui donne une indication sur le contenu de l'image.

Dans la section qui suit, nous présenterons quelques méthodes s'appuyant sur ces outils et nous verrons, dans la section suivante, dans quelles mesures elles sont applicables aux images de documents anciens.

Quelques méthodes utilisant les coefficients de la transformée en ondelettes :

- Dans [EDC97] les auteurs expliquent qu'il est possible de séparer les zones de dessins de celles de texte en appliquant à différentes résolutions de l'image la transformée en ondelettes. En partant de l'image basse résolution, ils appliquent et déplacent une fenêtre de taille prédéfinie sur l'image et calculent (après transformation par ondelette) une fonction d'énergie. Les auteurs proposent ensuite un système de classification floue permettant de déterminer la classe de chaque pixel (dessin/texte/photo).
- [GVCJ00] propose une approche relativement similaire. Pour différentes résolutions d'une image, il calcule une fonction d'énergie (somme au carré des coefficients de la transformée). L'auteur montre clairement que ces diagrammes diffèrent selon la zone analysée. Sur ce principe, 4 moments sont calculés et lui permettent de caractériser la forme des diagrammes. Une séparation texte/non texte est ainsi réalisée.
- Dans [LG00] les auteurs utilisent, eux aussi, la transformée en ondelette à différentes résolutions mais cette fois-ci ils proposent une alternative aux classiques moments calculés sur les coefficients. Ils proposent deux caractéristiques : la première est définie en fonction d'une étude des résultats des coefficients des ondelettes à haute résolution (qui d'après les auteurs

sont sensés suivre une distribution laplacienne quand ils sont calculés sur une photographie). La deuxième hypothèse est, qu'à haute résolution, les coefficients sont concentrés quasi-uniquement autour de 3 ou 4 valeurs lorsque l'image est composée uniquement de texte, alors que dans le cas où il ne sagit que d'une photographie les coefficients sont concentrés autour d'une seule valeur.

Nous allons le voir, les principales différences, dans le cadre de l'analyse d'images de documents, se situent au niveau du choix des paramètres du filtre et des post-traitements.

Voici quelques méthodes utilisant des bancs de filtres de Gabor :

- Dans [CC01], les auteurs décrivent un banc de filtres de Gabor composé de 4 orientations et de trois (hautes) fréquences. Les paramètres liés aux bancs sont calculés en fonction de la taille de l'image et de résultats de tests effectués manuellement. Pour chaque pixel (utilisation d'une fenêtre glissante), une énergie relative à la somme des énergies locales est calculée pour chaque réponse du banc $\sum_{l=1}^{3}\sum_{k=1}^{4} E(U_l, \theta_k)$. Finalement, cette énergie est comparée à un seuil fixé de manière heuristique (le seuil est égal à 40% de l'énergie locale) et permet ainsi de séparer le texte du dessin.
- On retrouve exactement le même fonctionnement dans [BSN04] et [RPR05]. En ce qui concerne [RPR05], les indices de Gabor sont calculés avec $\theta = 0, 45, 90, 135$ *et* $U =$ 0.2 0.3 0.5. Des informations relatives aux couleurs et aux composantes connexes extraites viennent compléter ces informations.
- [WMR97] calcule 9 caractéristiques avec son banc de filtres. A la différence de ce que nous avons déjà vu, l'auteur de [WMR97] construit un vecteur caractéristique (de dimension 9) pour chaque pixel de l'image et s'en sert pour alimenter un algorithme de classification des pixels (un 3ppv).

Discussion :

Les filtres de Gabor (ou ondelettes) semblent bien adaptés dans le cadre d'une analyse texture des images de documents. Afin de tester leur efficacité sur les documents anciens et permettre une perception de la structure quelque soit l'orientation ou les caractéristiques du texte et des dessins, nous avons choisi de nous inspirer des algorithmes présentés dans [BSN04, CC01, RPR05]. Toute la difficulté réside dans le bon choix de la forme du filtre et des paramètres de fréquences et orientations. Dans [IKK05], les auteurs décrivent comment paramétrer au mieux un banc de filtre de Gabor afin de recouvrir l'espace des fréquences. Les auteurs conseillent, entre autres, de choisir les orientations θ_i en respectant la formule $\theta_l = \frac{l2\pi}{n}$ avec $l = \{0, ..., n-1\}$ et n le nombre d'orientations choisies. Les fréquences f_i sont déterminées selon $f_i = (\sqrt{2})^{-i} f_{max}$ avec f_{max} la fréquence maximum désirée et $f_i = \{0, ..., m-1\}$ le nombre de fréquences désirées. Pour les autres paramètres nous nous sommes directement inspirés des articles cités précédemment. Le banc de filtres a été testé sur une vingtaine de documents contemporains et une vingtaine de documents anciens. Afin d'évaluer les performances du banc, nous avons volontairement utilisé des images de taille, d'origine et de contenu différents. Ces caractéristiques, combinées aux recommandations trouvées dans la littérature, nous a amené à construire un filtre composé de 5 orientations $\theta_l = \{0^\circ, 30^\circ, 60^\circ, 90^\circ, 120^\circ\}$ et de 6 fréquences $f_i = \{1, 2\sqrt{2}, 4, 32\sqrt{2}, 64\sqrt{2}, 128\sqrt{2}\}$. Après application du banc de filtres chaque pixel est décrit par 30 caractéristiques. Nous avons alors soumis ces données une classification de type Kmeans (avec distance euclidienne) et un nombre d'itérations maximum fixé de manière heuristique à 500. En pratique, le nombre d'itérations est faible (en moyenne il en faut moins de 15). Le nombre de classes est fixé à 3 afin d'évaluer la capacité de séparation texte/fond/dessin.

La figure **Fig. 2.25**, montre quelques résultats obtenus sur des documents contemporains. La figure **Fig. 2.26** montre deux résultats obtenus sur des documents un peu plus complexes (surtout d'un point de vue mise en page). Ces documents sont ceux qu'utilisent les auteurs de [EDC97] et dont la segmentation est réalisée par une approche analogue à la notre (ondelettes + multirésolution). La qualité de résultats obtenue correspond à celle décrite dans la littérature. Les zones de textes (quelque soit leur orientation) sont correctement détectées et il en va de même pour les photos. Les erreurs principales proviennent des zones de titres dont les caractéristiques fréquentielles sont plus proches de celles d'une photo que de celles d'une zone de texte (peu de transitions). Le fait que l'image soit traitée en niveaux de gris est à l'origine du mauvais classement des pixels clairs (du fond de la photo) qui sont considérés comme étant des pixels de fond.



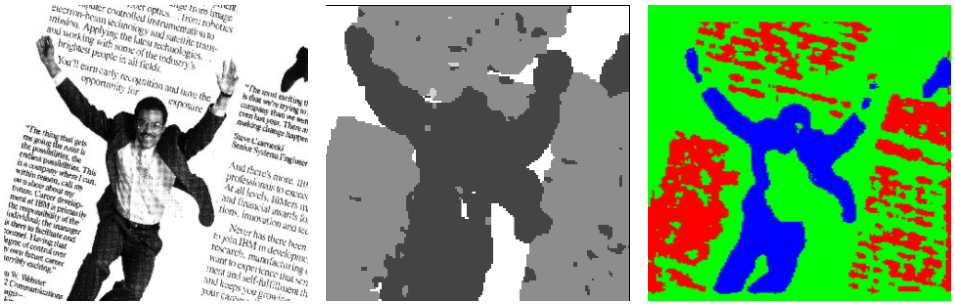
(a) image d'origine *(b) Résultat obtenu par filtres de Gabor* *(c) image d'origine* *(d) Résultat obtenu par filtres de Gabor*

FIG. *2.25: Segmentation texte/dessin avec Gabor*

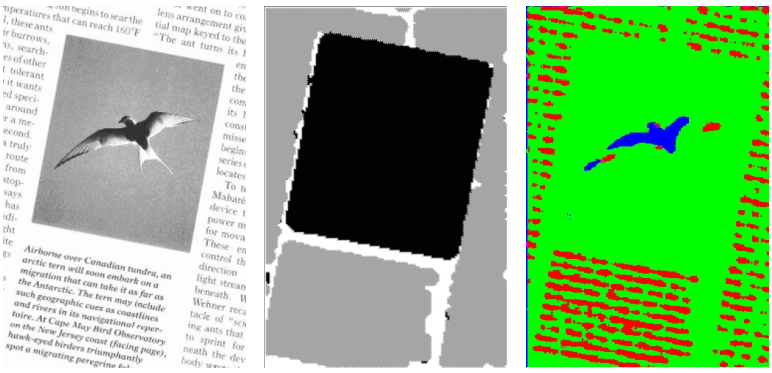
Nous avons appliqué le même banc de filtres sur des images de documents anciens. La figure **Fig. 2.27** montre deux résultats de classification résumant la qualité des résultats obtenus sur les documents anciens. La figure **Fig. 2.28** permet de détailler le problème récurrent lié à la détection de dessins de traits. En effet, si la détection de zones de textes (multi-orientées) ne pose pas de problèmes, c'est au niveau des illustrations que les erreurs de classification sont visibles. Il se trouve que les dessins de traits sont composés d'une multitude de petits segments plus ou moins rapprochés les uns des autres selon l'effet désiré par le concepteur. Cette caractéristique physique se traduit par une forte quantité de transitions entre l'encre et le fond, ce qui est synonyme de hautes fréquences. On le voit sur la figure **Fig. 2.28**, les deux blasons (composés de peu de transitions) sont globalement bien reconnus en tant que dessins alors que les deux lettrines sont assimilées à du texte.

2.3.6 Conclusion

Cette section a permis de mettre en évidence l'intérêt de l'utilisation d'outils textures. Selon nous, le principal avantage se situe dans la plus grande généricité que peuvent offrir ces outils. En effet, le fait qu'ils utilisent principalement des informations (très) bas niveau, permet de s'affranchir d'un bon nombre de connaissances a priori qu'utilisent les méthodes ascendantes/descendantes/mixtes. Parmi les autres avantages, on peut citer le fait que dans la plupart

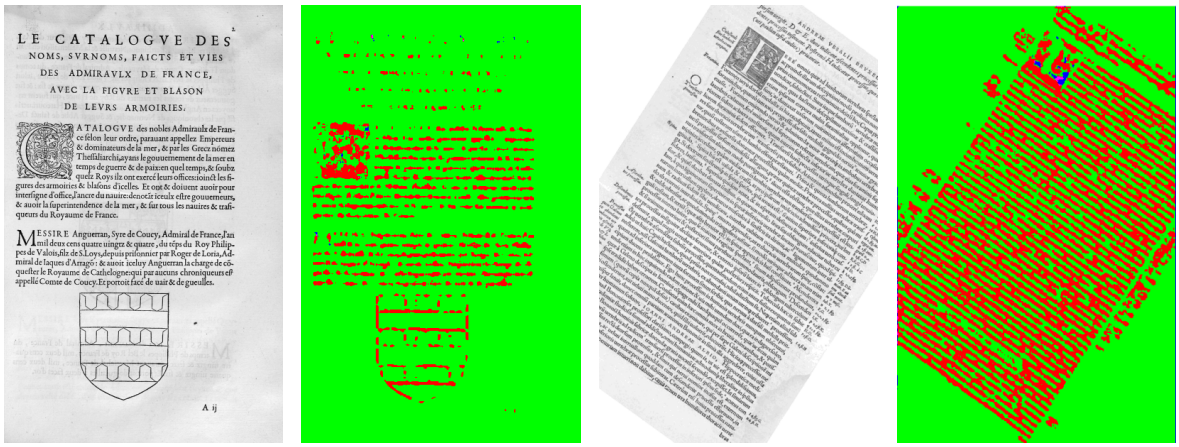


(a) image d'origine (b) Résultat obtenu par [EDC97] (c) Résultat obtenu par filtres de Gabor



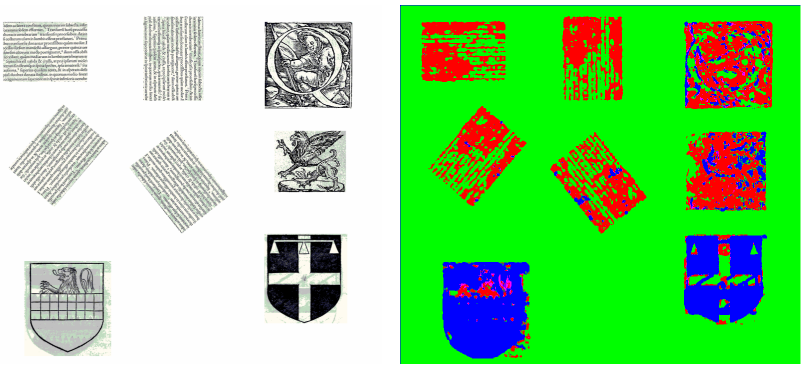
(d) image d'origine (e) Résultat obtenu par [EDC97] (f) Résultat obtenu par filtres de Gabor

FIG. 2.26: Segmentation texte/dessin avec Gabor



(a) image d'origine (b) Résultat obtenu par filtres de Gabor (c) image d'origine (d) Résultat obtenu par filtres de Gabor

FIG. 2.27: Segmentation de documents anciens avec Gabor



(a) image d'origine *(b) Résultat obtenu par filtres de Gabor*

FIG. *2.28: Segmentation de documents anciens avec Gabor*

des cas, ces outils fonctionnent sur des images en niveau de gris. Une binarisation n'est donc pas systématiquement nécessaire. Il est à noter que le type d'informations extrait par ces méthodes n'est pas de même nature que celle des approches classiques (RLSA, Composantes,...). Ces dernières fournissent un découpage en blocs permettant d'accéder à la structure d'une page. La caractérisation par approche texture nécessite l'application de prétraitements pour réellement segmenter en lignes, paragraphes, mots...

2.4 Caractérisation de blocs segmentés

Les outils textures présentés précédemment ne sont pas utilisés uniquement pour la segmentation d'images de documents. En effet, il est possible d'aller plus loin en mettant en oeuvre des méthodes qui permettent de reconnaître des fontes (OFR), d'identifier le type de textes (latin, asiatique, manuscrit ou typographié...) ou encore de différencier les styles de textes (gras, italique, gros caractères...) ou d'illustrations. Les besoins en analyse et indexation d'images de documents anciens exprimés plus tôt dans ce manuscrit, montrent bien que la segmentation texte/dessin n'est pas un besoin, mais plutôt un passage indispensable. Par exemple, une reconnaissance (ou une classification) de fontes permettrait d'améliorer les résultats d'un OCR. Une reconnaissance de style de textes ou d'illustrations pourrait permettre une navigation simplifiée dans un ouvrage (identification des titres, comparaison d'illustrations, des sommaires...). La reconnaissance de la structure passe, en effet, par la capacité à analyser finement le contenu d'une page, que ce soit le texte ou les illustrations.

2.4.1 Caractérisation de fontes ou de style de textes

Un peu à la manière des méthodes de segmentation, la caractérisation de texte utilise des outils fonctionnant dans le domaine spatial et d'autres dans le domaine fréquentiel. Les références détaillées par la suite ne sont pas des méthodes de segmentation, puisque la plupart d'entre elles fonctionnent sur des blocs homogènes prédécoupés.

2.4.1.1 Méthodes spatiales pour la caractérisation de fontes

D'une manière générale, ces méthodes s'appuient essentiellement sur des informations de taille et de positions de composantes connexes.

Dans [KKS04], l'auteur propose une méthode visant à identifier des caractéristiques bien spécifiques d'une zone de texte : la nature du script (latin/idéogrammes), le style (italic, gras...), la typographie (Gothic ou myung-jo) et enfin la taille de la police (10 à14). Son approche se base sur l'extraction des composantes connexes. A partir des composantes extraites il met en place un système de règles basées sur l'étude des tailles, des alignements, des projections (horizontales ou verticales) de pixels de ces composantes. Pour la discrimination de fontes, l'auteur se base sur 3 caractéristiques : moyenne des projections horizontales pour la détection de la graisse (un caractère gras possède plus de pixels noirs qu'un caractère non gras), le nombre maximum de pixels noirs horizontaux consécutif identifiable pour la détection des caractères soulignés et enfin, l'étude de la projection horizontale des composantes pour l'identification de l'italique.

Dans [FWT98], les auteurs proposent une méthode pour déterminer si un bloc de texte (déjà segmenté) est manuscrit ou non (latin ou chinois). Leur étude s'appuie sur l'analyse de projections horizontales et verticales dans l'optique de détecter si le texte est écrit horizontalement ou non. Dans un deuxième temps, l'auteur extrait les composantes connexes du texte et étudie, par ligne, leur position relative les unes par rapport aux autres (**Fig. 2.29**). En posant l'hypothèse qu'une écriture manuscrite est moins régulière dans son tracé que de l'écriture typographiée (taille et position des caractères), l'auteur propose un algorithme de classification qui permet de discerner les deux types d'écriture avec un taux de reconnaissance près de 90%.

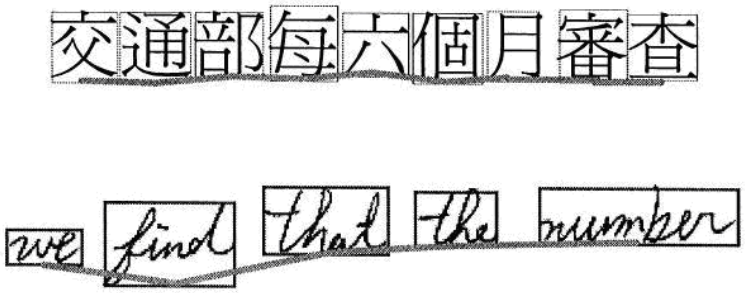


FIG. *2.29: [FWT98] Discrimination de différents types de textes à l'aide de l'étude de l'alignements des caractères*

Dans [PC01], l'auteur propose lui aussi une méthode de classification de zones manuscrites ou imprimées de texte indiens. Après une étude très précise des caractéristiques de ce type de textes, l'auteur propose un schéma de classification basée sur l'étude successive de ces différentes spécificités. Tout d'abord, il détermine le sens de lecture à l'aide de projections horizontales ou verticales. Ensuite, l'auteur se sert d'informations telles que l'intensité maximale d'une projection de pixels (horizontale ou verticale selon le sens de la lecture), de la connexité des lettres autour d'un axe précis et enfin de leur alignement pour finalement classer chaque bloc de texte.

Dans [IW98] l'auteur propose une méthode permettant de différencier des zones issues de documents bien spécifiques (math, logo, pub...). Celle-ci s'appuie sur l'extraction de composantes connexes suivie du calcul de 7 caractéristiques (nombre de composantes, périmètre, densité

pixels...).

2.4.1.2 Méthodes d'ordre fréquentiel pour la caractérisation de fontes

En marge des méthodes s'appuyant sur le calcul de descripteurs presque structurels (taille, position, nombre de pixels...), certaines méthodes utilisent des outils textures. Comme pour la segmentation, l'analyse de fontes de texte par approche texture, permet de s'affranchir de nombreux défauts des approches spatiales (essentiellement la nécessité de connaître les caractéristiques des images étudiées).

Dans [ZTW01], l'auteur propose une approche texture pour caractériser les fontes. Après avoir séparé les zones de textes les unes des autres, il crée dans un premier temps pour chaque zone de 128X128 une texture artificielle homogène (**Fig. 2.30**). Pour cela, l'auteur cherche à minimiser les espaces blancs (inter mots, inter lignes, inter lettres...) à l'aide des caractères présents dans la zone étudiée (et qui est sensée être uniquement composée du même type de caractères). Dans un deuxième temps, l'auteur applique sur cette texture un banc de filtres de Gabor basé sur l'étude de 4 fréquences et 4 orientations à l'aide duquel il calcule la moyenne et l'écart type de la réponse de chaque sortie du banc. Ceci donne donc 32 caractéristiques, signant cette texture homogène. Cette technique permet de discriminer plus d'une cinquantaine de fontes avec un taux de reconnaissance supérieur à 97%.



FIG. *2.30: [ZTW01] Discrimination de différents types de textes à l'aide de filtres de Gabor*

Des travaux récents réalisés par Ma et Doermann dans [MD05], les auteurs comparent deux approches par filtres de Gabor. La première est une approche classique (construction d'un banc de filtres composé de 4 fréquences et 4 orientations). La deuxième est basée sur une conception nouvelle de la perception visuelle qui permettrait de caractériser des aspect typographiques très fins que les filtres classiques ne peuvent percevoir. L'originalité tient en partie à l'étape de paramétrisation ou chaque valeur est supposée liée aux autres. Au final, les auteurs utilisent deux approches différentes pour classer les vecteurs générés par l'application des filtres (distance euclidienne et réseau de neurones).

[Rob01b] propose deux approches pour classer des fontes de journaux. Son but n'est pas d'identifier le nom des fontes mais plutôt de regrouper les caractères qui ont la même typographie (même fonte, même style). L'auteur propose, dans un premier temps, une solution pour classer les fontes selon leur taille. Pour ce faire, l'auteur utilise un algorithme de classification doué d'apprentissage et qui met en relation, pour chaque bloc utilisé pour apprendre, la hauteur de ligne (en pixel) en fonction du nombre de lignes de la zone apprise. Ainsi, lorsqu'elles sont utilisées, les petites fontes vont avoir pour caractéristiques de générer des zones de texte avec beaucoup de lignes et peu de hauteur de ligne (à l'inverse des grosses fontes). Après avoir séparé

les fontes selon leurs tailles, l'auteur classe directement les caractères selon leur inclinaison, leur graisse et le type de fonte. Pour cela l'auteur se base sur deux postulats : "tous les caractères composant un mot appartiennent à la même famille typographique et chaque caractère d'une famille typographique possède un dessin unique au pixel près". Donc, en appareillant les formes des caractères avec pour contrainte les deux postulats précédemment mentionnés, l'auteur arrive à classer les fontes des journaux qu'il analyse.

Dans [Egl98], l'auteur propose un traitement basé sur une analyse "macroscopique" des texture permettant d'établir une différence entre diverses polices de texte(taille, graisse...). L'auteur propose 3 indices textures permettant de réaliser cet objectif.

1. Le premier est basé sur une étude du nombre d'intersections d'une droite avec les lignes de texte. Pour chaque ligne (au sens pixel) de la zone étudiée on compte le nombre d'intersections entre une ligne horizontale et les pixels noirs. Ce principe permet de calculer une probabilité P_n qui se calcule de la manière suivante : $P_n = L(n)/N_l$ avec $L(n)$ le nombre de lignes comportant n intersections et N_l le nombre de lignes étudiées de la zone. Cela permet au final de calculer une mesure d'entropie ($E = \sum_{n=1}^{\infty} P_n.log(1/P_n)$) qui donne une information sur la nature du texte (permet de classer les zones de texte selon la graisse)
2. Le deuxième est basé sur un double calcul de compacité. Dans un premier temps l'auteur calcule une densité horizontale basée sue le dénombrement de toutes les intersection noires entre le texte et les lignes horizontales de la zone. Le même procédé est répété sur les lignes verticales. En normalisant ces deux mesures par la largeur et la hauteur du bloc étudié l'auteur dispose d'une signature qui lui permet de classer 17 polices différentes en taille, graisse, formes (latine, orientale...)
3. Pour ce dernier indice, [Egl98] propose une signature basée sur une analyse du relief des formes de la zone. Pour ce faire, il appose un pavage (isotropique ou anisotropique, vertical ou horizontal) sur l'image et calcule dans chaque pavé le niveau de gris moyen. En étudiant la répartition des densités sur tout le pavage (au moyen d'un écarte type), il est possible de se renseigner sur la variation des densités de niveaux de gris et finalement de signer plusieurs types de polices.

Les travaux [All04] reprennent ceux de [Egl98] et [WCW82] et proposent une extension. L'auteur suggère notamment de fusionner les indices de compacité horizontale et verticale, de prendre en compte l'excentricité du bloc (rapport hauteur sur largeur) et des informations sur les composantes connexes (nombre, surface...). De toutes ces propositions, l'auteur retient au final 28 valeurs à calculer sur l'ensemble des blocs de test. Finalement l'auteur classe plus de 300 blocs à l'aide de vecteurs à support machine avec un taux de reconnaissance proche de 77%

2.4.2 Caractérisations d'illustrations

Dans la section précédente, nous avons déjà abordé le problème de l'indexation (ou caractérisation) d'images naturelles. Or, nous l'avons évoqué, les dessins de traits ont des caractéristiques bien différentes de celles des photos. D'un point de vue visuel, là où les photos sont composées de grandes zones caractérisées par une grande variété de niveaux de gris, les illustrations ou dessins ne présentent pas la même caractéristique. De plus, la conception même d'une illustration (ensemble plus ou moins dense de petits segments), induit des caractéristiques de fréquences et de formes différentes de celles relatives aux photos (cf. tests réalisés dans les sections précédentes).

Si à l'évidence, les outils classiques semblent peu appropriés à l'analyse d'illustrations, il existe néanmoins peu de références traitant spécifiquement de ce sujet. Nous proposons de détailler deux

références dont l'objectif est d'indexer une base d'images de lettrines.

Dans [PVU+06], les auteurs décrivent une méthode de caractérisation de lettrines à l'aide d'une loi puissance (Zipf). Sur le même principe que [CCMV03], les auteurs appliquent un masque, mais cette fois-ci, non rectangulaire sur une image où l'espace des niveaux de gris a été discrétisé en 3 niveaux. De ce fait, il y a au maximum 3^5 motifs différents possibles dans un dessin. Appliqué aux images de traits, le diagramme de Zipf (échelle Log-Log représentant la fréquence d'apparition des motifs), fait apparaître trois segments de droite au sein de ce diagramme. Chacun d'entre eux a un sens précis (cf **Fig. 2.31**). Le segment S1 correspond aux régions homogènes dont le motif revient fréquemment, alors que le segment S3 correspond aux contours (zones de transitions). Selon les auteurs, c'est la nature de l'illustration étudiée qui va influer sur la forme de la courbe. Le segment S2 diffère, selon le nombre de motifs différents et leurs fréquences d'apparition. Ce sont les pentes de ces segments qui sont extraites pour caractériser les lettrines. A l'aide d'une distance de Hamming entre les vecteurs caractéristiques, les auteurs testent la capacité de leur système à différencier 3 styles de lettrines et obtiennent près 95% de bon classement.

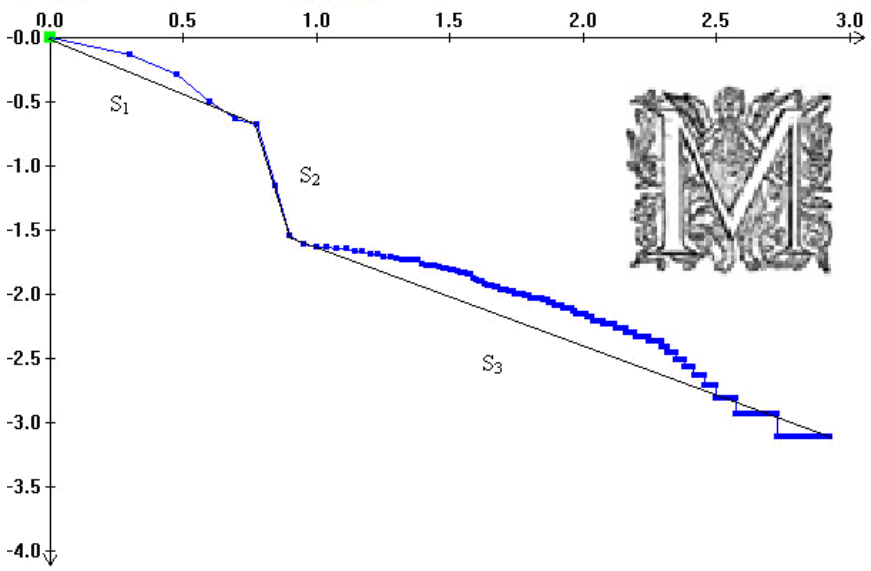


FIG. *2.31: Caractérisation de lettrine par la loi de Zipf [PVU+06]*

Les auteurs de [UOL05] ont la même ambition de comparaison de lettrines, mais leur proposition est d'ordre structurel. Dans un premier temps, la lettrine est segmentée de manière à identifier les zones texturées. Cette opération est réalisée à l'aide de la fonction d'autocorrélation. Appliquée sur une zone texturée, cette fonction décroît plus rapidement que sur une zone de fond. La figure **Fig. 2.32**.b donne un exemple d'extraction des zones texturées. Chaque couche d'information extraite, va être caractérisée par une signature structurelle (organisation spatiale des couches d'information). Les auteurs comparent deux approches.

1. Le minimum spanning tree[10]. Pour chaque couche d'information, les composantes connexes sont extraites. Le centre de gravité de chacun d'entre elles (**Fig. 2.32**.c) permet de construire

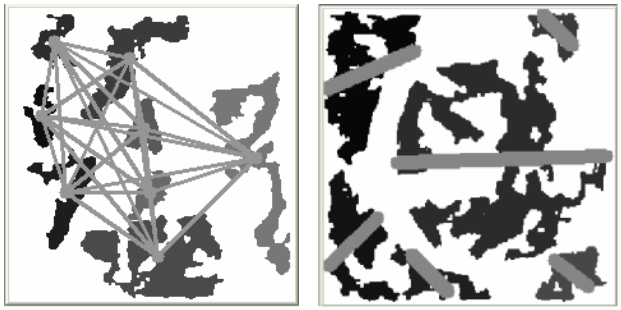
[10] *http : //en.wikipedia.org/wiki/Minimum_spanning_tree*

un arbre (**Fig. 2.32**.d). Le vecteur caractéristique de la lettrine est composé du résultat de l'algorithme MST appliqué sur chacun de ces arbres (un par couche d'information : zones texturées, zones de contours, zones de fond).

2. Pour chaque composante connexe d'une couche d'information, est calculé un axe d'inertie (**Fig. 2.32**.f). L'image est signée en fonction de la distance et de l'angle entre chaque paire de segments.



(a) Lettrine d'origine *(b) Extraction des zones texturées* *(c) Centre de gravité des composantes connexes*



(d) Construction d'un arbre sur la base d'une étude des centres de gravité *(e) Axes d'inertie des composantes*

FIG. *2.32: Caractérisation de lettrine par approche texture [PVU+06]*

2.4.3 Conclusion

Dans cette section, nous avons présenté une partie des méthodes permettant la caractérisation du texte ou des illustrations. Ces techniques permettent de détailler finement le contenu des images. Cependant, le fait que ces méthodes soient appliquées sur des blocs pré-segmentés pose le problème de leur application sur des documents non segmentés.

2.5 Comparaison de structure de documents

Les méthodes d'analyse de contenu d'images de documents que nous avons détaillées précédemment, montrent qu'il est possible de segmenter ou de caractériser une grande variété d'images.

Selon le cas, ces méthodes permettent d'extraire des informations relatives à la structure physique ou fonctionnelle intermédiaire de l'image de document. Dans la suite de ce chapitre, nous présentons des méthodes permettant d'exprimer, de modéliser et de comparer ces structures. Nous nous focaliserons plus particulièrement sur celles traitant des images de documents dont le modèle n'est pas connu à l'avance (eg : pas de DTD). Ainsi, nous allons essentiellement présenter des approches utilisant uniquement des informations bas niveau. Au travers d'applications de type document image retrieval, nous présenterons comment il est à la fois possible de modéliser et de comparer ces structures.

2.5.1 Introduction

Par définition, un document a ceci de différent d'une image naturelle, qu'il possède un contenu fortement structuré. A l'image des algorithmes de segmentation, les méthodes de DIR sont étroitement liées aux images qu'elles se proposent d'analyser. Dans les état de l'art [Kau99, Doe98b], les auteurs recensent une grande variété de méthodes et algorithmes. Certains auteurs utilisent (en complément d'une analyse d'image) l'information textuelle du document. Cette information est bien entendue très importante, mais selon le cas il n'est pas toujours possible d'accéder à ce contenu (OCR impossible). Cet état de l'art détaille principalement les approches n'utilisant pas l'information textuelle. Globalement, la question que se posent les auteurs est toujours la même : comment comparer deux images ? C'est-à-dire, comment modéliser la structure et le contenu de la page afin de permettre une comparaison ?
La section qui suit revient sur deux façons de procéder. La première consiste à extraire toute une quantité d'informations bas niveau et de comparer ces caractéristiques. La deuxième consiste à découper une page en zones similaires et de la modéliser par des graphes. Cette dernière technique nécessite d'effectuer une analyse de plus haut niveau (sémantique) du contenu de l'image. Elle implique donc souvent l'utilisation d'information a priori et se heurte donc au problème de perte de généricité.

2.5.2 Signature de mise en page par extraction de caractéristiques

Une manière très répandue d'extraire des indices bas niveau tout en gardant l'information topologique, est l'utilisation d'une approche par pavage. Cette technique revient à poser une grille de taille prédéfinie sur l'image et de calculer ensuite, pour chaque case, toute une quantité d'informations. Selon la méthode, la comparaison peut se faire cellule à cellule, ligne à ligne ou encore page à page.

Dans [KSP97], les auteurs utilisent un pavage dont la taille est calculée en fonction de la résolution et des dimensions de l'image. Ensuite, pour chaque cellule toute une liste de caractéristiques physiques du bloc est extraite (rapport pixels noirs/blancs, longueur de plage de pixels noirs...). A partir de ces informations, un algorithme de classification définit l'étiquette de chacune des cellules (texte, fond, dessin, ligne). Toute l'originalité de cette publication se situe dans la prise en compte du voisinage pour définir définitivement le label des blocs. La structure se résume ici à l'établissement d'un vecteur de caractéristiques relatif aux informations extraites précédemment. Une comparaison entre deux structures est donc possible en calculant une distance euclidienne entre deux vecteurs. L'objectif des auteurs est, dans une base de 70.000 documents, de pouvoir différencier des documents de type journaux, de documents de type lettres ou formulaires, publicités, cartes...

Dans le même esprit, l'auteur de [HKW99] s'appuie sur une analyse, via un pavage, de la

disposition et des caractéristiques des lignes de texte extraites d'un document. La méthode se base sur une segmentation en blocs rectangulaires et qui est supposée être parfaite. A partir de là, l'auteur applique, comme dans la méthode précédente, un pavage dont chaque cellule est de taille prédéfinie. Dans un premier temps, chacune d'entre elles est étiquetée " texte/non-texte " selon un critère de densité de pixels. Une distance de similarité est calculée entre toutes les cellules des deux documents. Un algorithme de programmation dynamique, permet de trouver une correspondance ligne à ligne entre deux pages afin de minimiser la somme totale des distances. Cette technique permet donc à la fois de caractériser la structure du document mais également de comparer deux documents.

Toujours en utilisant une méthode par pavage, [HC97] a réussi à mettre en place une méthode de comparaison de documents envoyés par fax. Le codage utilisé pour la compression de documents envoyés par fax, code la succession de chaînes de pixels noirs ou blancs par lignes. L'auteur se sert de cette information pour extraire des points caractéristiques. Ainsi, il repère chaque pixel noir de l'image qui ne possède pas de voisins inférieurs noirs. En appliquant un pavage sur cette image, il calcule ensuite le nombre de points caractéristiques pour chaque cellule, ce qui permet de construire un vecteur caractéristique. Pour comparer deux pages, l'auteur utilise une distance euclidienne ou une distance de Hausdorff. Cette technique a été testée sur une base constituée de près de 1000 documents. Elle permet dans près de 98% des cas, de discerner correctement, par exemple, des documents de type articles scientifiques de lettres manuscrites...

Dans [SD99a, SD99b], les auteurs ne cherchent pas exactement à comparer des pages entre elles : ils souhaitent être capables de déterminer automatiquement le " label " de la page à partir d'informations de mise en page contenues dans une base de données. A la différence des méthodes précédemment citées, les auteurs de [SD99b] utilisent une information plus riche et surtout plus détaillée pour arriver à décrire toute la finesse qu'il désire donner à leur méthode. Ainsi, ils calculent des caractéristiques sur trois niveaux différents : les composantes connexes (tailles, nombre...), projections horizontales et verticales, et une étude des caractéristiques sur toute la page. Des heuristiques et un arbre de décision dédiés au problème, permettent ensuite de déterminer certaines informations sur le label de la composante principale de la cellule (texte, dessin, gravure ou autre), le nombre et l'emplacement de colonnes, l'emplacement d'éventuels tableaux ... Toutes ces caractéristiques sont finalement données en entrée d'un arbre de décision qui permet d'apposer un label unique à l'image. L'objectif est donc de pouvoir trier automatiquement de grandes bases d'images de documents.

Dans ses travaux de thèse, [Bag04] adapte un indice de granulométrie pour signer, de manière très simple, la mise en page de documents. Cet indice est calculé à partir d'une succession d'ouvertures utilisant un élément structurant rectangulaire, dont la largeur et la longueur varient indépendamment. L'image étant binarisée, l'auteur étudie l'évolution du nombre de pixels noirs. Cette distribution, qui est fonction de la taille du rectangle, est la base de la comparaison puisque la similarité entre deux documents est donnée suite au calcul d'une distance euclidienne entre deux distributions (celle de l'image requête et celle contenue dans la base).

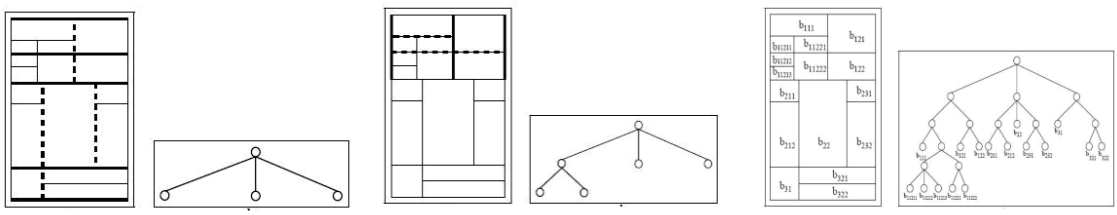
Dans ses travaux [PLCS01, MSB97a], l'auteur s'appuie à la fois sur une extraction de composantes connexes et sur une étude des orientations du texte pour établir un algorithme de comparaison de pages. Leurs caractéristiques (la taille, la position...) permettent de les chaîner les unes aux autres selon la surface croissante de leur rectangle englobant. Une page est donc modélisée par une liste chaînée. Un algorithme de comparaison de ces composantes, permet de calculer une distance entre une liste L1 (liste de l'image requête) et une liste L2 (liste de l'image de la base).

Discussion : Les approches que nous venons d'évoquer ont le mérite d'être peu coûteuses en temps de calcul et de se ramener à un modèle qui permet une comparaison simple des caractéristiques (distance entre vecteurs). Ces techniques tiennent plus de la comparaison de contenus que de la comparaison de structure de documents, mais ont néanmoins prouvé leur efficacité dans des cas de figures bien précis.
Dans la suite de cette section nous verrons qu'il est possible de décrire de manière plus structurée un document.

2.5.3 Comparaison par graphes

L'utilisation de graphes, permet une modélisation d'un document. En effet, à l'aide d'un noeud dans lequel on répertorie des caractéristiques propres à une zone de l'image on peut associer un degré d'abstraction supplémentaire, en modélisant des relations logiques entre les différents noeuds (une zone est à coté d'une autre, celle-ci est incluse dans celle là...). Ce choix permet d'être plus proche de la notion de mise en page puisqu'on peut détailler finement toute une quantité d'informations physiques et logiques. De plus, la littérature offre toute une gamme de méthodes de "graphe-matching" utilisable durant l'étape de comparaison.

Par exemple, l'auteur de [DA02] propose une méthode de comparaison de formulaires. Ces documents sont fortement structurés, avec notamment des lignes noires permettant de séparer clairement les différents champs ou zones. L'algorithme de construction de l'arbre modélisant la mise en page, fonctionne selon le principe du XY-cut. Le cut est effectué lorsqu'une ligne verticale ou horizontale de la partie pré-impriméee du formulaire du document est détectée. Imaginons que deux lignes horizontales soient détectées, alors l'algorithme de construction rajoute au noeud initial trois fils (**Fig. 2.33**.a). On a ainsi modélisé la relation : " le document est composé de trois zones horizontales". Comme l'algorithme XY-cut est algorithme récursif, le prochain séparateur est recherché dans le fils le plus à gauche dans l'arbre (parcours en profondeur d'abord). On choisit de couper selon un séparateur vertical. On crée donc deux fils à ce noeud (**Fig. 2.33**.b). Et ainsi de suite, jusqu'à créer un arbre modélisant toutes les relations contenant/contenus (une feuille correspond à un bloc, un noeud correspond à un regroupement de blocs). Afin d'éviter une (coûteuse) comparaison terme à terme entre deux arbres, l'auteur crée une matrice M_{ij} où chaque indice indique le nombre de noeuds ayant i enfants au niveau j. Pour mesurer une distance entre deux matrices, l'auteur propose une formule basée sur une somme des différences, terme à terme, des deux matrices.



(a) Construction du premier niveau *(b) Construction du deuxième niveau* *(c) Modélisation complète du formulaire*

FIG. *2.33: Construction d'un graphe modélisant la structure d'un formulaire [DA02]*

Les auteurs de [BMS03, MMS05] présentent un système permettant une comparaison de

documents sur la base de leur similarité de mise en page. Un peu à la manière de [DA02], l'image est découpée selon l'identification de certains séparateurs (une grammaire est définie et utilisée pour les retrouver). L'arbre est construit de manière classique. Un noeud racine symbolise une zone. Si celle-ci possède un séparateur et qu'elle est donc divisée en deux, on crée deux noeuds fils. Une feuille est labellisée texte (T),image (I), ligne verticale (vL) ou ligne horizontale (hL). Partant du principe que beaucoup d'arbres ont une structure similaire, l'auteur propose d'effectuer une réduction des arbres afin de simplifier la phase de comparaison. Cette réduction est le fruit de l'application de 11 règles prédéfinies. Certaines règles de réduction sont paramétriques : "Une petite région I est remplacée par une région T ", " Une petite ligne (hL ou vL) est enlevée de l'arbre "... D'autres visent à simplifier des parties d'arbre dont l'information est jugée redondante ou inutile : " Si un noeud interne possède des fils qui ont tous la même étiquette T, alors ce noeud prend l'étiquette T et tous ses fils sont détruits", "Si une feuille a pour étiquette hL ou vL alors elle est détruite "... La figure **Fig. 2.34** illustre la réduction d'un arbre. On voit par exemple, sur la partie gauche de l'arbre, que l'algorithme de réduction permet de se ramener à un sous arbre à 1 noeud, contre 7 avant l'opération. Si l'on regarde l'image référence, on remarque que l'information sur la mise en page n'est pas dénaturée puisque l'on modélise toujours le fait que la page se divise verticalement en deux blocs : d'un côté, un bloc de texte sur la gauche, de l'autre un bloc de texte et une image séparés par une ligne verticale. La comparaison entre deux graphes est réalisée à l'aide d'une distance d'édition. Les auteurs travaillent sur des images de documents datant du $XIX^{\text{ème}}$ siècle.

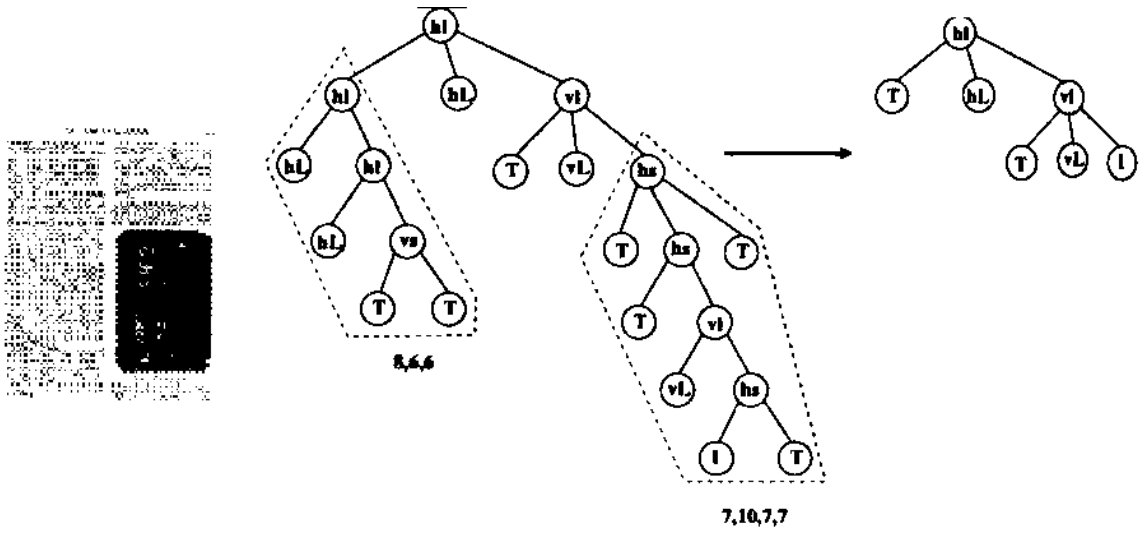


FIG. *2.34: Construction et réduction d'un graphe modélisant la structure d'une page [BMS03]*

Dans [BFH+04], les auteurs considèrent qu'une mise en page est l'association d'une combinaison spatiale d'entités ayant un aspect visuel spécifique. En premier lieu, une segmentation de l'image en régions homogènes (dessins, paragraphes, titres et fond) est effectuée. A l'inverse de [DA02], les auteurs ne se basent pas uniquement sur la position relative des blocs pour créer l'arbre (modélisation contenant/contenu). En effet, les documents analysés sont des journaux. La structure n'est donc pas identifiable à l'aide de marqueurs clairement identifiables (lignes noires). Pour construire l'arbre binaire correspondant à la structure du document, les auteurs séparent les différentes entités à l'aide des composantes connexes. Ensuite, pour chaque bloc, sont extraites

des caractéristiques de taille, position, couleur, classe (texte, dessin, fond) et enfin le contenu du texte (quantité de texte). Ces caractéristiques sont à la base de l'algorithme de construction de l'arbre. Une distance de similarité est calculée entre chaque zone, ce qui permet de déterminer à quels niveaux les noeuds vont être placés. Opérer de cette manière permet de construire un arbre dont les frères (noeud de même niveaux) et les fils (noeuds de niveau inférieur) ont des caractéristiques communes.
La distance entre deux arbres (deux pages) est effectuée à l'aide d'une formule inspirée d'une distance d'édition. Cette distance (dont la complexité est en O(NM) avec N et M le nombre de noeuds des arbres), est une formule récursive étudiant la forme de l'arbre.

Discussion :

La modélisation par arbre possède des atouts séduisants (simplicité de modélisation, utilisation d'outils de comparaison d'arbres...), mais présente également certains défauts comme la taille de l'arbre et sa complexité à la construire. Quoiqu'il en soit, notre réflexion nous pousse à conclure que les méthodes présentées dans cette section nécessitent une phase de segmentation précise des images en éléments de contenu. La recherche, la comparaison, la modélisation ou l'indexation de documents ne semblent donc pas (ou peu) avoir été traitées dans la littérature.

2.6 Conclusion générale

Ce chapitre a permis de revenir sur les points importants généralement abordés en analyse d'images de documents. Nous avons ainsi donné les définitions des différentes structures permettant de décrire et de modéliser le contenu d'un document. Nous avons également détaillé les grandes familles de méthodes permettant d'accéder à ces différents contenus.
La question que soulève cet état de l'art, est de savoir s'il est possible d'analyser et d'indexer le contenu d'une image de document sans analyser sa structure ? En effet, la majorité des techniques présentées cherche en majeur partie à segmenter l'image et à remonter vers la structure du document. Ce choix réduit généralement la généricité des techniques, dès lors que les documents traités présentent une forte variabilité.
Les approches textures proposent un premier élément de réponses à cette question, puisqu'elles permettent de caractériser le contenu d'une image sans émettre d'hypothèses sur la structure du document. Cependant, les difficultés d'adaptation de certains de ces outils sur notre corpus de documents anciens, nous a poussés à trouver une autre manière de caractériser le contenu de nos images.
Ainsi, dans le chapitre suivant, nous allons décrire notre propre système d'analyse de contenu d'images de documents anciens par approche texture multirésolution. Enfin, nous verrons dans le dernier chapitre, au travers d'expériences, qu'il est possible d'indexer le contenu d'images de documents anciens, sans pour autant devoir le segmenter ou retrouver sa structure.

Chapitre 3

Notre approche texture pour la caractérisation du contenu

3.1 Objectifs

3.1.1 Quels sont les enjeux ?

Les deux premiers chapitres de ce manuscrit ont permis de mettre en évidence, d'une part l'enjeu majeur qu'est la conservation des documents du patrimoine, et d'autre part l'importance du développement et de la mise en place de nouveaux outils de caractérisation d'images de documents à structure variable. Dans ce cas précis, les objectifs consistant à accéder, questionner ou retrouver des informations relatives aux contenus des images numérisées nécessite donc une nouvelle réflexion. De prime abord, la réalisation d'outils de traitements d'images de documents anciens (et plus globalement de documents à structure variable) soulèvent trois problèmes majeurs.

- Le premier est celui relatif à une caractérisation du contenu de ces images qui se voudrait générique et utilisable sur un large panel de documents. En effet, cet objectif nécessite la réalisation d'outils d'analyse et caractérisation d'images dédiés aux documents. Or, nous l'avons présenté dans le chapitre précédent, il reste complexe de transposer sur ce type de documents, des méthodes et des outils développés initialement pour des images naturelles ou des documents contemporains.
- La deuxième difficulté provient de la définition et de l'intégration de l'utilisateur dans le système qu'il sera amené à utiliser. En effet, ces applications seront principalement manipulées par des personnes n'étant pas spécialistes en traitement d'images. Il faut donc prendre en compte cette spécificité et ne pas proposer d'outils nécessitant, par exemple, une phase de paramétrisation trop importante dont dépendrait la qualité des futurs résultats.
- Enfin, ces objectifs nous confrontent directement au problème classique du traitement de masse de données très volumineuses. La constitution de corpus implique le traitement de quantités d'images importantes. Il convient donc de réfléchir à un système capable non seulement de retrouver l'information pertinente qui se trouve être noyée dans la masse, mais également de réfléchir à une organisation permettant de traiter de manipuler et d'interroger un grand nombre de données et d'images.

Cette réflexion sur les attentes des usagers et la complexité de la réalisation d'outils de caractérisation d'images qui en découlent, nous a amené à proposer une nouvelle approche de

caractérisation de contenu d'images. Nous présentons dans la suite de ce chapitre notre contribution à cette problématique qui se base sur l'utilisation de nouvelles caractéristiques textures pour décrire le contenu des images. A la suite de ce chapitre, nous détaillerons des expérimentations (utilisant ces indices de caractérisation que nous proposons) et qui permettent d'analyser le contenu des images numérisées sans pour autant segmenter et retrouver la structure des images analysées.

3.1.2 Présentation de notre système de caractérisation de contenu

Ce chapitre présente une démarche originale de caractérisation d'images de documents. A l'instar de ce qui ce 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 des textures qui composent l'image.
Face aux caractéristiques des images de notre corpus, nous avons décidé de porter notre choix sur l'extraction d'informations bas niveau. La raison principale de ce choix, est notre volonté de ne pas interpréter le contenu des images. Plus précisément, notre approche consiste à réaliser une caractérisation robuste du contenu des documents issus d'ouvrages anciens provenant d'horizons très différents, donc fortement hétérogènes entre eux. Etant donné que l'usage de l'outil est orienté vers des non-spécialistes en traitement d'image, il ne doit comporter ni seuils, ni modèles, ni structures explicites dans le processus d'analyse. Le processus global consiste donc à caractériser précisément les contenus des pages d'un ouvrage, et ceci à l'aide de nouveaux algorithmes d'extraction d'indices de textures dédiés à l'analyse de documents en niveaux de gris. Cette approche n'intègre pas de connaissances sur les caractéristiques physiques et sémantiques des images traitées. Nous cherchons juste à exprimer la variété des textures présentes dans les images de documents anciens. Cette caractérisation des pixels de l'image, sera la base de la création ultérieure d'outils permettant d'analyser précisément le contenu des images.

Le fonctionnement de notre système de caractérisation est décrit dans le schéma **Fig. 3.1**. Il se compose de deux parties distinctes. La première correspond à une phase de calcul d'indices textures sur des images de documents. La deuxième correspond aux usages potentiels que l'on peut faire de ces indices. Voici en détail les différents points du schéma **Fig. 3.1** :

1. Point A : Lorsqu'un nouvel ouvrage est numérisé, il est automatiquement mis en entrée de notre système d'analyse.
2. Point B : Pour chaque page un ensemble d'indices textures est calculé pour chaque image de l'ouvrage. Visuellement, une texture est un arrangement de pixels dont une vision globale se traduit par des caractéristiques d'orientations et de fréquences similaires. Les indices textures que nous proposons, ont été imaginés dans le but d'extraire des informations en relation avec ces deux caractéristiques. Nous proposons ainsi le calcul de 5 indices différents. Les 3 premiers sont relatifs aux informations liées aux orientations, les deux autres sont relatifs aux informations de fréquence. Ces indices sont calculés à différentes résolutions de l'image. Dans le cadre d'une analyse d'images de documents, ce choix permet de percevoir des structures de tailles différentes dans l'image, la résolution du dessin étant limitée par la main de l'artiste et celle du texte par les caractéristiques physiques de la presse et les choix de mise en page. L'utilisation d'algorithmes de caractérisation de texture n'est pas limitée à ceux que nous proposons, et il est tout à fait envisageable d'enrichir le système que nous proposons.
3. Point C : Une fois l'ensemble des pages de l'ouvrage analysé, les informations extraites

sont stockées dans une base. Cette base peut, soit répertorier pour chaque pixel d'une image les caractéristiques textures lui correspondant, soit correspondre à des métadonnées synthétisant brièvement les informations apportées par les indices textures proposées.

4. Point D : Cette étape correspond à une phase d'analyse des résultats. A l'aide d'un algorithme de classification, il est possible de visualiser ou comparer les différents éléments composant les pages des livres.
5. Point E : Cette partie correspond à l'ensemble des applications potentiellement réalisables à l'aide, entre autre, des indices textures stockés dans la base. Cette liste n'est pas exhaustive. Ces outils sont réalisés par des traiteurs d'images mais ont pour objectif d'être manipulés principalement par des personnes de sciences humaines et sociales (SHS). Cette situation rend parfois complexe les échanges d'informations et la spécification des besoins. C'est pourquoi, dans le cadre de ces travaux de thèse, nous n'avons pas répondu à un besoin spécifique, nous avons préféré mettre en place des expérimentations. Celles-ci permettent d'illustrer la pertinence des indices extraits, mais également de mettre en avant le fait qu'il n'est pas nécessaire de redévelopper un processus d'analyse d'images pour chaque nouveau besoin. Sur la base des indices extraits, il est tout à fait envisageable d'imaginer la mise en place de plusieurs applications répondant à des besoins spécifiques.

Dans ce chapitre, nous abordons les points B,C et D. Ces parties correspondent à notre proposition de calcul d'indices textures. Le dernier chapitre détaillera plus particulièrement le point E.

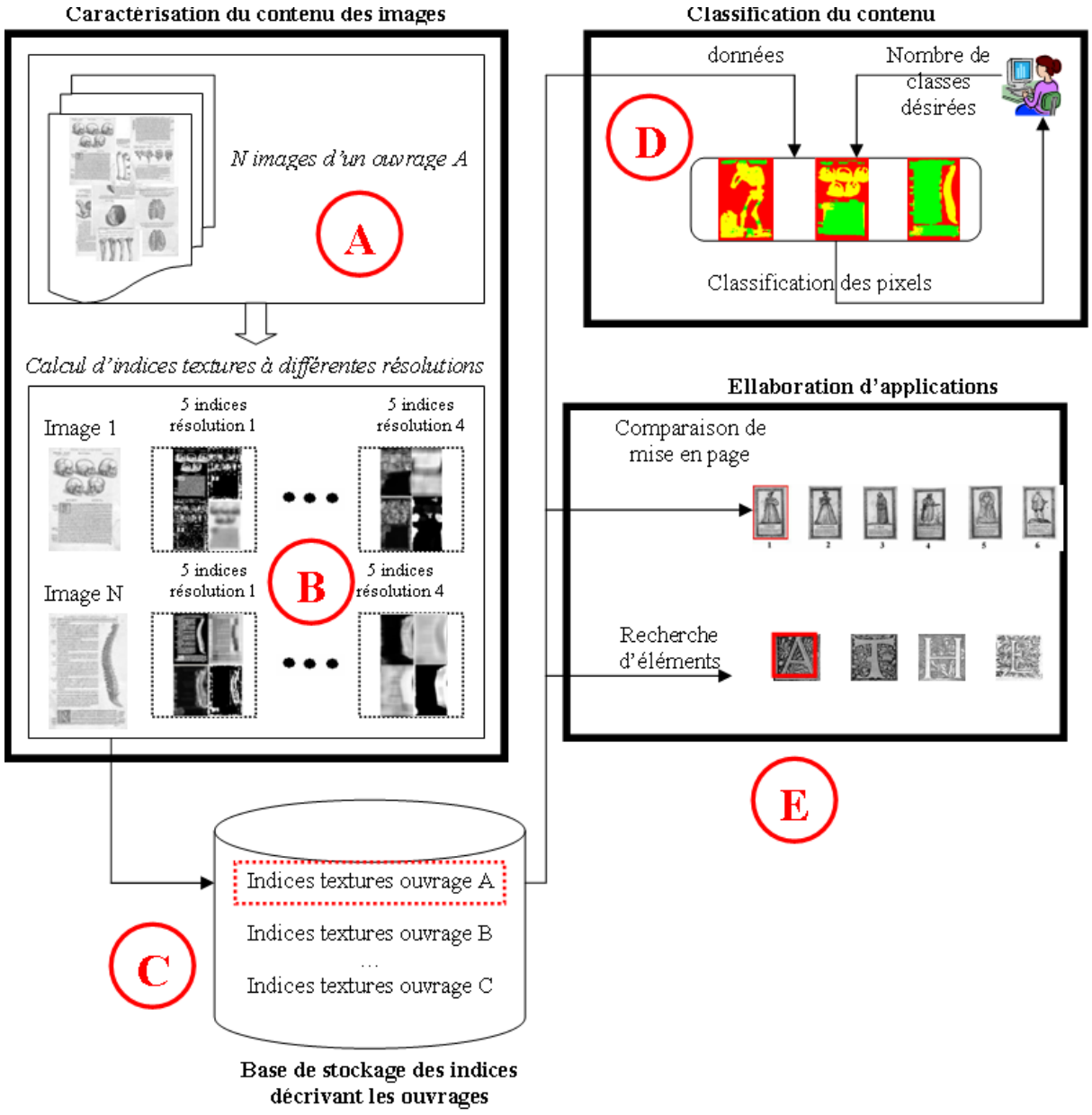


FIG. 3.1: *Présentation de notre système*

3.2 Extraction d'indices textures dédiés à l'analyse d'images de documents

3.2.1 Principe global

3.2.1.1 Formalisation

Dans cette section , nous formalisons la construction des attributs textures sur les images d'un corpus. Les notations données ici, sont celles que l'on retrouve dans la suite de ce manuscrit.

Il est tout d'abord possible de définir notre corpus :

$$Corpus_{IJ} = \{\Theta / dim(\Theta) = (I, J)\} \quad i \in \overset{*}{\mathbb{N}}_I = \mathcal{I}, j \in \overset{*}{\mathbb{N}}_J = \mathcal{J}$$
avec I et J largeur et hauteur d'une image.

Le corpus est composé de k pages.

$$\Theta = \{ \bigcup_{k=1}^{card\Theta} P_k \} \quad \forall \text{k largeur}(P_k)\text{=I et hauteur}(P_k)\text{=J}$$

Ce qui donne sous forme matricielle :

$$P_k = \underset{i \in \mathcal{I}, j \in \mathcal{J}}{(p_{ij}^k)}$$

Soit $Texture_{s,r}$ la fonction qui pour une image P_k associe l'application d'un algorithme d'extraction d'indice texture s à une résolution r.

$$Texture_{s,r} : P_k \longmapsto Texture_{s,r}(P_k) \quad \begin{array}{l} avec \;\; s \in \mathcal{A} = \{A_1,...,A_n\} \; pour \; n \; indices \; textures \\ et \;\; m \in \mathcal{R} = \{R_1,...,R_m\} \; pour \; m \; résolutions \end{array}$$

Ce qui, sous forme matricielle donne pour une image P_k :

$$\begin{aligned} Texture_{sr}(P_k) &= \underset{i \in \mathcal{I}, j \in \mathcal{J}}{Texture_{sr} \; (p_{ij}^k)} \\ &= \underset{s \in \mathcal{A}, r \in \mathcal{R}}{C_{i,j}^{s,r}} \end{aligned}$$

Pour une image P_k, si on calcule l'ensemble des attributs textures à toutes les résolutions on a donc une matrice en trois dimensions de taille IxJx$(s * r)$.

$$Texture(P_k) = \Big(\underset{s \in \mathcal{A}, r \in \mathcal{R}, i \in \mathcal{I}, j \in \mathcal{J}}{\times C_{i,j}^{s,r}} \Big)$$

Pour un pixel d'une image k, on a donc un vecteur de dimension $s * r$ qui contient toutes les informations textures à toutes les résolutions.

$$\forall i, j, k \;\; C_{i,j}^k = (C_{i,j}^{s_1,r_1}, ..., C_{i,j}^{s_n,r_m})$$

3.2.1.2 Outil utilisé pour la caractérisation des orientations

L'orientation est l'une des caractéristiques visuelles impliquée dans la vision prête attentive. Le modèle de Itti [IK00] l'utilise pour caractériser les points de saillance dans les images naturelles pour son côté discriminant. Nous nous sommes intéressés à cette caractérisation afin de proposer trois indices textures liés aux informations d'orientation. Dans la littérature, la caractérisation des orientations est souvent réalisée à travers un filtre directionnel de type Gabor ou ondelettes [EDC97, BSN04]. Cependant, ce type de filtres nécessite un choix judicieux du banc de filtres

à appliquer afin de mettre en évidence les réponses de la convolution à des fréquences et des orientations précises. Notre ligne de conduite étant de minimiser le nombre de paramètres à fixer, nous avons écarté les approches de type *"filtre"* puisqu'il reste très difficile de choisir la forme des bancs de filtres, les paramètres liés aux fréquences et aux orientations à étudier. De plus, il est parfois nécessaire de binariser...

Ainsi, nous avons choisi d'utiliser un outil non paramétrique basé sur la fonction d'autocorrélation : la rose des directions (proposée par Bres dans [Bre94]). La rose des directions est un diagramme polaire se basant sur l'étude de la réponse de la fonction d'autocorrélation lorsqu'elle est appliquée sur une image.
Dans ses travaux, [Egl98] définit la fonction d'autocorrélation comme étant le regroupement de l'ensemble des valeurs que l'on peut obtenir en faisant la somme de tous les produits des niveaux de gris des points en correspondance après translation de l'image I par rapport à elle-même. Ainsi, un point $C_{xx}(k,l)$ de la fonction d'autocorrélation contient la valeur de la somme des produits des niveaux de gris des points en correspondance après une translation de vecteur (i, j). Ces différentes translations permettent d'inspecter l'image selon ses différentes directions. Toutes les translations selon des vecteurs colinéaires donnent des indications selon la direction correspondante. Sur la fonction d'autocorrélation, ces données relatives à une même direction seront situées sur une même droite, ayant aussi cette direction, et passant par l'origine. La figure **Fig. 3.2** donne l'exemple de calculs d'autocorrélation effectuée sur trois images différentes. Sur ces formes simples, on voit clairement que cette fonction permet d'identifier les orientations principales. La translation d'une droite dans sa propre direction va conduire à un fort niveau de correspondance qui se traduit par une valeur importante de la fonction d'autocorrélation dans la direction de celle-ci. Ce qui ne sera pas le cas si la direction du calcul est dans une autre direction.
Cette fonction a déjà été utilisée dans [Pra78, HOP+95] afin de caractériser des textures naturelles.



(a) Images testées



(b) Autocorrélations correspondantes

FIG. *3.2: Exemples de la capacité de la fonction d'autocorrélation à faire ressortir les orientations principales*

La définition de la fonction d'autocorrélation pour un signal bi-dimensionel est définie par

Eq. 3.1 :

$$C_{xx}(k,l) = \sum_{k'=-\infty}^{+\infty} \sum_{l'=-\infty}^{+\infty} x(k',l').x(k'+k,l'+l) \tag{3.1}$$

Pour une raison de temps de calcul, la fonction d'autocorrélation n'est pas, en pratique, calculée dans le domaine spatial. Pour cela on peut s'appuyer sur le théorème de Plancherel [11] afin d'effectuer les calculs dans l'espace des fréquences (via l'utilisation d'une FFT).

La rose des directions est un diagramme polaire qui permet d'analyser le résultat de la fonction d'autocorrélation. Soit (u,v) le point central de l'image après autocorrélation (par exemple les images de la figure **Fig. 3.2**.b) et la droite $D_{origine}$ l'axe des abscisses passant par ce point. Soit θ_i l'orientation étudiée, on calcule alors la droite Di telle que l'ensemble de ses points (a,b) respecte la relation suivante : angle formé par la droite (a,b) et passant par $D_{origine} = \theta_i$. Pour chaque orientation θ_i on calcule ainsi la somme des différentes valeurs de la fonction d'autocorrélation (**Eq. 3.2**).

$$R(\theta_i) = \sum_{D_i} C_{xx}(a,b) \tag{3.2}$$

Ces valeurs sont ensuite normalisées (**Eq. 3.3**) pour ne garder qu'un aspect relatif de la contribution de chaque orientation.

$$R'(\theta_i) = \frac{R(\theta_i) - R_{min}}{R_{max} - R_{min}} \quad avec \quad R_{max} \neq R_{min} \tag{3.3}$$

La figure **Fig. 3.3** revient sur la construction de cette rose des directions pour une forme simple. On remarque, entre autre, comment le calcul de $R'(\theta_i)$ permet de construire la rose pour l'orientation θ_i.

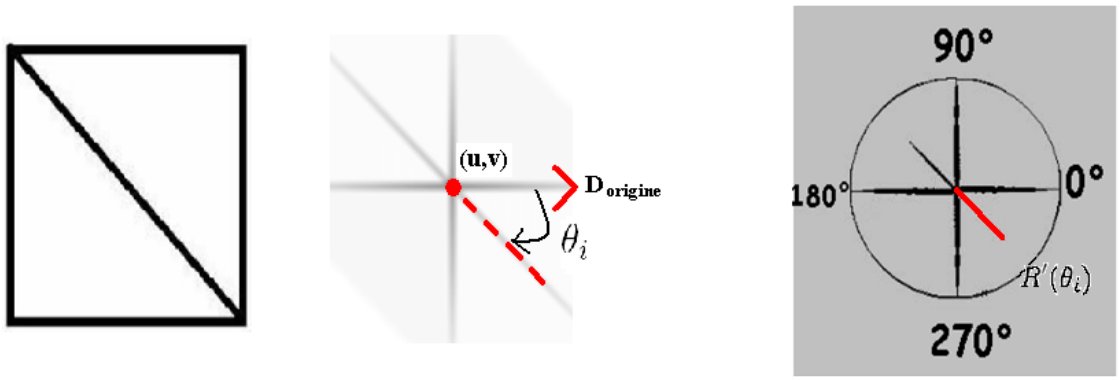


FIG. *3.3: Construction de la rose pour un angle θ_i*

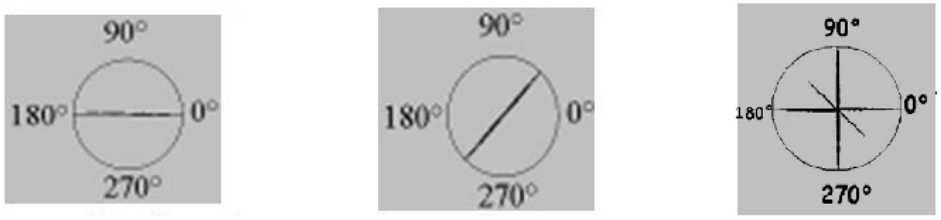
La figure **Fig. 3.4** illustre quelques exemples de rose des directions lorsqu'elle est appliquée sur des images. Ce diagramme polaire se lit comme un cercle trigonométrique. On se place au centre du cercle et on va vers l'extérieur suivant un angle θ afin de connaître la "quantité" de cette orientation θ dans l'image d'origine. Sur le cercle, plus un trait se rapproche du bord, plus

[11] matworld.wolfram.com/PlancherelsTheorem.html

l'orientation est présente dans l'image d'origine. Dans la figure **Fig. 3.4**, la rose correspondant à la flèche horizontale indique clairement que seules les orientations 0° et 180° sont présentes dans l'image d'origine.



(a) Images testées



(b) roses correspondantes

FIG. *3.4: Exemples de roses*

Nous allons maintenant étudier le comportement de la rose des directions sur des images de documents.

Par définition, la rose permet d'obtenir des informations sur les orientations privilégiées présentes dans l'image. La figure **Fig. 3.5**.a montre le comportement de la rose, dans le cadre de l'analyse d'une image composée uniquement d'un bruit dont la distribution est uniforme. Comme on peut s'y attendre, il n'y a pas de directions privilégiées et la rose prend la forme d'une boule pleine régulière. Si l'on ajoute une ligne noire horizontale à ce bruit (**Fig. 3.5**.b), la forme de la rose se modifie. Une analyse de cette rose indique que l'orientation la plus présente est l'orientation horizontale (le pic). Toutes les autres orientations sont présentes , mais de manière beaucoup moins significative que celle horizontale. Appliquée sur une zone homogène de texte (**Fig. 3.5**.c), la rose montre que les orientations présentes dans une telle image ont les mêmes caractéristiques que celle de la figure **Fig. 3.5**.b. Le "pic" est dû à la forte corrélation horizontale des pixels entre eux. Les autres orientations sont présentes, mais dans une moindre mesure. elles sont dues aux corrélations entre les lignes et à la forme des caractères. La caractérisation des textes par cet outils semble donc possible.

Comme le montre la figure **Fig. 3.6**, il n'existe pas un motif précis de rose qui permette de signer des zones de manière certaine. Il n'est donc pas possible de mettre en place un système de segmentation ou de caractérisation fonctionnant sur le principe de recherche de motif. Pour le texte, la forme de la rose est tributaire du nombre de lignes, de la taille des caractères, de l'orientation du texte (**Fig. 3.6**.d-f)... Pour les dessins, la même remarque peut être faite. La grande variété des illustrations existantes ne permet pas de définir un modèle homogène de rose des direction sensé "signer" une zone de dessin (**Fig. 3.6**.a-c).
Néanmoins, le calcul de la rose permet d'extraire des indices très riches en informations. Nous verrons par la suite, que l'orientation principale, la forme et l'intensité de la rose sont des indices

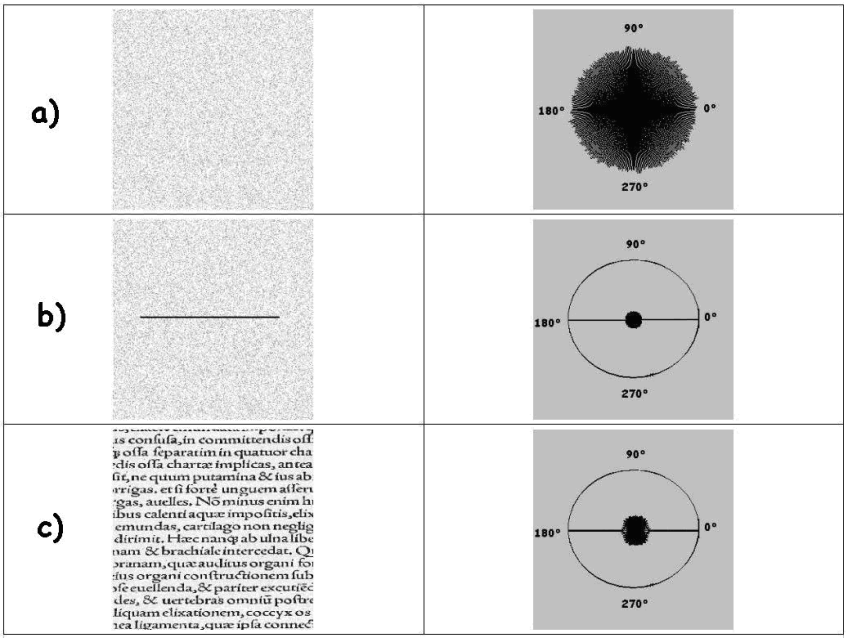


FIG. 3.5: *Exemples du comportement de la rose sur une image de document*

permettant d'effectuer une caractérisation fine du contenu.

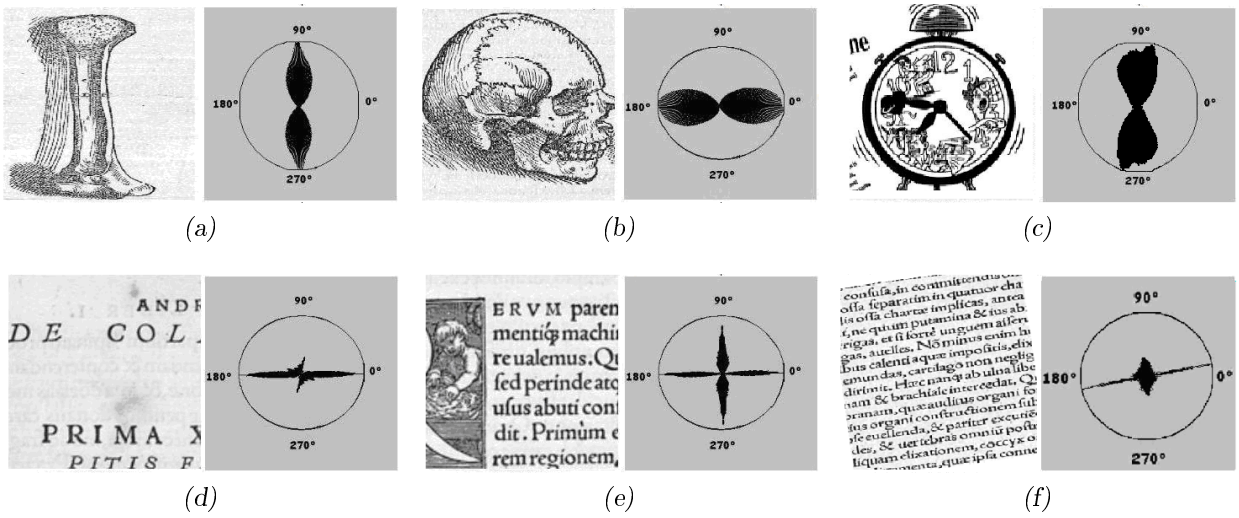


FIG. 3.6: *Exemple de roses des directions*

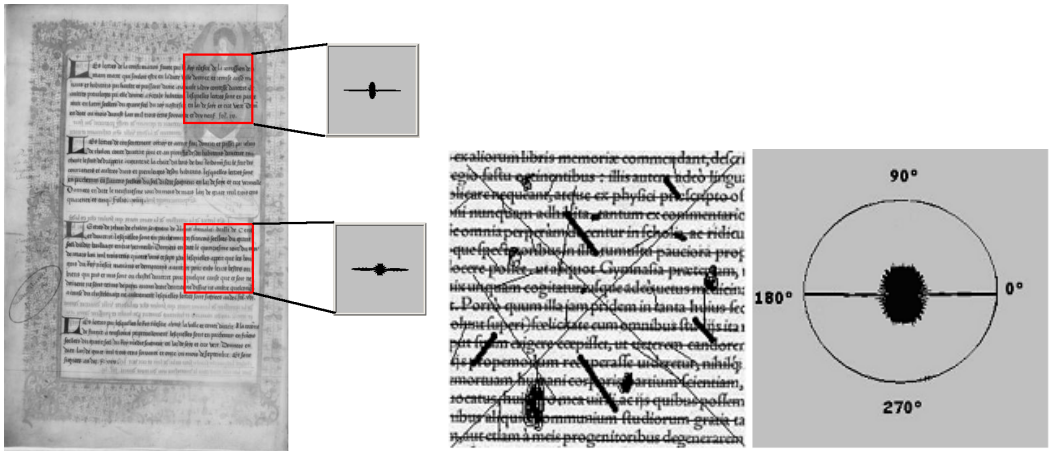
La figure **Fig. 3.8** montre le comportement de la rose des directions lorsqu'elle est calculée sur des images bruitées. Un défaut que l'on retrouve très fréquemment dans les documents anciens, est celui de l'apparition de l'encre du recto sur le verso de la feuille. Sur l'image **Fig. 3.7**.b on voit apparaître les illustrations du recto **Fig. 3.7**.a. Lors d'une binarisation, cette information peut être prise pour du dessin ou du texte **Fig. 3.7**.c. Le fait que la rose soit calculée en niveau de gris permet de ne pas être dépendant de ce type de bruit. On voit sur la figure **Fig. 3.8**.b, que même si la rose est légèrement différente (la boule du centre est légèrement difforme), l'information

principale (orientation horizontale importante) reste clairement identifiable.



(a) Image recto *(b) Image verso* *(c) Binarisation de l'image verso*

FIG. *3.7: Exemples du comportement de la rose sur une image de document bruitée*



(a) Rose sur une image en niveaux de gris *(b) Rose sur une zone de texte artificiellement bruitée*

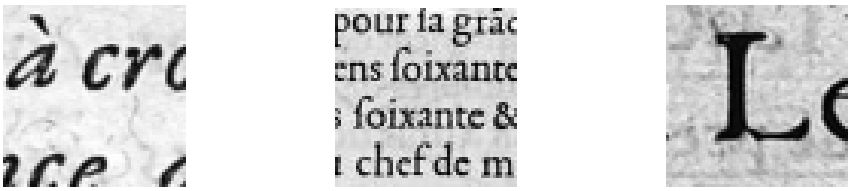
FIG. *3.8: Exemples du comportement de la rose sur une image de document bruitée*

Cet outil semble donc être adapté et robuste aux bruits inhérents aux documents anciens. il reste donc à extraire des caractéristiques pertinentes pour la caractérisation du contenu.

3.2.1.3 La rose des directions et la multirésolution

Comme la plupart des méthodes par approche texture, notre choix s'est porté sur une étude locale des caractéristiques de texture. En effet, la notion de texture sur un seul pixel n'a pas de sens. C'est une vision globale d'un pixel et de ses voisins qui donne cette notion de texture. Ainsi, notre approche se traduit par l'affectation d'une valeur numérique à un pixel de l'image, à l'aide d'une analyse par fenêtre glissante. Cette dernière est de taille fixe, elle est tout d'abord positionnée sur le coin haut gauche de l'image. Les 5 attributs textures sont alors calculés et le pixel central de la fenêtre est alors " marqué " des 5 valeurs numériques calculées. Enfin, la fenêtre se déplace d'un pixel et ainsi de suite jusqu'à avoir parcouru l'ensemble de la page.

Ce choix de marquage par fenêtre glissante pose la question douloureuse du choix de la taille de la fenêtre d'analyse. La figure **Fig. 3.9** montre à quel point la taille de la fenêtre est difficile à fixer. Chaque imagette est un extrait de trois corps de texte tirés de trois ouvrages différents. Chacune de ces imagettes fait 60X60 pixels et on voit que de l'une à l'autre on passe de 1 à 2 à 4 lignes pour une même taille de fenêtre d'analyse ce qui se traduit par des caractéristiques "textures" très différentes dans chacun de ces trois cas.



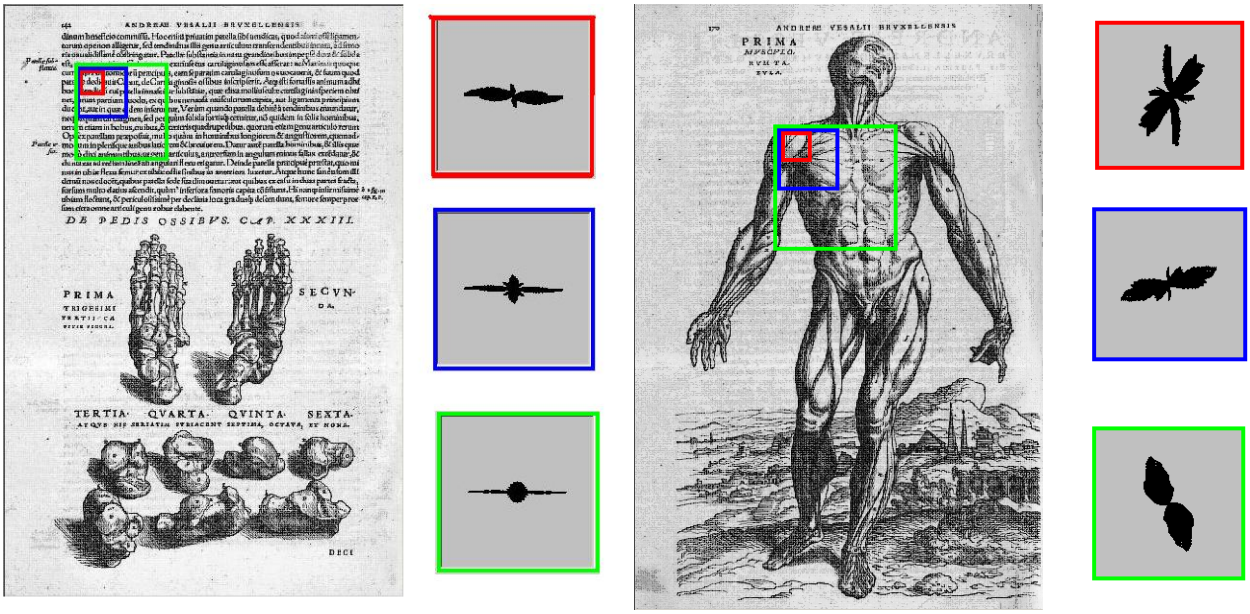
(a) première image *(b) deuxième image* *(c) troisième image*

FIG. *3.9: Difficulté de choisir la taille d'une fenêtre d'analyse*

Face à ce constat, nous avons décidé d'implanter une approche multirésolution. Ce choix a été aussi fait dans des publications traitant de segmentation texture par approche fenêtres. Dans [TZ00] [SG05] [SG04] et [CLM98], les auteurs utilisent la propriété suivante : en baissant la résolution d'une image, on obtient une vision plus approximative de cette dernière. Ainsi, on ne discerne pas l'information telle qu'elle existe réellement mais plutôt une vision grossière, un peu comme si l'on regarde une image de loin : on ne discerne que certaines caractéristiques de l'image (proportion des couleurs, présence de bruits, certaines formes...). Cette propriété est encore plus intéressante sur les images de documents car comme le dit [TZ00] : *"la beauté de la multi-résolution, c'est qu'elle a tendance à grouper, au fur et à mesure, les caractères en mots, les mots en lignes et les lignes en paragraphes"*.

L'intérêt de l'approche multirésolution vaut donc dans l'aspect perceptuel et dans son indépendance à la taille de l'image. La figure **Fig. 3.10**.a illustre l'intérêt d'un calcul de la rose sur une zone de texte pour différentes résolutions de l'image. Les 3 tailles de fenêtres utilisées génèrent 3 roses de formes différentes. On remarquera néanmoins que l'information de l'horizontalité ne varie pas. La figure **Fig. 3.10**.b illustre le même principe, mais cette fois-ci lorsque la rose est calculée sur une illustration. A l'inverse du calcul sur du texte, la rose présente de fortes variations dès lors que la fenêtre est de taille différente.

L'implantation de l'algorithme d'analyse multirésolution, a été réalisé en faisant varier la taille de la fenêtre d'analyse. En effet, il y a une dualité entre une première solution qui consiste



(a) Multirésolution sur du texte *(b) Multirésolution sur une illustration*

FIG. *3.10: Importance d'un calcul à différentes résolutions*

à fixer une taille de fenêtre et changer la taille de l'image et une deuxième solution qui consiste à garder la taille de l'image d'origine et de simplement faire varier la taille de la fenêtre d'analyse. La figure **Fig. 3.11** illustre cette équivalence. Pour le test, nous avons choisis une image de taille 889X600 sur laquelle on imagine avoir calculé des attributs pour une fenêtre de taille 64X64 **Fig. 3.11**.a). L'approche multirésolution reviendrait, par exemple, à diminuer la taille de l'image par 2 et à recalculer de nouveau les mêmes attributs **Fig. 3.11**.b). Pour simplifier l'analyse multirésolution, nous préférons garder la même taille d'image et changer la taille de la fenêtre d'analyse. On remarquera que cela ne change pas la propriété d'un calcul multirésolution. **Fig. 3.11**.c).

Pour effectuer notre analyse multirésolution des textures, nous avons choisi de prendre des fenêtres allant de 32X32 pixels à 256X256 (on multiplie la taille de la fenêtre par 2 à chaque fois), et ceci quelque soit la taille et le contenu de l'image d'origine. Ce nombre de fenêtres a été choisi par expérimentation. Il permet de traiter un large éventail de résolutions d'images.

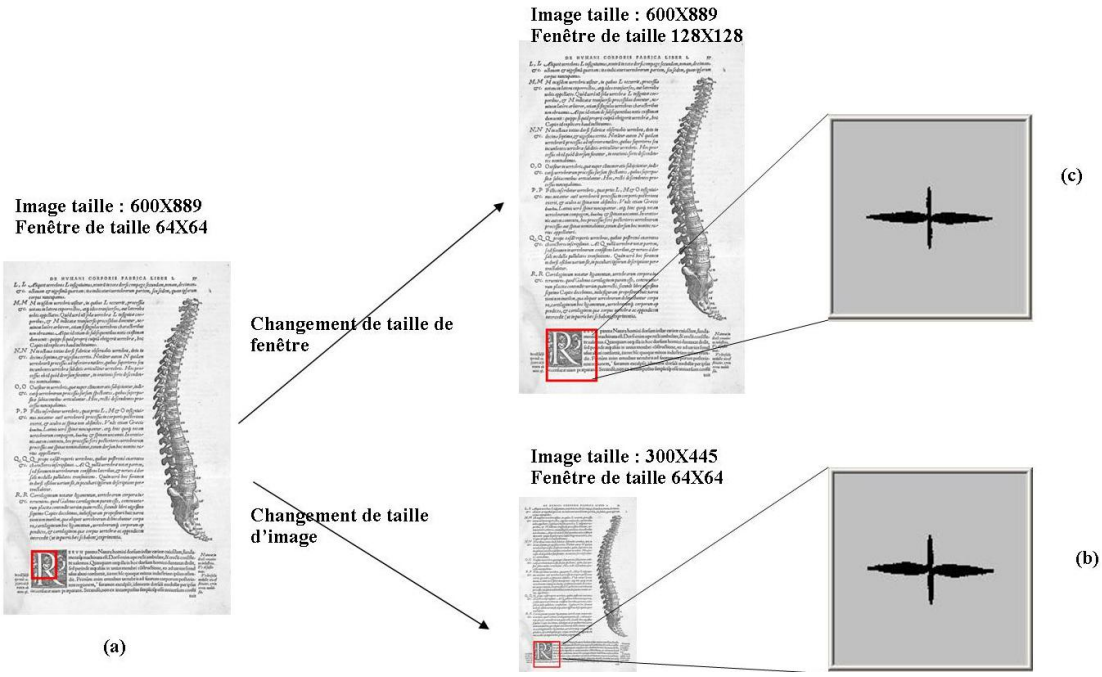


FIG. *3.11: Dualité entre changement de taille de fenêtre ou de taille d'image pour un calcul multirésolution*

3.2.2 Indices textures liés aux orientations

Comme nous l'avons vu dans les exemples précédents, le calcul de la rose permet d'identifier une orientation principale (orientation d'un texte, orientation principale d'une illustration) **Fig. 3.12**. La rose permet également de signer l'orientation de petites zones. La figure **Fig. 3.13** montre que sur une petite zone de texte l'orientation principale est horizontale. Cependant il se trouve que la forme de la boule au centre diffère. Dans le premier cas le texte est de style italique (**Fig. 3.13**.a) et la rose indique la présence significative d'orientations "oblique". Dans le deuxième exemple (**Fig. 3.13**.b), le texte est droit et la forme de la rose ne présente plus ce petit pic oblique.

On retrouve le même type de caractéristiques de la forme de la rose lorsqu'on l'applique sur des dessins de traits. La figure **Fig. 3.14** montre que selon les illustrations, la forme de la rose diffère. Cette double observation permet donc de conclure que non seulement, l'orientation principale est une information indispensable, mais que la forme de la rose l'est tout autant.

Une autre caractéristique riche en information, est l'intensité non normée de la fonction d'autocorrélation. En effet, pour des raisons de lisibilité, l'équation **Eq. 3.3** est utilisée pour normer les différentes valeurs et ainsi atténuer les différences. Par définition l'autocorrélation donne une information sur les évènements répétés, mais en aucun cas ne peut donner d'information sur la localisation dans l'image de ces évènements répétés. Dans notre cas, le calcul de la rose revient à étudier l'association des niveaux de gris de pixels selon une orientation précise. De ce fait, la valeur "R" calculée par **Eq. 3.2** est plus important quand l'image présente une forte anisotropie. Pour donner un ordre d'idée, sur les deux images testées dans la figure **Fig. 3.15**, l'autocorrélation moyenne (R) est presque 10 fois plus grande pour l'image **Fig. 3.15**.a que pour l'image

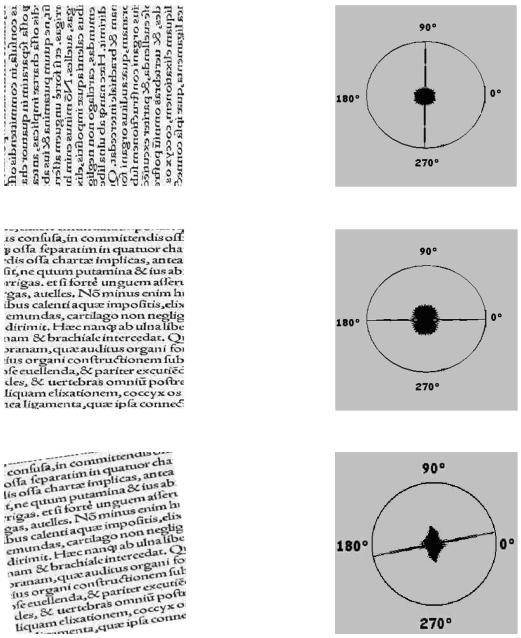


FIG. *3.12: Détection de l'orientation principale d'une image à l'aide de la rose des directions*

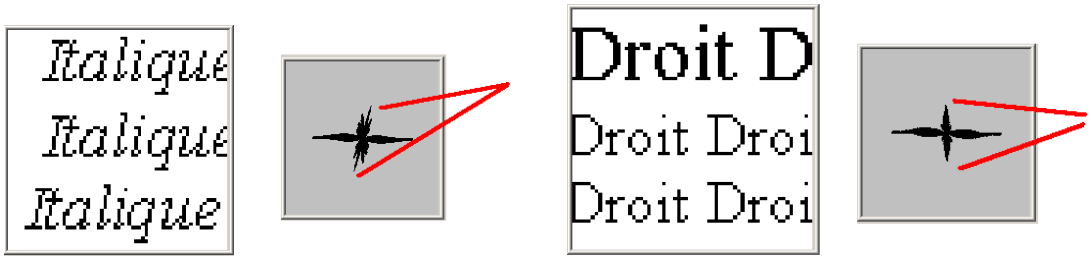


FIG. *3.13: Analyse des orientations à faible échelle*

Fig. 3.15.b. Il en va de même pour une image composée de lignes de textes, caractérisée par une forte anisotropie horizontale (**Fig. 3.15**.c), alors qu'une image composée d'un dessin de traits **Fig. 3.15**.d se caractérise par une forte isotropie. Ce constat s'explique de manière très intuitive par le fait qu'en tout point de l'image **Fig. 3.15**.c, il y a une forte corrélation qui se trouve être horizontale. Dans le cas de l'image **Fig. 3.15**.d il n'y a pas d'orientation privilégiée. La rose permet de trouver des corrélations dans les orientations portées par les traits noirs, mais cette isotropie de l'orientation se traduit par une faible corrélation.

Nous avons donc décidé d'extraire 3 indices permettant d'extraire les informations relatives aux orientations que nous venons d'exposer. Le premier indice extrait, est l'angle correspondant à l'orientation principale de la rose des directions (**Eq. 3.4**). Pour ne pas avoir à manipuler de

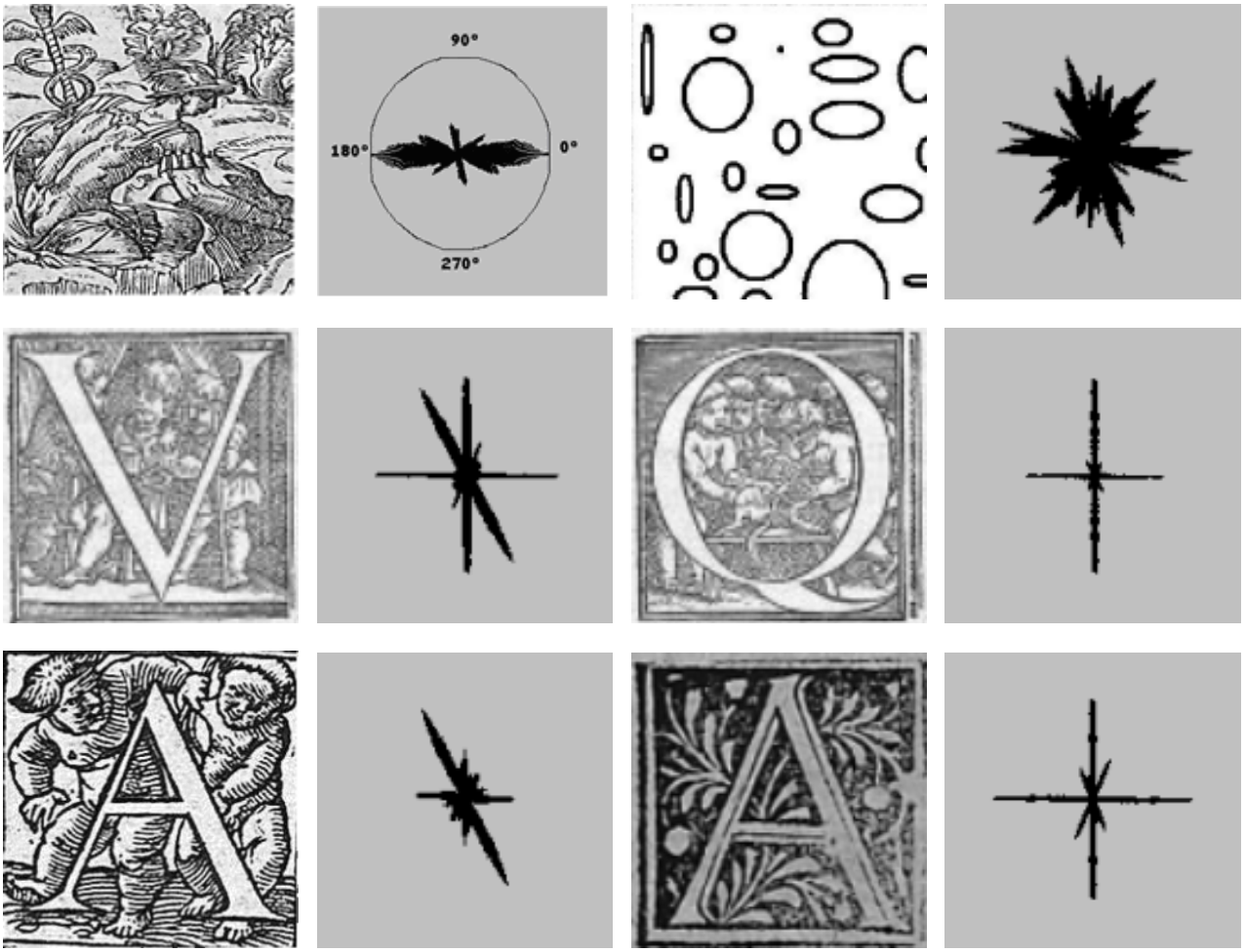


FIG. 3.14: Roses calculées sur des illustrations

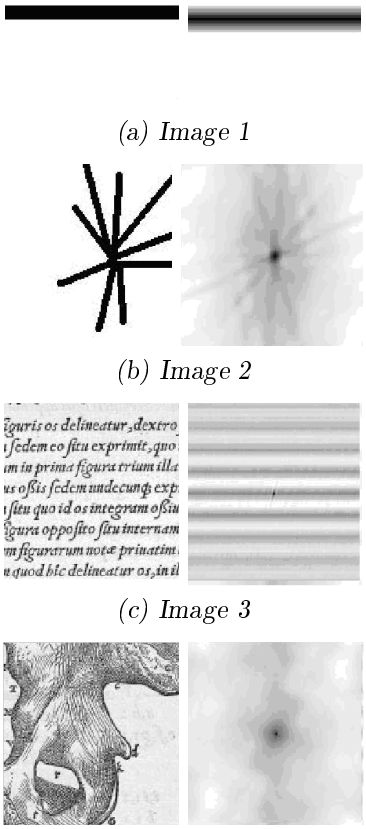
données circulaires, cet angle est normalisé en fonction de l'écart à l'angle horizontal.

$$\underset{s=A_1,\mathrm{r}\in\mathcal{R}}{C^{k}_{i,j}} = |180 - ArgMax(R'(i,j))| \tag{3.4}$$

L'isotropie de l'image est évaluée en fonction de l'intensité de la fonction d'autocorrélation. Ainsi, pour l'orientation principale trouvée par **Eq. 3.4**, chaque pixel sera caractérisé à l'aide de **Eq. 3.5**). Ce calcul est effectué sur la valeur non normée de la fonction d'autocorrélation.

$$\underset{s=A_2,\mathrm{r}\in\mathcal{R}}{C^{k}_{i,j}} = R(ArgMax(R'(i,j))) \tag{3.5}$$

Le dernier indice lié aux orientations caractérise la forme globale de la rose. Pour cela on calcule la variance des intensités de la rose des directions, excepté pour l'orientation d'intensité maximale. Ainsi, si la variance est faible, cela signifie que l'orientation principale est significativement plus présente que les autres orientations. Au contraire, si la variance est forte, cela



(a) Image 1

(b) Image 2

(c) Image 3

(d) Image 4

FIG. *3.15: Lien entre l'intensité de la réponse d'autocorrélation et l'isotropie/anisotropie des directions. Colonne de gauche : image d'origine, Colonne de droite : Autocorrélation de l'image*

signifie que la rose est difforme et qu'un grand nombre d'orientations sont présentes dans des proportions diverses.

$$\begin{aligned} \underset{s=A_3,\mathrm{r}\in\mathcal{R}}{C_{i,j}^{k}} &= Variance(R'(\theta)) \text{ avec } \theta \in [0,\pi] \\ &\text{Et avec } \theta \neq ArgMax(Rose(i,j)) \end{aligned} \tag{3.6}$$

Quelques exemples de la pertinence de ces trois premiers indices peuvent être illustrés. Pris indépendamment les uns des autres, ces indices ne permettent pas de segmenter ou de caractériser les contenus des images de documents anciens. C'est leur combinaison qui le permettra. Cependant, sur des images bien précises, il est possible de confirmer le pouvoir discriminant de certaines de ces caractéristiques textures. La figure **Fig. 3.16** est obtenue après application de l'équation **Eq. 3.4**. Pour l'affichage, chaque indice est normé entre 0 et 255. Plus un pixel de marquage est noir, plus il est " horizontal ", plus il est clair moins ils est " horizontal ". Le carré rouge correspond à la taille de la fenêtre d'analyse utilisée pour cet exemple. Dans ce cas assez simple, où 3 blocs de texture ont des orientations différentes, on retrouve bien l'information désirée. En effet, chaque bloc de texte est de couleur différente.

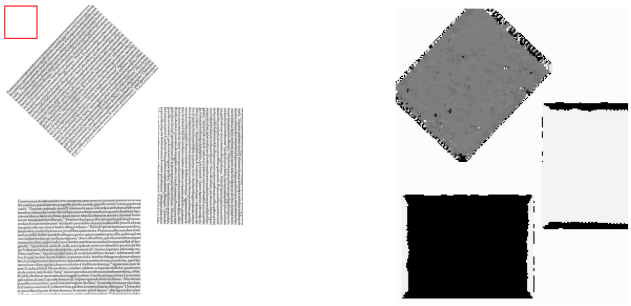


FIG. 3.16: *Comportement de l'indice* ***Eq. 3.4*** *sur des textures orientées*

La figure **Fig. 3.17** est obtenue après application de l'équation **Eq. 3.5**. Plus un pixel est noir plus **R** est faible. On aperçoit que les éléments graphiques apparaissent plus foncés que le texte et le fond. Ceci correspond aux hypothèses avancées précédemment. En effet, les dessins composant cette page ne comportent pas d'orientations privilégiées. De ce fait, les zones de texte auront tendance à générer une intensité de rose plus forte que sur les illustrations. Il est à noter que, si les illustrations étaient fortement directionnelles (par exemple des os, des traits horizontaux, ...), ce serait les illustrations qui présenteraient une intensité de rose élevée.

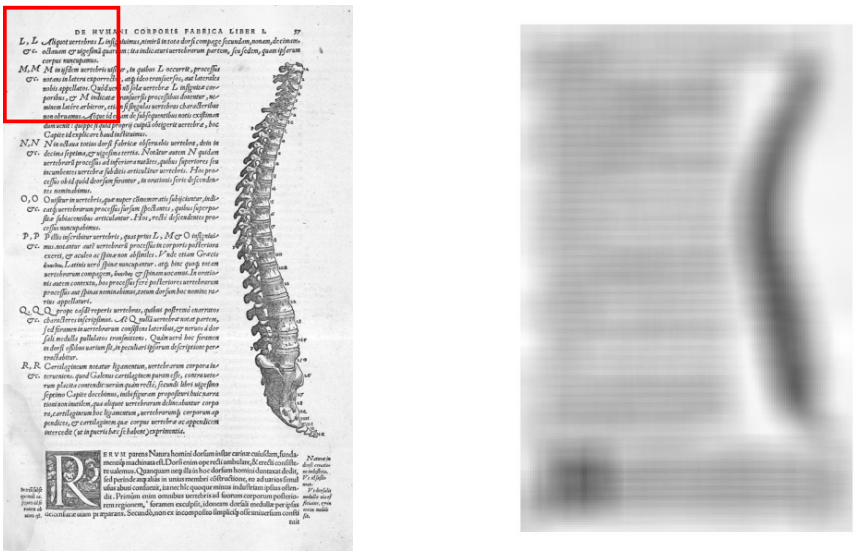


FIG. 3.17: *Comportement de l'indice* ***Eq. 3.5*** *sur des textures composées de zones isotropiques*

La figure **Fig. 3.18** est obtenue après application de l'équation **Eq. 3.6**. La partie supérieure est composée d'une multitude de cercles disposés aléatoirement, la partie inférieure est composée de lignes de textes horizontales. Après normalisation, plus un pixel est noir, plus la variance est faible. Ce résultat reste conforme a ce qui a été montré précédemment. Les cercles n'ont pas de directions privilégiées et ont, de ce fait, une rose d'apparence irrégulière.

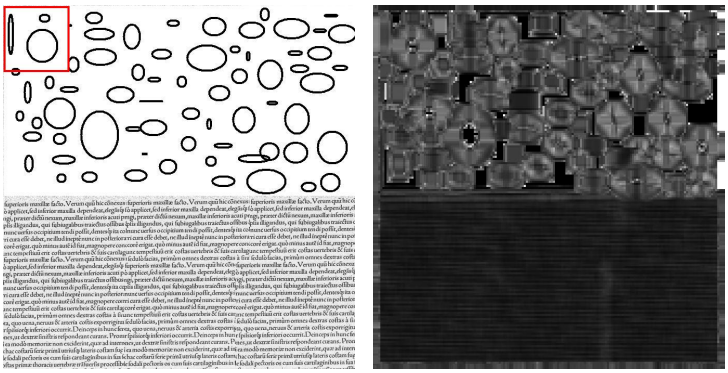


FIG. *3.18: Comportement de l'indice* ***Eq. 3.6*** *sur des textures composées*

3.2.3 Indices textures liés aux fréquences

En complément des informations liées aux orientations, nous allons extraire en supplément des indices liés aux fréquences. Nous l'avons évoqué dans le chapitre précédent, de nombreux auteurs s'appuient sur cette catégorie d'indices afin, de segmenter des images (de documents ou non) ou de caractériser des blocs segmentés. La notion de "fréquence" sur des images de documents, est lié à la fréquence de transition entre le papier et l'encre. Les tests réalisés dans le chapitre précédent ont montré les limites d'outils tel que Gabor. Qui plus est, la paramétrisation qu'implique l'utilisation de ces outils, rend complexe leur utilisation.
Afin de caractériser les fréquences de transitions, nous avons préféré nous inspirer des travaux de [Egl98, All04, CWS03]. Ces auteurs 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. Ce qui revient à exploiter des indices de fréquences.
Notre apport se situe à deux niveaux différents. Tout d'abord, les travaux cités ci-dessus exposent des méthodes de caractérisation sur des blocs pré-segmentés. Il faut donc un indice qui soit capable d'identifier les fréquences de transitions présentes dans une image, sans que celui-ci ne soit dédié à l'analyse de police, de styles de caractères, d'illustrations... En deuxième lieu, il se trouve que la plupart de ces travaux travaillent sur des images binarisées. Par exemple, [Egl98] caractérise des blocs de texte à l'aide d'une mesure d'entropie. L'auteur calcule les probabilités de transition noir/blanc par ligne, et conclut sur la nature des textes étudiés. D'une manière générale, la difficulté de la phase de binarisation pousse à la retarder au maximum dans une chaîne de traitements. C'est pour cela, que nous proposons deux mesures calculables sur des images en niveaux de gris.

Le premier indice que nous allons utiliser permet de caractériser les fréquences de transition entre l'encre et le papier. Pour chaque ligne de la zone analysée par la fenêtre glissante, on somme la différence de niveau de gris d'un pixel et de son voisin de gauche. Plus la somme est élevée, plus le nombre de transitions sur une ligne est élevé. Un simple calcul de moyenne permet d'obtenir un indice sur les transitions de la zone étudiée (**Eq. 3.7**).

$$\underset{s=A_4,\mathrm{r}\in\mathcal{R}}{C^{s,r}_{i,j}} = Avg(\sum_{i\in I'}(\sum_{j\in J'}(p^k_{ij} - p^k_{ij+1})) \tag{3.7}$$

Avec I' et J' la taille de la fenêtre d'analyse.

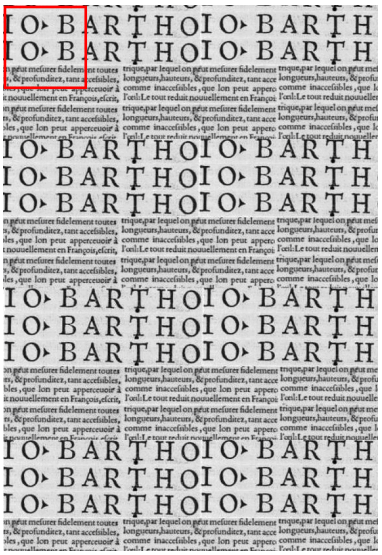
La figure **Fig. 3.19** donne quelques exemples du comportement de l'indice **Eq. 3.7** appliqué à différents types de textures. La normalisation a été choisie de telle sorte que plus un pixel appartient à une zone de hautes transitions, plus il est noir. Sur la figure **Fig. 3.19**.a-b, on constate que les zones composées de petits caractères sont noires après calcul de l'indice. Ceci correspond à ce qui était attendu. En effet, une petite police de caractères implique un plus grand nombre de transitions entre l'encre et le papier, qu'une grande police. La figure **Fig. 3.19**.c, est composée de deux textures naturelles, d'une zone de texte oblique, et enfin d'une zone composée de quelques traits noirs horizontaux. Après application de l'indice, les traits noirs horizontaux ont disparu. Ce résultat s'explique tout simplement par le fait qu'il n'y a pas de transitions horizontales(excepté aux bords). On voit ensuite clairement, que la zone de texte possède des transitions plus importantes que les images naturelles, puisqu'après calcul de l'indice tous les pixels du coin supérieur haut sont noirs. Ce qui signifie que cette zone possède des transitions très importantes.

Le dernier indice texture calculé, est inspiré des travaux d'analyse des longueurs de plage. Il est inspiré de [RBD06] qui propose un algorithme de caractérisation des plages blanches entre les composantes connexes. Nous recherchons ainsi un moyen d'obtenir des informations sur l'étendue des diverses zones de fond qui jalonnent les pages. Comme pour l'indice précédent nous proposons, une adaptation de cet indice afin de pouvoir le calculer sur une image en niveaux de gris. Nous avons aussi adopté une approche récursive de cet algorithme. Cet indice n'est donc pas calculé par une approche par fenêtre glissante et multirésolution. Notre approche consiste ainsi à utiliser 4 phases d'un algorithme XY-cut récursif. A chaque itération on coupe en quatre zones de taille identique celle qui vient d'être analysée et on applique à nouveau le même indice. Cet indice est, pour chaque pixel, égal à la moyenne de la somme des niveaux de gris en colonne et en ligne (**Eq. 3.8**).

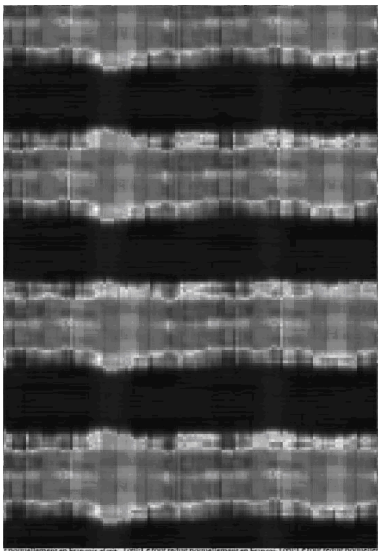
$$\underset{s=A_5,r\in\mathcal{R}}{C^{k}_{i,j}} = \frac{\sum_{l\in J'} p^k_{il} + \sum_{h\in I'} p^k_{il}}{2} \tag{3.8}$$

Avec I' et J' la taille de la fenêtre d'analyse à l'itération k de l'algorithme récursif.

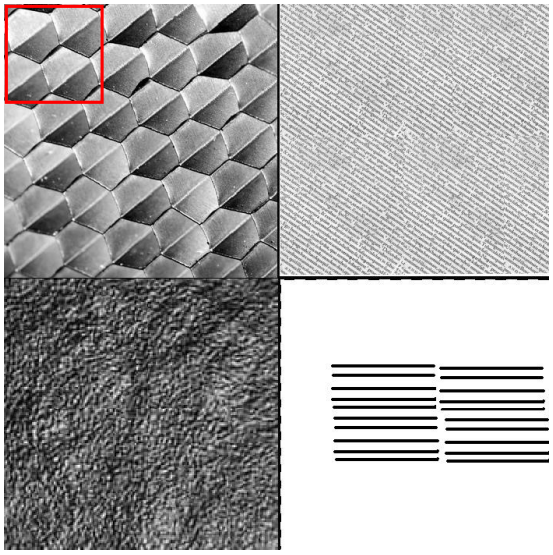
La figure **Fig. 3.20** donne quelques exemples du comportement de l'indice **Eq. 3.8** appliqué à différents types de textures. La normalisation a été choisie en fonction de la longueur des plages blanches de la zone étudiée. On aperçoit sur la figure **Fig. 3.20**.a-e, des variations de niveau de gris dans les différentes zones (les espaces inter-lignes sont "moins blancs" que les marges). Le fait d'avoir calculé l'indice à différentes itérations de l'algorithme récursif, permet de d'obtenir des informations différentes sur la structure de la page. Même si le texte n'est pas horizontal (**Fig. 3.20**.f-j) cet indice permet d'obtenir des informations pertinentes sur la structure.



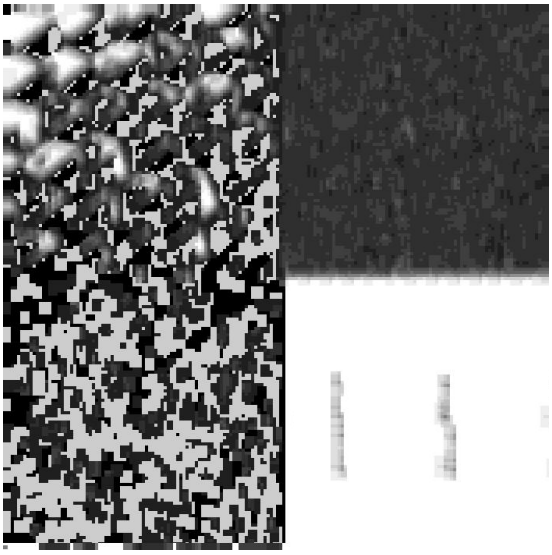
(a) Image composée de deux tailles de texte différentes



(b) Indice calculé ***Eq. 3.7***



(c) Image composée de textures différentes

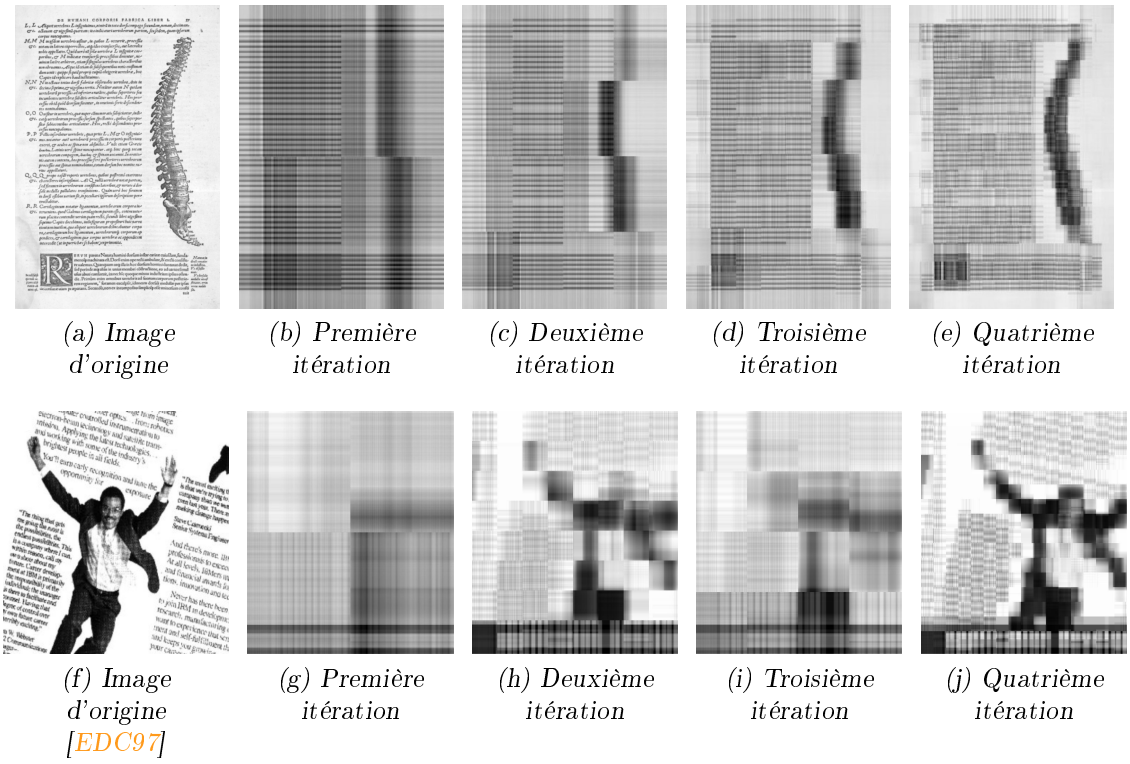


(d) Indice calculés ***Eq. 3.7***

FIG. *3.19: Comportement de l'indice* ***Eq. 3.7*** *sur des textures composées de fréquences de transitions différentes*

3.2.4 Conclusion

Dans cette section, nous avons détaillé notre proposition d'attributs textures dédiés aux documents. Chacun d'entre eux permet d'exprimer une caractéristique liée aux fréquences et aux orientations des motifs présents dans les images. Ils répondent à des informations visibles liées à la distribution des traits sur une page de texte ou une page composite. ils permettent ainsi d'exprimer une forme de régularité (dans la distribution des transitions aux frontières des lettres dans les zones de texte) ou au contraire un plus grand désordre (plus généralement dans



(a) Image d'origine (b) Première itération (c) Deuxième itération (d) Troisième itération (e) Quatrième itération

(f) Image d'origine [EDC97] (g) Première itération (h) Deuxième itération (i) Troisième itération (j) Quatrième itération

FIG. 3.20: Exemple de l'indice récursif **Eq. 3.8**, calculé sur deux images différentes.

les zones graphiques).
A travers quelques exemples, nous avons montré que ces attributs semblent être pertinents. Cependant, pris indépendamment les uns des autres, ces attributs ne sont pas suffisants pour caractériser le contenu d'une image. Dans la section suivante, nous allons voir dans quelle mesure, en associant ces attributs, il est possible de réaliser une caractérisation générique des contenus d'images de documents sans pour autant segmenter ou extraire cette structure.

3.3 Analyse et pertinence des données extraites pour la caractérisation des contenus

Dans la section précédente, nous avons décrit l'ensemble des indices textures de caractérisation de contenu d'images de documents. Sur quelques exemples bien précis, nous avons illustré la pertinence des algorithmes de caractérisation. Mais ces figures, restent des images choisies pour illustrer au mieux la pertinence de nos extracteurs de textures. Cependant, ils ne garantissent en aucun cas leur utilisabilité dans le cadre d'une exploitation à plus grande échelle. Cette section détaille deux points de réflexion, qui nous semblent important de vérifier, avant de passer à l'étape finale consistant à expérimenter des outils s'appuyant sur ces nouveaux indices. La première partie est dédiée à la l'évaluation de la catégorisation du contenu au travers d'une classification des pixels sur la base des 20 indices de textures proposés. La deuxième partie détaille les résultats provenant d'une analyse factorielle de nos données. En effet, même si nous avons voulu réaliser une approche la moins paramétrique possible (pas de seuils, pas de description des documents

traités...), il n'en reste pas moins que certains choix demeurent obligatoirement arbitraires. Ces choix concernent essentiellement les interrogations suivantes : quelle est l'influence de la taille et de la résolution de numérisation des images sur notre méthode ? Comment vont se comporter les algorithmes dès lors qu'ils seront appliqués à grande échelle ? le calcul des indices à différentes résolutions apporte t'il une réelle information supplémentaire ? etc...

3.3.1 Classification automatique du contenu

Classifier les éléments de contenu des ouvrages, permettra d'une part de vérifier la pertinence des informations extraites, et d'autre part de vérifier si la caractérisation des contenus est conforme à l'objectif de séparation de l'information en couches, lorsqu'elle est opérée sur un ouvrage complet (ce qui correspond à un cas concret d'utilisation envisagé).

3.3.1.1 Algorithme de classification choisi

La première étape consiste à vérifier la caractérisation des contenus en prenant en compte l'ensemble des indices calculés. Dans cette hypothèse, chaque pixel de l'image dispose de 20 valeurs issues des 5 indices calculés à 4 résolutions différentes. Notre objectif étant de regrouper les pixels de l'image correspondant à des zones homogènes, ce qui revient à regrouper des vecteurs caractéristiques sensés être proches au sens d'une métrique. C'est un problème de classification non supervisée pour lequel nous ne connaissons pas a priori les étiquettes des points permettant de construire les classes. Notons ici que c'est un problème complexe vu la dimension du vecteur. Nous suivrons ici une démarche pragmatique sans pondération spécifique des valeurs du vecteur.

Nous avons testé 3 algorithmes de classification par centres mobiles.

- **Clara (Clustering LARge Applications)** : Clara permet de manipuler des vecteurs de grande dimension et un nombre important de données. Cet algorithme opère en deux étapes. Dans un premier temps la classification d'un échantillon de pixels (choisis aléatoirement) est calculée en utilisant l'algorithme PAM. Le nombre de classes est un paramètre de l'algorithme. Une fois cette première étape achevée, les pixels n'appartenant pas à l'échantillon sont classés dans une des partitions, en fonction du plus proche voisin (sans rejet possible). Clara est décrit en détail dans [KR90]. Intéressons nous plus précisément à cette première étape. L'algorithme PAM recherche les objets représentatifs qui sont situés au centre des classes qu'ils définissent. L'objet représentatif d'une classe, le médoïd, est l'objet pour lequel la dissimilarité moyenne par rapport à tous les objets dans la classe est la plus petite. En réalité, l'algorithme PAM minimise la somme des dissimilarités au lieu de la dissimilarité moyenne. Le critère calculé est la somme de distance de tous les points à leur médoïd. C'est cette distance qui est à minimiser. La version de Clara utilisée est en $O(kn)$. Avec k le nombre de classes entrées en paramètre, et n le nombre total de points.
- **K-means :** Il en existe plusieurs versions, mais le principe reste le même. Dans les tests que nous avons réalisés, chaque classe est représentée par sa moyenne des points. De manière itérative, cet algorithme va affecter chaque point dans l'une des k classes et recalculer les centres. Dans notre implantation, la classification se termine quand il n'y a plus de changements dans les centres de classe. La complexité est en $O(nki)$ avec le n nombre de points, k le nombre de clusters et i le nombre d'itérations.
- **K-median :** Même principe que K-means, excepté que c'est un point médian qui est calculé. La complexité est donc plus élevée. Cette version, au contraire des deux autres, permet d'être robuste aux points aberrants.

Il existe d'autres algorithmes de classifications que nous n'avons pas testés. Ce qui a motivé le choix des 3 algorithmes présentés ci-dessus est avant tout la capacité de ces algorithmes à traiter de gros volumes de données.

3.3.1.2 Classification des pixels d'une page

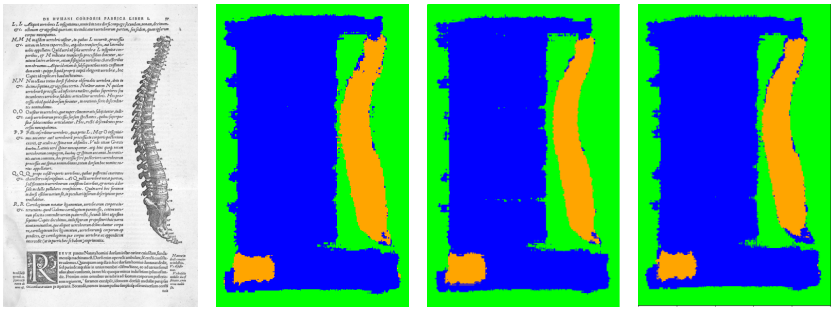
Appliquer un algorithme de classification sur les données extraites était avant tout pour nous, un moyen d'avoir un premier retour sur la pertinence des indices calculés.

La figure **Fig. 3.21** présente un échantillon d'une centaine de pages testées avec les trois algorithmes. Notre choix s'est finalement porté sur l'algorithme Clara. Les résultats étant de même facture, nous avons opté pour celui dont la complexité était la plus faible. L'autre avantage de Clara est que le nombre d'échantillons nécessaires est faible. Les auteurs de Clara indiquent que $40+2k$ (k le nombre de classes) échantillons sont suffisants pour obtenir un bon partitionnement.

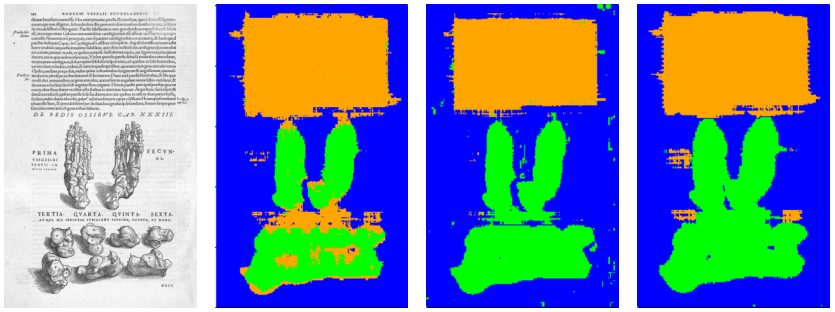
Les résultats que nous allons montrer par la suite ne doivent pas être vus comme étant des résultats de segmentation mais plutôt comme une illustration de la pertinence des indices textures proposés. En effet, comme la plupart des approches textures, c'est un marquage des pixels qui est effectué et non pas une segmentation en zones homogènes (cette différence n'est pas toujours mentionnée dans certains articles de la littérature). Pour segmenter en zones homogènes, il faudrait, par exemple, effectuer des post-traitements comme ceux que fait Doermann dans [EDC97] après un marquage par approche texture.

La figure **Fig. 3.22** donne une première indication sur le pouvoir discriminent des indices. Si dans ces exemples le choix du nombre de classes reste arbitraire, on voit néanmoins d'autres informations intéressantes apparaître en fonction de ce paramètre. 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 composition très variable sont parfois moins bien extraits. Lorsque l'on augmente le nombre de classes à séparer, on remarque d'autres informations intéressantes apparaître comme par exemple les espaces inter-lignes (partie c de **Fig. 3.22**)

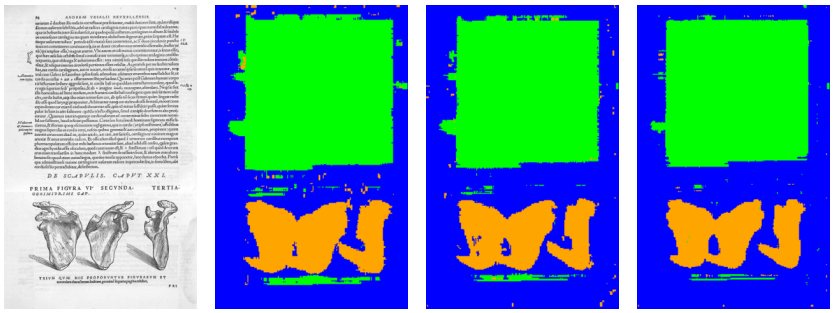
Nous avons également testé cette classification de pixels sur des images de documents contemporains. La figure **Fig. 3.23** montre des exemples de résultats obtenus sur des images extraites de [MD05, KIM99]. Globalement, les résultats de marquage sont corrects (bonne séparation texte/fond/dessin). Cependant, ces résultats semblent de moins bonne qualité que ceux obtenus dans les références dont les images sont tirées.



(a) Image d'origine (b) Clara (c) kmeans (d) kmedian

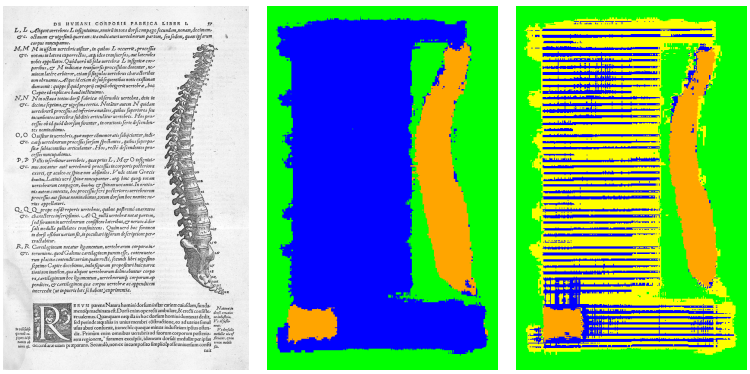


(e) Image d'origine (f) Clara (g) kmeans (h) kmedian

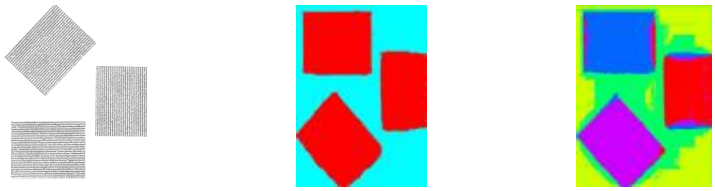


(i) Image d'origine (j) Clara (k) kmeans (l) kmedian

FIG. 3.21: Classification de pixels avec différents algorithmes

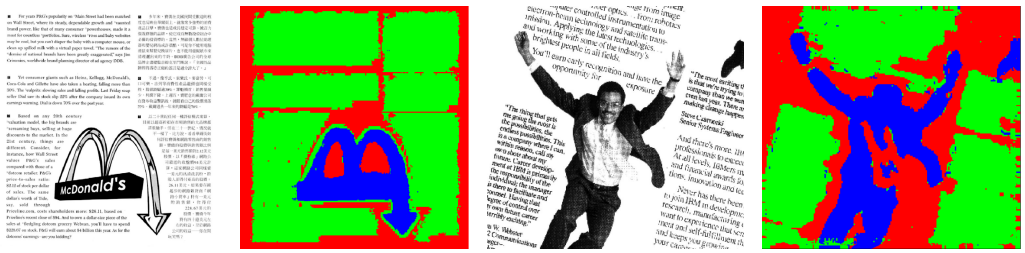


(a) Image d'origine (b) classification à 3 classes (c) classification à 4 classes

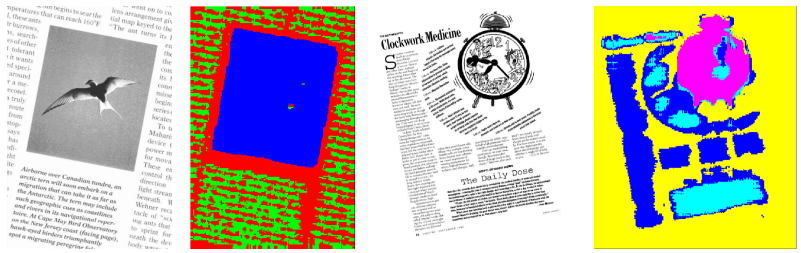


(d) Image d'origine (e) classification à 2 classes (f) classification à 5 classes

FIG. 3.22: Classification à 3 classes de pixels avec Clara



(a) Image d'origine (b) Clara à 3 classes (c) Image d'origine (d) Clara à 3 classes



(e) Image d'origine (f) Clara à 3 classes (g) Image d'origine (h) Clara à 4 classes

FIG. 3.23: Classification de pixels de documents contemporains (images issues de [MD05, KIM99])

3.3.1.3 Classification des pixels d'un ouvrage complet

La figure **Fig. 3.24** montre le type de résultats que l'on obtient lorsque l'on effectue une classification sur un ouvrage complet. Techniquement cela revient à considérer un ouvrage comme une seule image ou toutes les pages seraient "collées" les unes à la suite des autres. De ce fait, si deux pixels ont la même couleur, cela signifie qu'ils appartiennent à la même classe. Cette classification a un réel sens. En effet, elle permet d'obtenir un point de vue global des textures se ressemblant dans un ouvrage complet. Cela permet donc de fixer un nombre de classes pour un ouvrage, et non pas page par page. En effet, si la classification était réalisée page par page, il serait alors complexe de déterminer, a priori, le nombre de classes présentes dans l'image analysée. Effectuer une classification à 3 classes sur l'ouvrage complet permet donc, très simplement, d'obtenir une séparation texte/fond/dessin sur l'ouvrage. On notera par exemple sur la figure **Fig. 3.24** certaines pages comportent 3 classes et d'autres uniquement 2 car il n'y a pas d'illustrations présentes.

D'un point de vue purement qualitatif, nous obtenons des résultats encourageants, d'autant plus que nous utilisons l'algorithme Clara sans adaptation et sans traitement des données (exceptée une opération de centrage/réduction). Ces tests ont, avant tout permis de mettre en évidence un réel pouvoir séparateur cohérent des indices extraits. En ce qui concerne les principales limites de cette étude, elles se localisent au niveau de l'analyse de zones de transition entre textes et images, mais aussi de titres contenant de gros caractères isolés. De ce fait, une grande partie des titres (isolés du corps de texte) sont identifiés comme étant du dessin. De même, des dessins dont le trait est très fin ou de faible densité (par exemple les os de **Fig. 3.24**) ou encore les dessins qui sont très proches d'une zone de texte ne sont pas clairement marqués (par exemple une lettrine dans **Fig. 3.24**).

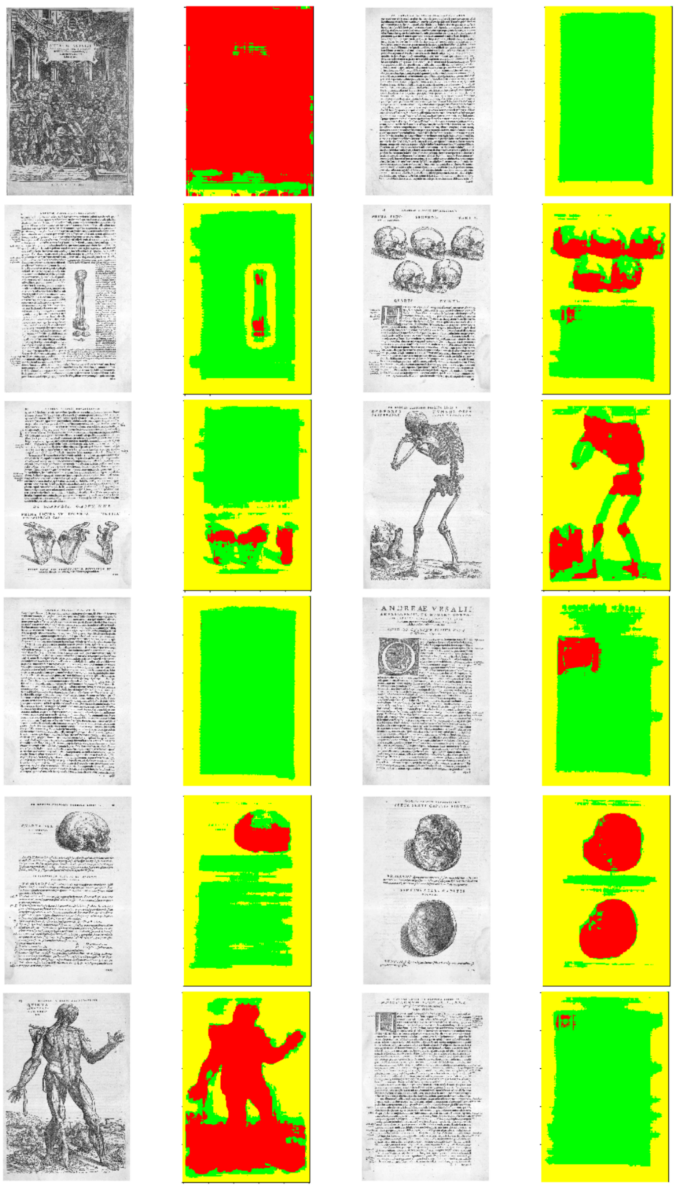


FIG. 3.24: *Classification de pixels d'un ouvrage avec Clara*

3.3.1.4 Proposition d'une évaluation de la catégorisation

Dans cette section, nous proposons une évaluation simple qui n'a pas pour but d'être comparée aux autres algorithmes de segmentation de la littérature. L'intérêt est de pouvoir donner une tendance sur les capacités de nos données plutôt que de donner une grande quantité de résultats visuels. Nous avons donc décidé, dans cette partie, de nous caler sur ce qui se fait généralement dans la littérature c'est à dire l'évaluation d'une séparation des pixels en 3 classes : texte/dessin/fond.

Pour ce faire, nous avons saisi manuellement une vérité terrain à l'aide d'une application que nous avons développée et qui permet de délimiter à la souris les contours des dessins et des zones

de texte. L'intérieur de la zone délimitée est ensuite remplie automatiquement. Une fois toute la vérité saisie, un fichier est créé afin de stocker cette vérité terrain pour être finalement comparée à la classification calculée par Clara **Fig. 3.25**.

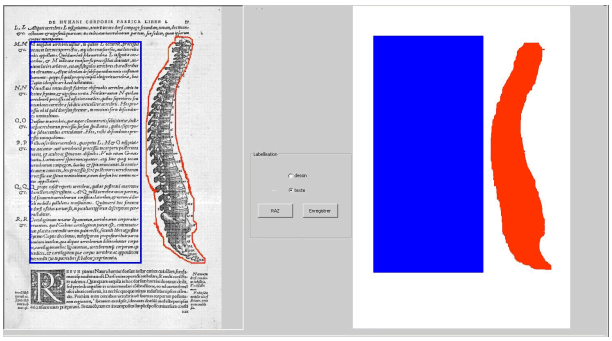


FIG. *3.25: Logiciel de saisie terrain*

Nos tests ont été réalisés sur 200 pages de documents anciens, extraient de 9 ouvrages différents. Naturellement, ce chiffre de 200 pages peut paraître surprenant si l'on tient compte que notre corpus est composé de plusieurs centaines d'images. Ce choix est à mettre en relation avec le contenu des images du corpus. Lorsqu'on le détaille, on s'aperçoit que beaucoup de pages sont composées uniquement de texte (Environ 85% des pages du corpus). Hors, comme nous l'avons évoqué juste avant, les problèmes de marquage pour une séparation texte/dessin/fond viennent des zones de transitions entre le texte et les dessins de traits. Appliquée sur des pages composées uniquement de lignes de textes, il se trouve que le taux de bonne classification est proche de 100%(cf les exemples de **Fig. 3.24**). Etant donnée que nous voulions avoir une idée de la pertinence des indices extraits, nous avons travaillé sur des images au contenu le plus varié possible.

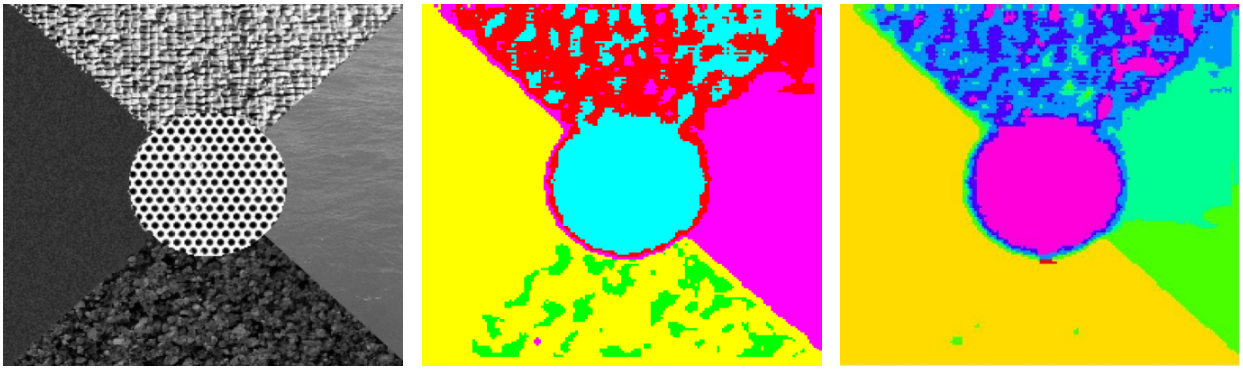
Les résultats **Table 3.1** montrent le potentiel de catégorisation de notre approche. En comparant la vérité terrain et la classification obtenue sur l'ensemble des pages de la base de tests, nous avons pu établir les taux de reconnaissance donnés par le tableau **Table 3.1**. Nous avons donc opté pour un "comptage" des pixels. Etant donné que l'image "vérité terrain" est de la même taille que l'image générée après classification, nous regardons simplement si les pixels sont étiquetés avec le même label. L'algorithme de classification utilisé est Clara sur un ouvrage complet (comme pour **Fig. 3.24**).
Dans la plupart des cas les différentes parties de l'image sont correctement détectées. Les erreurs principales proviennent de l'analyse de zones de transitions entre textes et images mais aussi de titres où de gros caractères sont utilisés. Exceptés ces deux cas spéciaux notre méthode donne de bons résultats et nous permet d'atteindre notre objectif principal qui est la caractérisation des pages du document afin de mettre en avant le contenu visuel du document.

Dessin	Texte
83%	92%

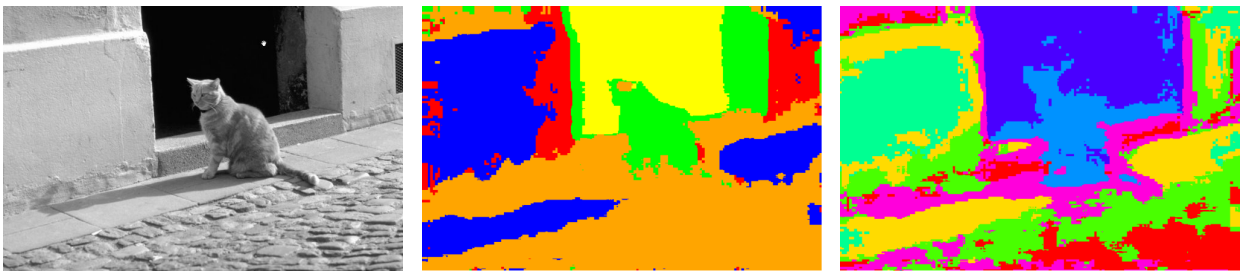
TAB. *3.1: Evaluation de la qualité du marquage*

3.3.1.5 Conclusion sur la classification de pixels

Les tests réalisés sur des images de documents, ont montré la pertinence des indices extraits. L'étude des orientations et des transitions entre niveaux de gris sont donc des informations permettant de discriminer les différents éléments du contenu. Ces indices ont été réalisés dans l'objectif de traiter des documents.
Nous avons réalisé quelques tests sur des images naturelles. Aux vues des résultats, les indices semblent utilisables pour le traitement de ce type d'images(**Fig. 3.26**.a-c). Le problème principal tient au fait qu'il est complexe de fixer le nombre de classes à segmenter. La figure **Fig. 3.26**.d-f montre qu'une classification à 5 classes semble visuellement correcte. Mais, Si l'on augmente le nombre de classes, les différentes zones de l'image sont nettement moins bien identifiées. En effet, les frontières sont moins nettes : la répartition des hautes et basses fréquences sur ces images est plus aléatoire, on n'a pas une distribution en noir et blanc (comme dans le cas du texte sur du fond).



(a) image d'origine [Ros99] *(b) Classification à 5 classes* *(c) Classification à 6 classes*



(d) image d'origine [CCMV03] *(e) Classification à 5 classes* *(f) Classification à 7 classes*

FIG. *3.26: Classification de pixels d'images naturelles*

3.3.2 Analyse factorielle des indices textures

Compte tenu de la quantité de données/indices générés, une analyse de nos données est une étape obligatoire. Le fait de calculer des indices à différentes résolutions nous fait prendre le risque de créer une information redondante (et donc inutile). Cette section permet d'apporter des réponses à ce type de questions et validera ainsi certains choix.

Dans une première partie nous présentons brièvement l'outil que nous allons utiliser. Nous donnons aussi certaines explications afin de permettre la lecture de nos "fichiers résumés". Dans une deuxième partie, nous détaillons les tests effectués sur nos données et les résultats obtenus.

3.3.2.1 Quelques explications sur notre façon de procéder

Etant donné les questions que nous nous posons sur les données générées, nous avons opté pour l'utilisation de l'analyse en composantes principales (ACP). Les tests effectués par la suite, s'appuient sur l'interprétation des résultats d'une ACP. Voici quelques rappels très succincts sur certaines notions utilisées dans cette section, dès lors que nous les appliquons sur nos indices. On trouvera de plus amples informations dans [BS78].

- **le pourcentage d'inertie :**
  Le % d'inertie correspond à la quantité d'informations portées par les axes factoriels. Si l'on désire compresser les données initiales à l'aide d'une ACP, l'inertie permet de donner une idée de la quantité d'informations perdues en réduisant le nombre de variables.

- **Les vecteurs propres :**
  Ce sont les vecteurs qui définissent le nouvel espace dans lequel sont projetées les données initiales. Ils permettent de passer d'une dimension p à une dimension q, avec p>=q.

- **Le cercle des corrélations :**
  Comme son nom l'indique, le cercle des corrélations facilite l'étude des liens existants entre les variables étudiées après projection dans le sous espace. Son étude permet aussi d'apprécier les liens existants entre variables initiales et les axes factoriels construits.

- **Carte factorielle des individus :**
  La carte factorielle permet d'obtenir des informations précieuses sur le nuage des individus (pixels) après projection. Cette carte permet de connaître les liens entre les individus. Ainsi, on pourra savoir quels individus sont liés entre eux (dans l'espace projeté), ou savoir quels points participent aux axes, ou encore savoir s'il y a des points aberrants.

Afin d'illustrer les tests réalisés dans cette section, nous présentons des images synthétisant les résultats de l'ACP. En ce qui concerne la notation utilisée, nous avons décidé de nommer les indices textures de A à E. Etant donnée que chaque indice est calculé pour 4 tailles de fenêtres différentes, on nommera A_i l'indice A calculé avec la i^{eme} taille de fenêtre (de la taille la plus faible à la plus grande).
Les indices textures sont les suivants :

- A : Indice relatif aux longueurs de plages (**Eq. 3.8**)
- B : Indice relatif aux transitions encre/papier (**Eq. 3.7**)
- C : Indice relatif à la forme de la rose (**Eq. 3.6**)
- D : Indice relatif à l'orientation principale (**Eq. 3.4**)
- E : Indice relatif à l'intensité de la rose pour l'orientation principale (**Eq. 3.5**)

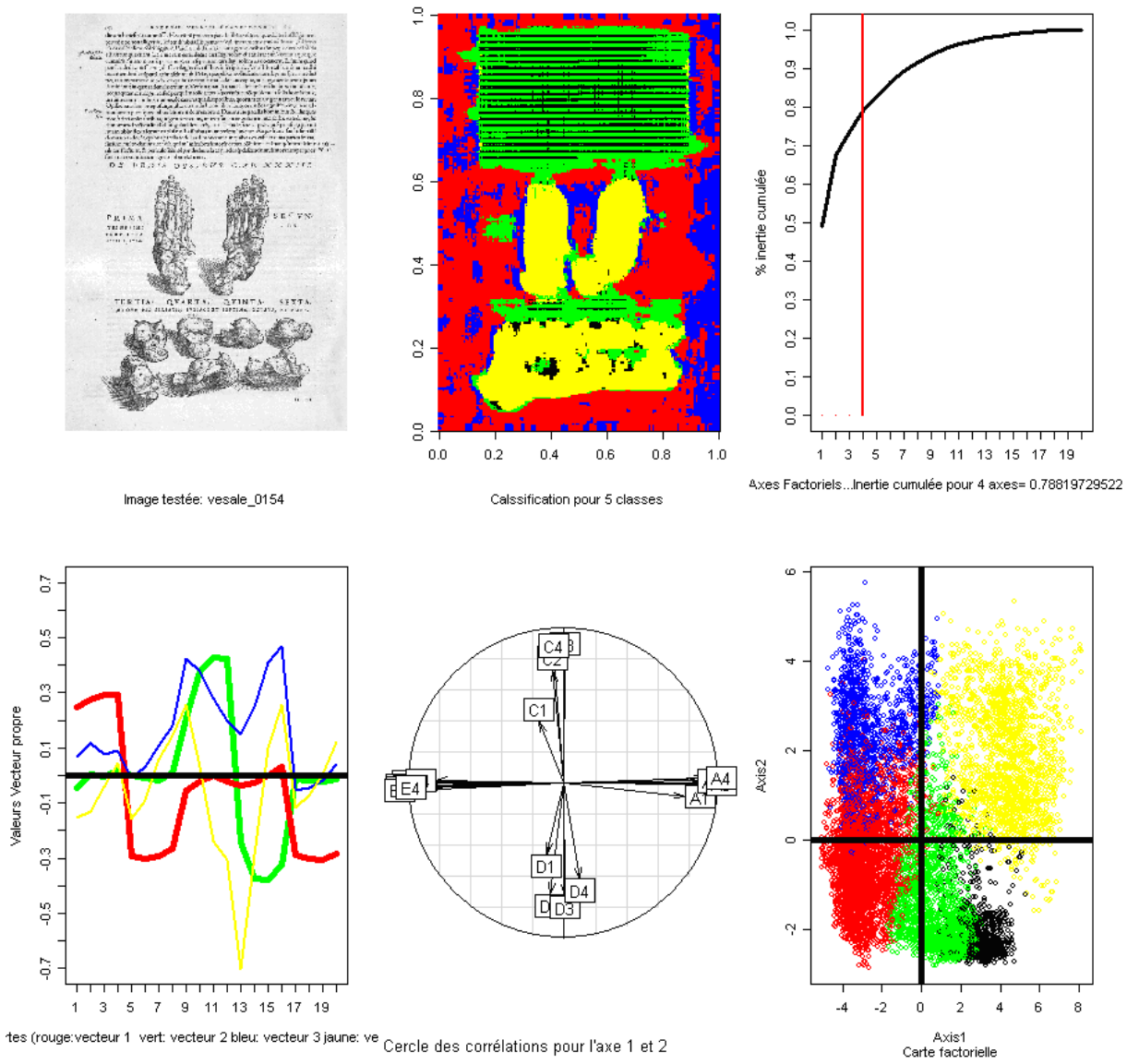


FIG. *3.27: Résumé des résultats d'un ACP sur une image de document*

Les informations que l'on retrouve sur ces images de résumé sont liées aux informations interprétables après un calcul d'ACP. La figure **Fig. 3.27** est un exemple de ces informations. En haut à gauche, on retrouve l'image testée. A sa droite est affichée le résultat d'une classification à k classes (k est indiqué sous l'image). Cette classification est fournie à titre indicatif car elle est effectuée sur une seule image. Il n'y a aucun lien d'une image à l'autre entre deux classifications. En haut à droite, on retrouve le pourcentage d'inertie cumulée pour les 20 axes (nos 20 variables). Une barre verticale est placée au quatrième axe pour visualiser le pourcentage d'inertie conservé lorsque l'on ne garde que les 4 premiers axes (ce qui est notre cas dans nos tests). Ce taux d'inertie est indiqué dans le sous-titre de l'image.
En bas à gauche, sont détaillés les coefficients des vecteurs propres des caractéristiques cor-

respondant aux quatre premiers axes. Une ligne noire horizontale indique l'axe 0. Il n'y a pas d'interprétation spéciale à faire sur ces valeurs. L'objectif de ce graphique est d'obtenir une information sur la stabilité de ces coefficients d'une image à l'autre. A sa droite, on retrouve le cercle des corrélations. On retrouve, sur chaque cercle, 20 étiquettes qui correspondent aux 20 variables. La dernière figure est la carte factorielle des individus (pixels classés). Les couleurs des points ont un sens. Elles correspondent aux couleurs utilisées lors de la classification. Dans l'exemple donné, les pixels verts sont des pixels de texte. Ces pixels sont majoritairement portés par l'axe 2. On voit aussi très nettement que les pixels de même couleur sont proches les uns des autres. Selon le test réalisé, on ne retrouvera pas forcément l'ensemble des imagettes relatives aux caractéristiques des données.

3.3.2.2 Etude des corrélations entre indices

La première interrogation à laquelle il nous semble important de répondre, est de savoir comment se comportent les indices. Plus précisément, nous souhaitons savoir dans quelle mesure le contenu d'une image influence les indices. Afin d'apporter un élément de réponse, nous avons décidé de créer 4 groupes d'images que nous classerons en fonction de leur contenu. Le premier groupe est composé d'images comportant en grosse majorité des illustrations. Le deuxième regroupe les images ne contenant que du texte. Le troisième est composé de pages comportant textes et dessins dans les mêmes proportions. Enfin, le dernier groupe est celui dans lequel on trouvera des images de documents contemporains. Nous avons essayé d'être le plus varié possible dans le choix des images. Certaines d'entre elles ont été crées manuellement, afin de mettre en avant certaines caractéristiques. Afin que d'autres paramètres ne viennent pas influencer les analyses de cette hypothèse toutes les images ont été numérisées à la même résolution et elles ont également la même taille (**Fig. 3.28**).

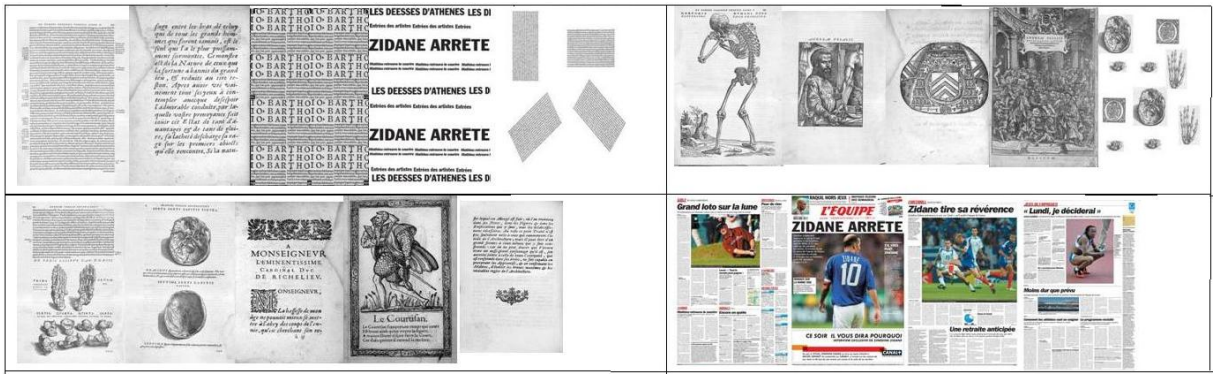


FIG. *3.28: Extrait du corpus de test*

La figure **Fig. 3.29** donne quelques exemples d'analyse des corrélations sur des images issues de trois ouvrages différents.

Il n'existe pas une méthode absolue permettant de conclure sur la nature des données. Les conclusions que nous avons tirées des différentes analyses 5ACP) réalisées sur ce corpus sont les suivantes :

1. L'inertie portée par les 4 premiers axes est toujours très bonne. Il est possible dans tous les cas de réduire nos données d'une dimension 20 à une dimension 4 tout en gardant environ

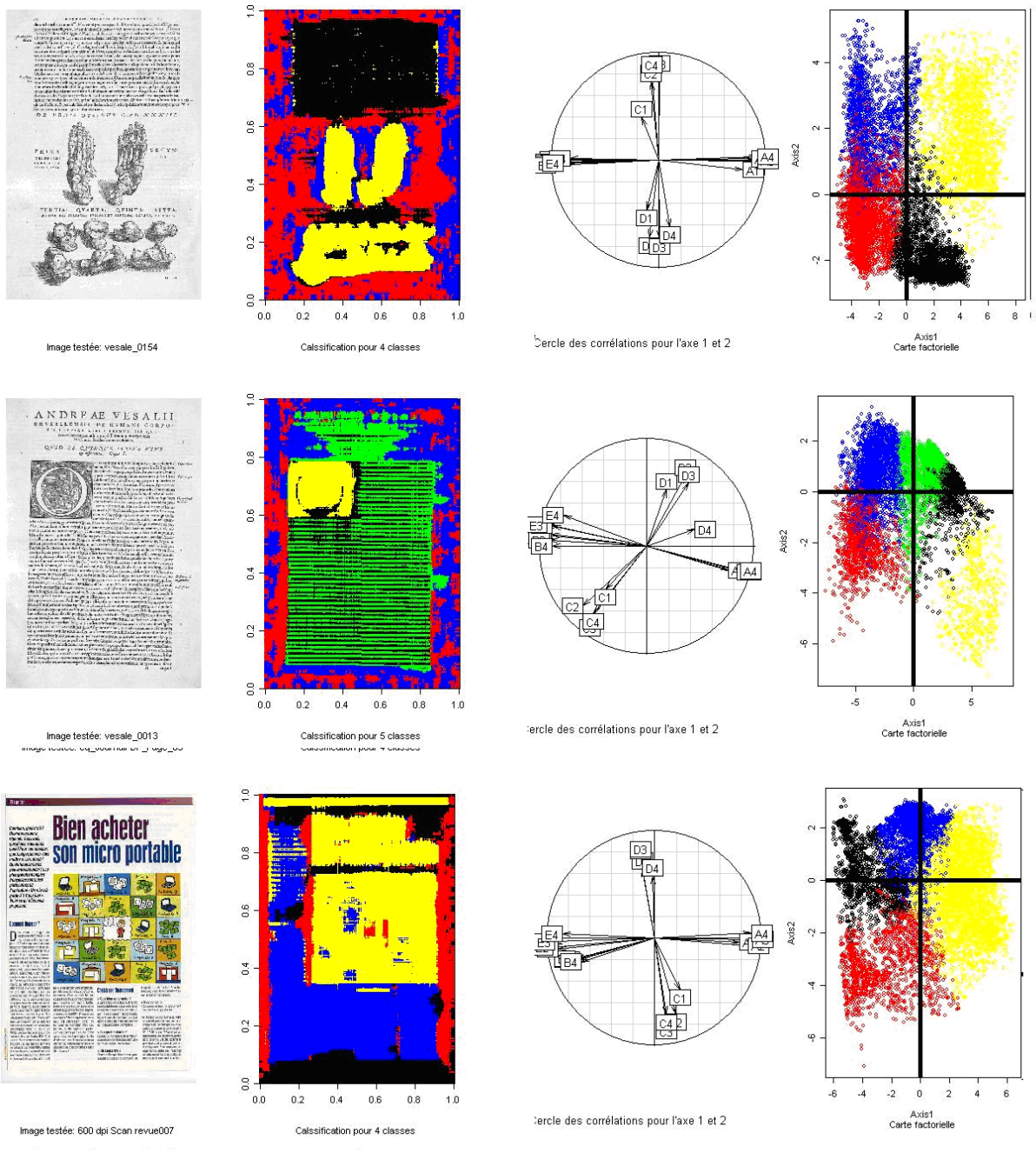


FIG. 3.29: *Comportement des données*

78% de l'information (cf. **Table 3.2** qui a été calculé sur 116 images provenant de plusieurs ouvrages).

% d'inertie cumulée (I) pour 4 axes	$I < 70\%$	$70\% < I < 75\%$	$75\% < I < 80\%$	$80\% < I$
Nombre d'images	1	10	74	31

TAB. 3.2: *Evaluation de l'inertie des axes factoriels sur 116 images*

2. Les indices B (transition encre/papier) et E (intensité de la corrélation pour l'orientation principale) sont fortement corrélées entre eux et anticorrélées avec l'indice A (longueur de plages). Ceci s'explique, par le fait que le texte est toujours horizontal et qu'il est très fortement présent dans nos images. L'indice B est sensible aux fortes transitions noir/blanc correspondant aux lignes de texte. Il se trouve que l'indice E est lui aussi sensible aux orientations privilégiées (ici horizontales). L'indice A leur est anticorrélé. La plus probable hypothèse étant que cet indice est sensible aux longues plages blanches (contrairement aux deux autres qui le seront aux longues plages noires).
3. Nous n'avons pas réussi à faire apparaître des relations stables entre variables (coefficients de la combinaison linéaire) permettant de passer d'une dimension 20 à 4. Il n'existe donc pas de relation stable entre variables initiales et les variables synthétiques.
4. Regarder en même temps le cercle des corrélations et la carte factorielle permet de faire certains rapprochements entre les variables et les individus. Etant donné le nombre de pixels que l'on projette sur les deux premiers plans factoriels, il est très difficile d'observer des corrélations précises entre les deux premiers axes et les individus dans l'espace projeté. Cependant on remarque certaines tendances qui se répètent d'une image à l'autre. Les pixels du fond semblent être principalement corrélés avec la partie gauche de l'axe 1. Quand on regarde les cercles des corrélations, on aperçoit que les indices E et B sont corrélés avec cet axe. De même, les pixels de texte ont tendance à être portés par l'axe 2 et l'on observe qu'ils sont fortement liés à l'indice D, soit l'orientation principale (même remarque pour les pixels de dessin avec l'indice A).
5. Les cartes factorielles montrent également que les différentes classes restent bien séparées même après projection. La figure **Fig. 3.30** montre ce qu'aurait pu donner une carte factorielle où la projection dénature l'information de départ (les individus sont k-séparables en dimension 20 mais ne le sont plus en dimension 2).

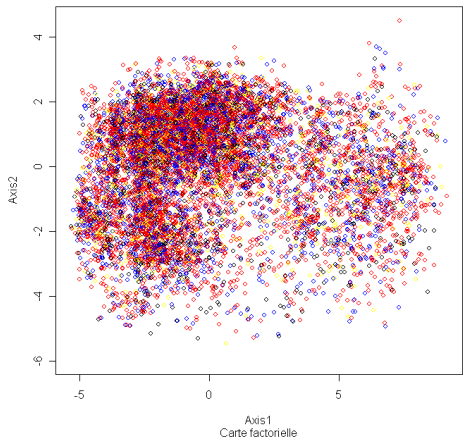


FIG. *3.30: Exemple d'une mauvaise projection d'un espace n à p avec n>p.*

Nous reviendrons plus en détail sur l'apport de la multi-résolution, mais on peut d'ores et déjà remarquer certaines caractéristiques. Sur les exemples donnés précédemment, on remarque une forte corrélation des indices d'une résolution à l'autre. Cependant, la figure **Fig. 3.31** montre que selon le contenu de l'image, les données d'un même indice ne sont pas toujours corrélées

entre elles. Dès lors que l'on amplifie artificiellement certaines caractéristiques des images (multi-orientation, plusieurs tailles de caractères...), on remarque que les indices calculés à différentes résolutions sont moins nettement corrélés. On remarque par ailleurs que les indices E,C,D sont décorrélés. Ces trois indices sont ceux relatifs à la rose des directions. On voit par exemple toute l'importance d'un calcul multirésolution sur la première image. En effet, les indices relatifs à la résolution 4 apportent peu d'informations sur les deux premiers plans. Il était donc nécessaire de calculer cet indice à plusieurs résolutions pour obtenir une information pertinente.

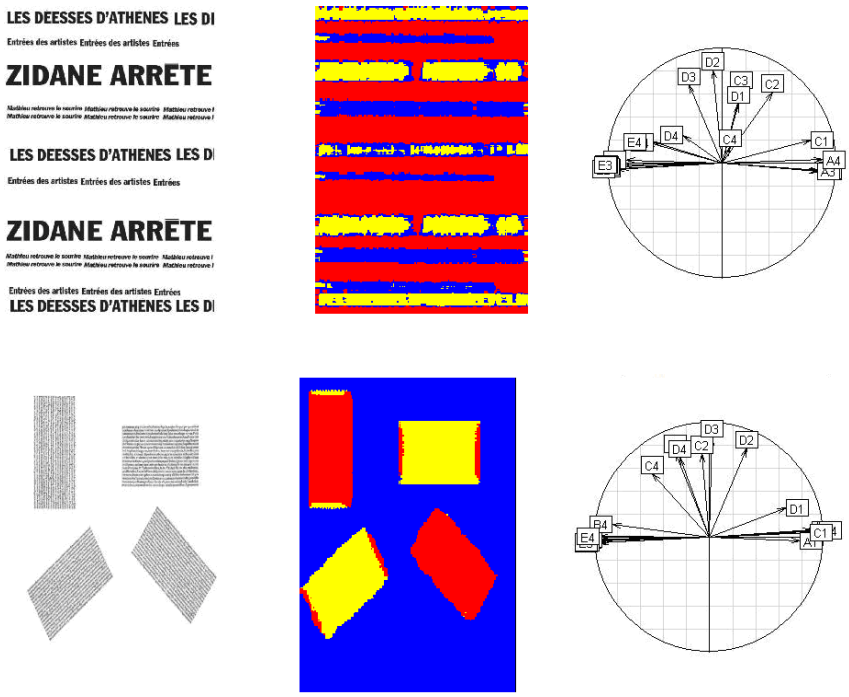
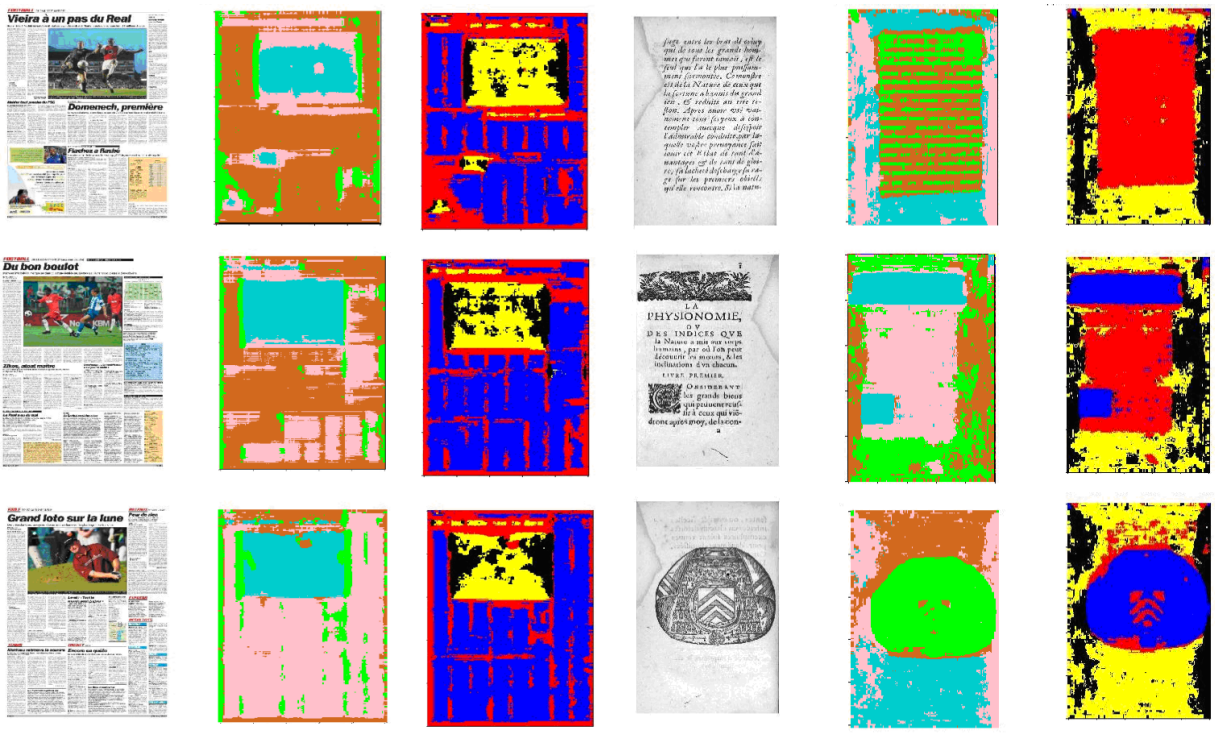


FIG. *3.31: Décorrélation des indices multirésolution.*

En conclusion, on peut donc dire que les indices proposés comportent de la redondance puisqu'il semble qu'il est possible de réduire la taille de l'espace de représentation des pixels. Il semble également envisageable de ne plus calculer l'indice B ou E. Malheureusement, il ne nous a pas été possible de construire une application linéaire permettant de passer de la dimension 20 à la dimension 4 pour tous les types de documents. Il faudrait construire plusieurs application linéaires (pour chaque type d'images) et être capable de connaître le type d'une image au préalable, ce qui n'est pas facilement envisageable.

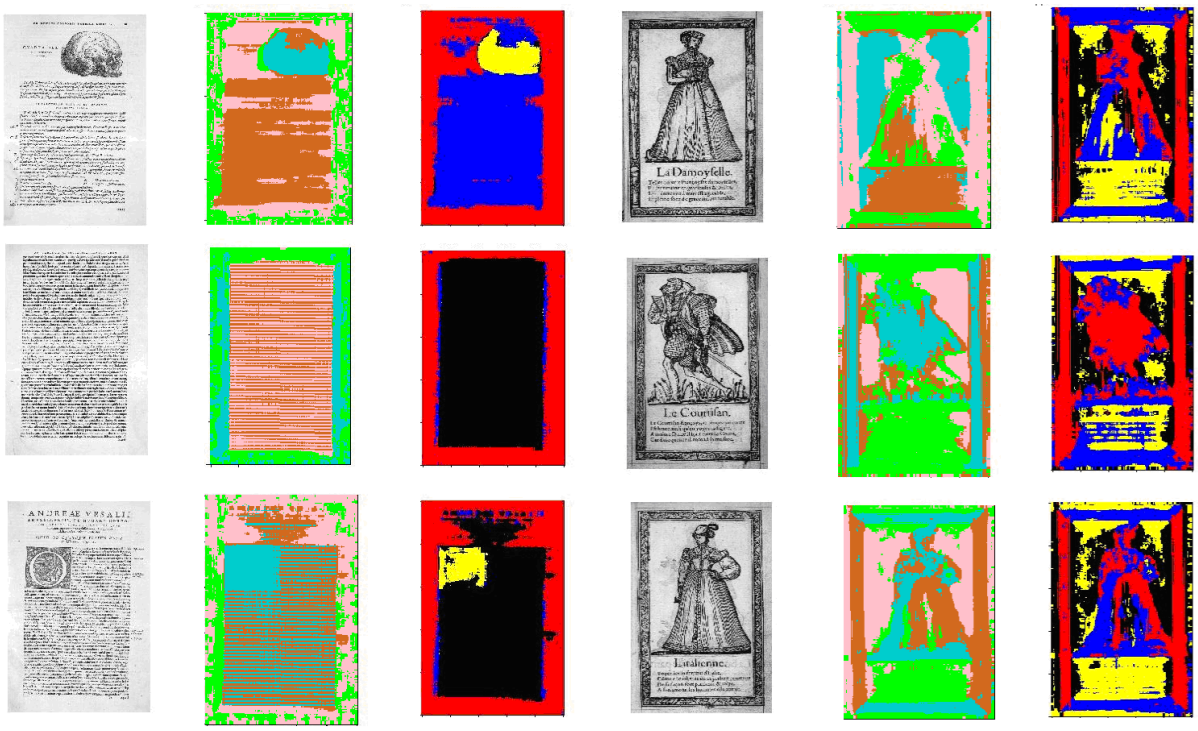
3.3.2.3 Granularité de l'analyse : de la page à l'ouvrage

Dans le cadre de notre travail, nous avons été amené à traiter un ensemble de pages plutôt qu'un traitement page par page. En effet, dans notre objectif de classification via la caractérisation du contenu, il est indispensable de traiter un ouvrage dans son ensemble. On ne va donc pas simplement chercher à comparer entre eux les pixels d'une page, mais plutôt comparer tous les pixels d'un ouvrage entre eux ; un peu comme si l'on analysait une seule image qui serait composée de toutes les pages " collées " l'une derrière l'autre. Les tests effectués au début de ce chapitre ont montré la pertinence qu'il y a à classifier les pixels d'un ouvrage complet, plutôt que de classer les pixels page par page. Un exemple tout simple de l'intérêt de cette approche, est que l'on ne connaît pas à l'avance le contenu de la page. Est elle composée de texte et de dessin ? Dans qu'elles proportions ? On ne peut donc pas envisager une approche où l'on demanderait, pour chaque page et quelque soit son contenu, une classification texte/fond/dessin. Dans les figures **Fig. 3.32** et **Fig. 3.33** on voit très nettement l'intérêt : non seulement la classification donne de " meilleurs " résultats visuels, mais en plus on est capable de dire qu'un pixel rouge de la page 1 est identique à un pixel rouge de la page 2 (ce qui n'est pas le cas d'un classification sur une page où d'un exemple à l'autre la couleur du marquage n'a **aucun lien**).



(a) Ouvrage contemporains *(b) Ouvrage ancien 1*

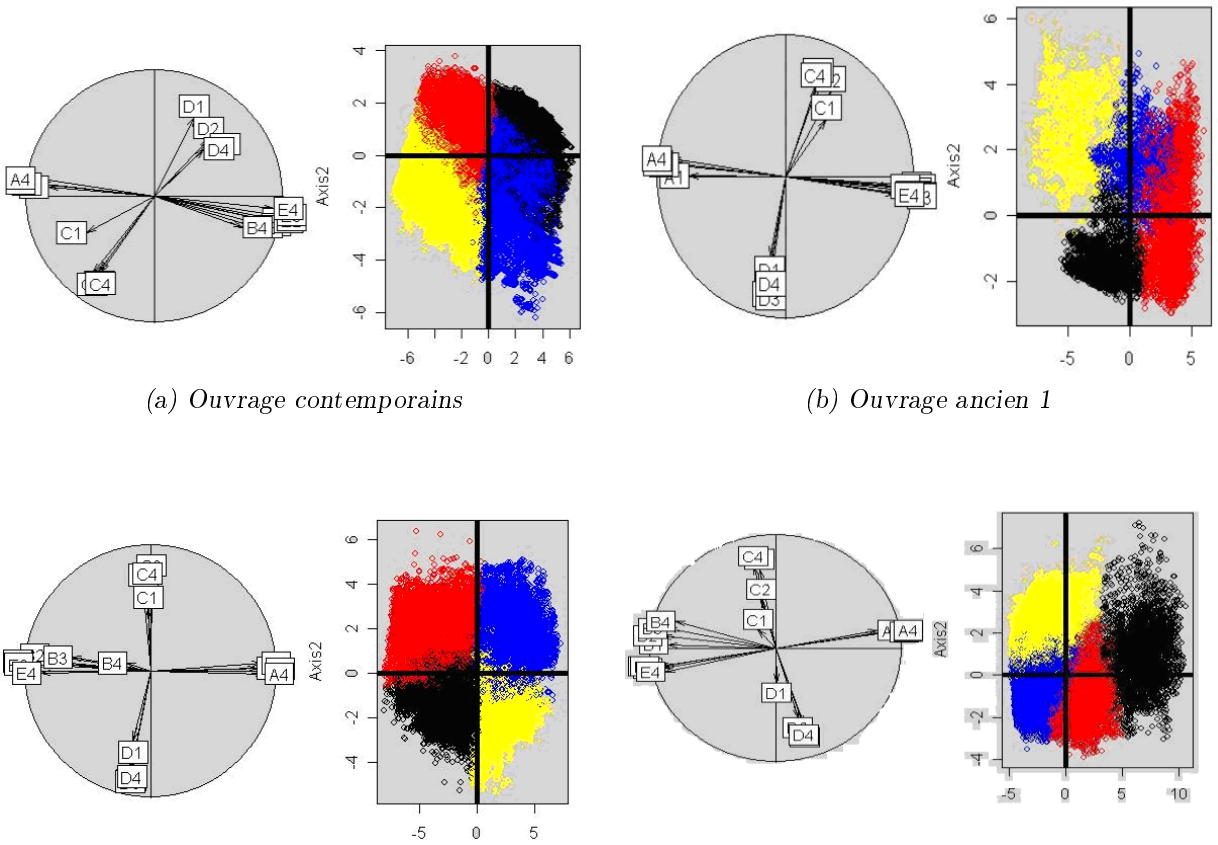
FIG. *3.32: Comparaison entre une classification page par page et une classification d'un ouvrage complet*



(a) Ouvrage ancien 2 (b) Ouvrage ancien 3

FIG. 3.33: Comparaison entre une classification page par page et une classification d'un ouvrage complet

Regardons maintenant à quoi ressemblent les données lorsqu'on les concatènent. On rappelle que par ouvrage on a un seul fichier de description et donc une seule ACP calculée **Fig. 3.34**.



(a) Ouvrage contemporains

(b) Ouvrage ancien 1

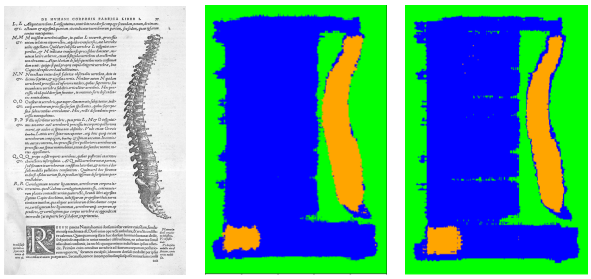
(c) Ouvrage ancien 2

(d) Ouvrage ancien 3

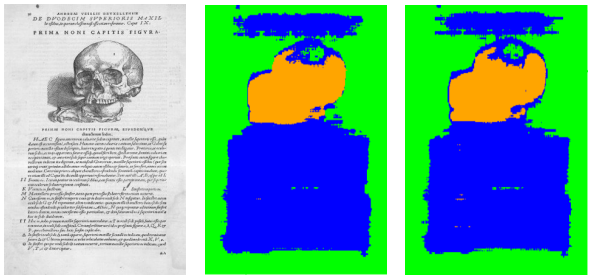
FIG. *3.34: Comportement des caractéristiques lorsqu'on classifie un ouvrage complet*

L'analyse a été réalisée sur 4 ouvrages différents. Il se trouve que l'on retrouve globalement les mêmes relations Axes/variables/pixels que pour une analyse page par page. On retrouve les indices B,E,A portés par l'axe 1 et les deux autres portés par l'axe 2. L'apport d'un calcul multirésolution se voit, entre autre, sur le cercle de l'ouvrage 3.

Une analyse page par page a montré que d'une image à l'autre, les relations entre variables, et leur apport aux deux premiers axes varie selon le contenu. Travailler sur un ensemble de pages a tendance à générer des relations plus "stables" entres les variables et les individus. Prendre un nombre important de pixels pour l'analyse, a tendance à "moyenner" l'information de certaines pages aux caractéristiques bien spécifiques.
Nos tests ont montré qu'il était possible de produire une combinaison linéaire permettant de passer de 20 variables à 4 variables (combinaison linéaire des 20) dans le cadre d'un travail par ouvrage. Cette combinaison linéaire n'est valable que pour les pages issues de l'ouvrage sur lequel a été calculé l'ACP. La figure **Fig. 3.35** montre que cette réduction de 20 à 4 variables ne détériore que très peu les résultats visuels. Cette réduction a avant tout pour intérêt de réduire la quantité de stockage des données.



(a) Image d'origine (b) Classification sans réduction (c) Classification après réduction



(d) Image d'origine (e) Classification sans réduction (f) Classification après réduction

FIG. 3.35: *Illustration de la capacité à réduire les données après la réalisation d'une ACP sur un ouvrage complet*

3.3.2.4 Influence de la taille et de la résolution des images

Les bases que nous avons à traiter sont composées d'images de tailles différentes et parfois numérisées à des résolutions différentes. Pour évaluer l'impact de la taille des images sur notre méthode, nous proposons d'analyser une image à 5 tailles différentes (150%, 100%, 80%, 70%, 60%) de la taille de l'image d'origine (600 X 995). Nous avons choisi l'image de la figure **Fig. 3.36** car elle possède à la fois du texte (plusieurs paragraphes) et deux illustrations différentes. Etant donné que nous ne changeons jamais la taille des fenêtres d'analyse, il y a une taille minimum pour les images. Dans le cas de l'image testée, en dessous de 60% de la taille d'origine, l'image testée est plus petite que la plus grande des fenêtres et donc les indices ne sont pas calculables.

Sur la figure **Fig. 3.36**, on observe que les résultats sont très similaires alors que la taille de l'image varie de 370X548 à 1334X900.

Pour être plus précis, la classification donne des résultats quasi identiques. Le taux d'inertie est toujours très bon et les cercles des corrélations sont identiques (l'inversion des deux indices D et C dans le cercle de l'image 1 n'a pas d'importance puisque la carte factorielle est elle aussi inversée.) Nous avons effectué les mêmes tests avec des images numérisées à différentes résolutions (1200, 300 et 150 dpi) et les conclusions restent les mêmes.

Ces tests permettent de penser que l'utilisation de la multirésolution est justifiée. Un exemple

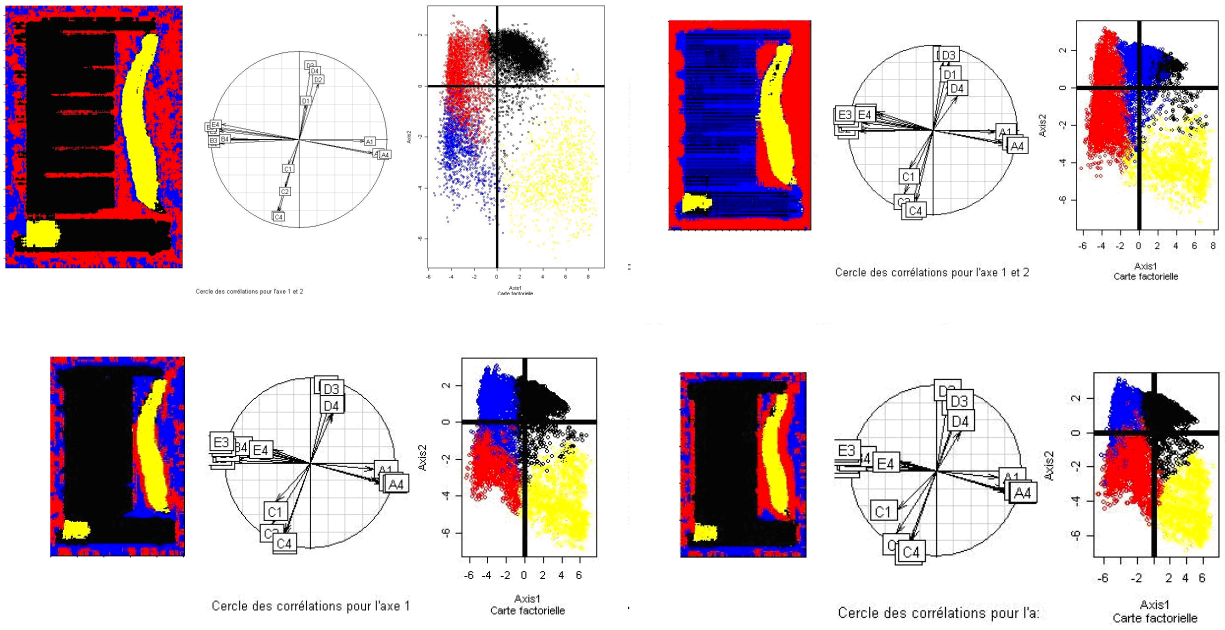
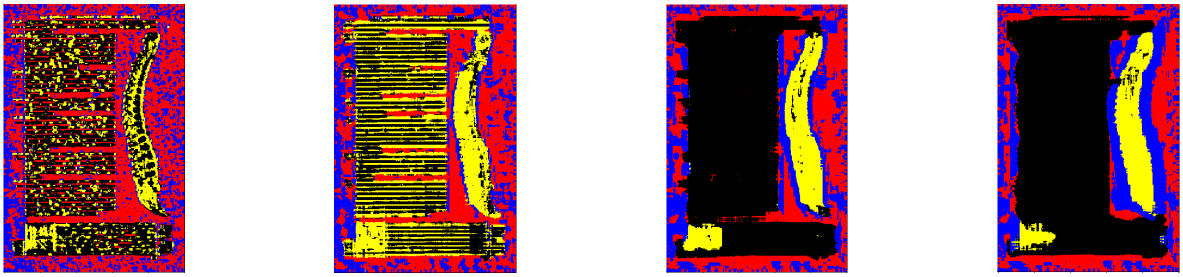


FIG. *3.36: Comportement des données selon la taille des images*

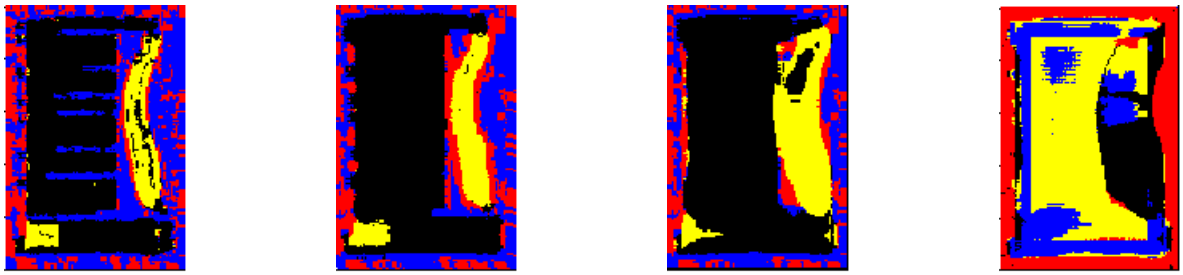
très simple montre les résultats que nous aurions si jamais nous n'avions utilisé qu'une résolution. La figure **Fig. 3.37** prouve que pour une image de grande taille (1334X900)la classification n'est pertinente que dans dans les deux derniers cas.



(a) Plus petite résolution *(b) Deuxième résolution* *(c) Troisième résolution* *(d) Quatrième résolution*

FIG. *3.37: Comportement des données selon la taille des images (cas d'une grande image)*

A l'inverse, la figure **Fig. 3.38** montre que pour une image de petite taille (370X548) la classification n'est pertinente que dans dans les deux premiers cas.



(a) Plus petite résolution (b) Deuxième résolution (c) Troisième résolution (d) Quatrième résolution

FIG. 3.38: Comportement des données selon la taille des images (cas d'une petite image)

3.3.2.5 Etude de l'échantillonage des données

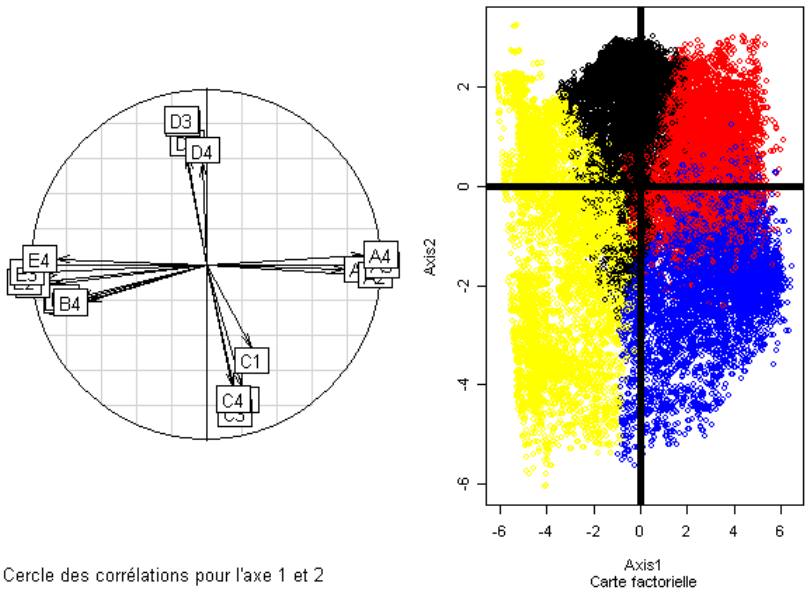
Nos images sont de grosse taille, les fichiers de descripteurs (après calcul des indices) sont volumineux. De ce fait toute opération sur ces fichiers de données est coûteuse en temps. La forte corrélation de nos données (cf. hypothèse précédente) et la forte corrélation spatiale des pixels, nous permet de supposer qu'un échantillonnage est réalisable sans détériorer la qualité des analyses. Nous vérifions cette hypothèse en comparant des résultats d'une ACP sur des données échantillonnées ou non. Les pixels sont choisis aléatoirement.

Les tests de la figure **Fig. 3.39** sont réalisés sur un ouvrage complet. L'inertie des 4 premiers axes ne varie pas et reste égale à 78%. La question est donc de savoir si prendre un échantillon (aléatoire) de pixels va dénaturer les relations entre variables. Un élément de réponse est visible dans les cercles des corrélations selon le nombre de points pris pour calculer l'ACP.

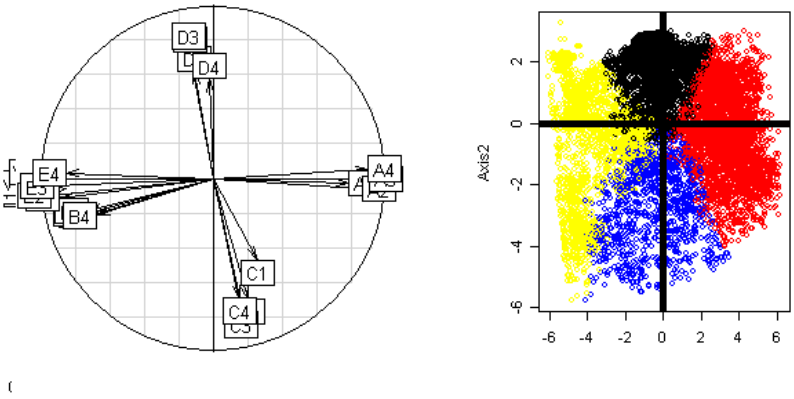
Sur les essais réalisés sur des images de 600X889,il faut descendre en dessous de 0.1% des points pour avoir des cercles des corrélations qui diffèrent. Ces tests montrent qu'un échantillonnage des données ont très peu d'incidence sur le calcul des ACP. De ce fait, on peut supposer que tout traitement d'analyse (classification hiérarchique, acp....), qui est difficilement réalisable sur de telles tailles de données, sera tout à fait réalisable et exploitable sur un échantillon des données.

3.3.3 Discussions sur la pertinence des indices textures

Cette partie a permis d'une part de valider la pertinence des attributs textures présentés au début de ce chapitre. Un algorithme de classification par centres mobiles, permet de visualiser le pouvoir discriminent des attributs. Selon les tests, nous avons vu qu'il est possible de séparer le texte des dessins, des gros caractères des petits, du texte multi-orienté... D'autre part, l'analyse des données a permis de faire ressortir certaines caractéristiques des indices extraits, ainsi que l'intérêt d'un calcul multirésolution d'autant plus que les images du corpus ne sont pas toutes de même taille. De cette étude ressort principalement le fait qu'il est possible de réduire la taille des données, dès lors que l'on effectue une seule ACP sur un ouvrage complet. Il ressort également qu'il est possible de procéder à des traitements (ACP, classification, ...) sur un échantillon des données plutôt que sur la totalité.



(a) 100% des pixels de l'ouvrage



(b) 10% des pixels de l'ouvrage

FIG. *3.39: Comportement des données selon la taille de l'échantillon*

3.4 Conclusion générale

Dans ce chapitre, nous avons détaillé notre proposition d'indices textures dédiés à l'analyse d'images de documents. La spécificité du corpus, nous a poussé à imaginer ou adapter certains algorithmes de caractérisation d'images. La création de ces indices et leur calcul à différentes résolutions représentent une première partie de notre contribution à l'analyse d'images de documents anciens. Ainsi, nous avons mis en évidence la pertinence de l'usage de la rose des directions et de l'étude des transitions encre/fond, mais également l'importance d'un calcul à différentes résolutions.

Le chapitre 2 a mis en évidence la difficulté d'utilisation de certains outils dès lors que nous les appliquons sur des documents anciens. Notre proposition répond à ces lacunes à l'aide d'une méthode de caractérisation des images de documents n'incluant pas de seuils (tailles des caractères, mise en page, niveaux de gris...) et permettant également de caractériser des ouvrages de contenus très variables. Cette proposition répond en particulier au besoin de plus en plus important de caractériser des ouvrages de contenus variables">pour ce faire sans structures connues a priori. L'étape d'étiquetage des pixels après classification, montre qu'il est possible de caractériser les contenus sans pour autant segmenter ou identifier la structure d'un document.
Dans le chapitre suivant, nous détaillons au travers de plusieurs expérimentations, comment il est possible d'utiliser, à l'aide de différents procédés, ces indices textures afin d'indexer le contenu d'images très variées.

Chapitre 4

Illustration de la pertinence des indices textures

4.1 Introduction

4.1.1 Contexte de l'étude de la pertinence des indices de texture

En terme d'usages, la liste des outils et applications souhaités par les utilisateurs des bibliothèques numériques est aussi diverse que variée. A ce propos, le premier chapitre a permis de faire un bilan relatif à ces attentes. La suite de ce mémoire a mis en évidence l'importance, mais également la complexité de la tâche que représente l'indexation d'images autrement que par l'utilisation de mots-clefs. Appliqué aux pages de documents anciens, cet objectif d'indexation nécessite ainsi de pouvoir caractériser finement le contenu d'un corpus complexe à traiter. Cette complexité de traitement est principalement liée à l'hétérogénéité des images qui le composent. Face à ce constat, nous avons proposé dans le troisième chapitre, différents indices textures permettant de caractériser le contenu des images de documents anciens. Ces indices se basent sur une analyse des orientations et des fréquences des motifs qui y sont présents.

Finalement, ces trois chapitres ouvrent (peut être) sur plus de questions qu'ils n'amènent de réponses. Ainsi, il reste toujours à savoir si sans segmenter ou retrouver la structure d'un document, il est possible de mettre en place des outils d'indexation, et finalement de juger si les indices que nous proposons apportent une solution au problème de caractérisation de contenus d'images de documents anciens.

Comme énoncée dans l'introduction de ce mémoire, cette thèse n'avait pas pour objectif de réaliser une application précise dans le domaine de l'indexation par le contenu. Cette liberté nous a ainsi permis d'explorer de nombreuses pistes sur l'utilisation que nous pouvions faire des indices présentés dans le chapitre précédent. Ainsi, nous avons testé plusieurs idées, permettant d'apporter des éléments de réponses aux interrogations évoquées précédemment.

Si l'on se réfère au schéma introduisant le chapitre 3, ces expérimentations sont présentées comme des modules indépendants venant utiliser les indices extraits dans la première phase de notre méthode. La philosophie de notre proposition, tient essentiellement au fait que nous pensons qu'en utilisant de la sorte ces indices, il est possible de réaliser diverses applications sans pour autant devoir redévelopper le processus de caractérisation à chaque nouvel objectif. Par ailleurs, elle est également marquée par la très faible charge de paramétrage manuel nécessaire à

l'exploitation de cette caractérisation. En pratique, seules les tailles de fenêtres de balayage des images lors de la construction des vecteurs de caractéristiques sont un facteur à fixer a priori.

4.1.2 Principe de l'exploitation des indices de texture

Dans le domaine de la recherche d'information dans des bases d'images, une des solutions les plus courantes est de rechercher des images en donnant comme requête une « image exemple » de ce qu'on recherche. Les systèmes dédiés à cette recherche proposent alors en réponse un ensemble d'images similaires à l'exemple donné. Les usages qui peuvent en être faits par la suite peuvent être de nature très diverse tout comme les outils ergonomiques d'interface permettant de les réaliser. Ces derniers aspects ne seront pas abordées dans ce chapitre.

Dans ce contexte, le problème est de savoir ce que l'on entend par « les images les plus similaires à un exemple donné » puisque la seule définition de la similarité est très subjective et que finalement seul l'utilisateur sait réellement ce qu'il recherche (un type de pages, une page précise de texte, un objet graphique..). Naturellement, le contenu d'une image est très complexe et la représentation de cette même image en réduit énormément le contenu : il faut donc trouver une bonne adéquation entre cette représentation et la similarité qui permettra de réaliser les comparaisons entre l'image requête et les images de la base.

Ce problème est d'autant plus complexe que les contenus des images sont hétérogènes et variables d'une image à l'autre, en particulier sur les images fortement texturées et les images de traits qui présentent de grandes variabilités internes mais également entre elles. Il faut donc définir des métriques à la fois riches et précises, ce qui peut être pénalisant en terme de complexité et de temps de calcul dans un système de recherche.

Notre travail s'inscrit donc dans ce contexte où il faut parvenir à trouver le meilleur compromis possible entre la nécessaire simplification des contenus des images (individuellement très volumineuses) et le besoin de justesse de réponse du système de recherche.

Nous allons montrer que les caractéristiques de bas niveau que nous avons définies dans le chapitre 3 peuvent constituer de façon très pertinente la base d'un système de recherche par le contenu en remplaçant le contenu réel des images de manière interne. De façon très générale à tout système, les caractéristiques ainsi que les métriques permettant de les manipuler (les mesures de similarité) peuvent à elles seules constituer une limite à la performance d'un système et cela, quelles que soient les techniques d'apprentissage, la taille de la base d'image de tests et celle de la base d'apprentissage utilisées. Dans ce chapitre, nous ne proposons pas de système de recherche impliquant l'ensemble de ces composantes mais nous portons toute notre attention sur les deux dimensions fondamentales de tout système : les caractéristiques de bas niveau et les métriques de similarité permettant de réaliser les comparaison entre images. Les approches "système" ne seront évoquées que comme des perspectives à plus long terme de ces travaux.

En conséquence, les résultats proposés dans chacune des expérimentations ci-après sont donc à analyser en relation avec ces hypothèses de départ. En particulier, les taux de précisions et les performances internes de l'outil d'analyse de nos indices de texture ne sont pas à comparer avec les résultats des systèmes complets de recherche d'images par le contenu (Les systèmes dits CBIR ou RIC) que l'on peut trouver en abondance dans le domaine.

La première partie de ce chapitre présente différentes mesures de similarités pouvant être utilisées lors de l'exploitation de nos indices textures. Les mesures que nous proposons dans un premier temps permettent d'utiliser soit directement, soit après classification des pixels, les

indices textures associés à chaque pixel pour comparer des pages de documents ou pour comparer les éléments de contenus (comme le montrent les expériences réalisées).
Ce chapitre se conclut sur des propositions d'exploitation potentielles des indices textures. Il peut s'agir alors de venir compléter la liste des descripteurs d'images utilisées dans un système CBIR ; il est alors nécessaire de synthétiser les données pour réduire la quantité d'informations à stocker dans les métadonnées de ces descriptions. Il peut aussi être question de regrouper les pages similaires d'ouvrages ou même venir en aide à un processus de segmentation et d'extraction de la structure d'images de document.

4.2 Indices de similarité proposées

4.2.1 introduction

Dans les systèmes d'informations permettant un accès au contenu des bases d'images, les objets sont la plupart du temps représentés par des vecteurs de grande dimension. Le mode d'exploitation de ces systèmes est basé sur l'image requête et une métrique de similarité exprimant la distance entre la requête et les images de la base. Des systèmes d'indexation de renom tels que le QBIC d'IBM ([FSN+95]) ou le projet PhotoBook du MIT ([PPS96]) peuvent être donnés en exemple. D'autres systèmes plus récents peuvent également être cités comme le système KIWI ([Lou00]), le système NETRA ([MM99]) ou encore les travaux de [Haf05]. Si chacun de ces systèmes possède ses spécificités (mode de requête, distances de similarités utilisées, place de l'utilisateur dans le système, mode de recherche...) , on retrouve souvent les mêmes caractéristiques sur lesquelles sont basées les mesures de similarité : couleur, formes et textures.

La mise en place d'un système d'information permettant un accès au contenu de bases d'images, nécessite généralement une analyse et un traitement en temps réel des informations recherchées. Cette contrainte implique donc de prendre en compte les critères que sont la dimension de la taille de la base et de l'espace des caractéristiques.
Pour ce qui est du nombre d'images conservées dans de tels systèmes, si ce nombre est petit, un processus de recherche linéaire est généralement suffisant pour produire des performances acceptables lors d'une recherche. Cependant, lorsqu'il s'agit de traiter des volumes de données de dimensions très supérieures (allant de quelques milliers à plusieurs millions d'images), ce qui est le cas dans les bibliothèques numériques, une telle simplicité de fonctionnement n'est plus suffisamment performante et ne permet pas de répondre à des besoins de réaction en temps réels. Dans ce contexte, on peut rappeler que certaines méthodes de recherche d'images dans de grandes bases utilisent des techniques de hachages et de structures d'arbres ([MM85]).

Dans une application de recherche d'images par le contenu, il est généralement admis de procéder en deux étapes : pour chaque image de la base, un vecteur d'attributs (ou un ensemble de vecteurs) caractéristiques de certaines propriétés de l'image est calculé et stocké dans une base d'attributs. Puis, considérant une image requête possédant ses caractéristiques propres, on extrait de la base les images les "plus proches". Les attributs et la mesure de similarité utilisés pour comparer les vecteurs d'attributs de deux images doivent être suffisamment précis pour parvenir à faire ressortir les images réellement similaires de la base et parallèlement d'écarter celles qui sont fortement dissimilaires. Il s'agit d'un objectif double qu'une seule métrique ne peut pas toujours satisfaire.

Nous avons tout d'abord tenté à travers les métriques de similarité que nous présentons dans ce chapitre, de répondre au mieux à ces deux exigences.

On peut citer deux types de mesures de ressemblance couramment employées dans la littérature : celles basées sur une similarité entre attributs et celles basées sur la mise en correspondance de graphes.
Pour ce qui est des mesures basées sur une similarité entre attributs, on discerne deux approches :

- Celles qui comparent des vecteurs de caractéristiques. Un vecteur décrit de manière globale les caractéristiques d'une image et la similarité entre deux images est induite par la distance entre deux vecteurs. Les distances euclidiennes ou distances de Minkowski sont celles que l'on retrouve le plus souvent dans la littérature.
- Celles qui comparent les distributions statistiques des caractéristiques des images. Si chaque pixel est décrit par un vecteur de n caractéristiques, alors il est possible de construire n histogrammes bi-dimensionnels correspondant à leur distribution dans l'image. La similarité entre images est induite par une distance entre histogrammes. Les plus connues sont les distances de Bhattacharyya, la divergence Kullback-Leiber... Les auteurs de [PRTB99] proposent un état de l'art de ces mesures et évaluent leur performances.

Les mesures de similarité basées sur le calcul et l'appariement de graphes permettent de réaliser une mesure de similarité d'ordre structurel. Les images sont segmentées en régions homogènes, ce qui permet de construire un graphe où les noeuds correspondent aux régions et les arcs à des informations relatives à ces régions (distance entre régions, position relative...). Là encore plusieurs méthodes sont proposées dans la littérature :

- Il est possible d'obtenir un (ou plusieurs) indices à partir d'une analyse du graphe. Par exemple, dans les travaux de [PVU+06], les auteurs segmentent en régions l'image analysée, ce qui permet de construire un graphe dont ils extraient plusieurs caractéristiques (chemin minimum reliant tous les sommets, positions des noeuds les uns par rapport aux autres...). Cette caractérisation est opérée pour chaque image, ce qui permet ensuite de les comparer entre elles.
- L'appariement de graphes est également très utilisé dans la littérature. Cela consiste à associer les régions de l'image testée, aux régions des images de la base. Une fois cette appariement réalisé, cela permet de calculer une mesure de similarité globale entre deux images. On trouvera dans [Pet02, GJ96, SK05] plusieurs exemples et comparaisons de méthodes utilisant les graphes. Parmi celles-ci on peut citer par exemple les appariement flexibles (Dynamic Time Warping), la mise en correspondance de graphes attribués (ARG)...

Afin de montrer la pertinence des indices extraits, nous avons imaginé deux expérimentations. Il ne s'agit pas ici de mettre en place une application répondant à un usage mais de vérifier d'une part que les indices sont exploitables et d'autre part de souligner les limites d'une telle approche. Dans notre travail, les deux applications envisagées sont les suivantes :

1. La première expérimentation consiste en une comparaison de pages d'images de documents anciens. L'objectif est de permettre le calcul d'une signature de l'ensemble des éléments contenus dans un document et de la comparer avec celle d'une autre page. Cette expérimentation permettra d'étudier s'il est possible, sans segmenter et sans identifier la structure d'une page, de caractériser l'ensemble de son contenu à partir d'informations textures. Les métriques globales (comparaisons de vecteurs) sont envisageables lorsque l'index de l'image est défini comme un ensemble unique d'attributs ou un vecteur de caractéristiques calculées sur l'image toute entière. Dans le cas des images très hétérogènes contenant des parties distinctes de textures différentes (telles que cela se trouve dans les images de documents hétérogènes de nos collections et plus fortement encore dans les images de textures naturelles), il n'est pas pertinent de ne disposer que d'un index global. Aussi, pour caractériser nos documents, nous avons choisi de nous inspirer des méthodes analysant la répartition

des différentes régions composant l'image. Nous proposons de fournir une description synthétique de la page à l'aide d'une classification des pixels. L'approche qui nous a semblé naturelle, se base sur nos travaux de caractérisation présentés dans le chapitre précédent et consiste en la définition d'une métrique de similarité à partir du découpage des images en partitions, et procédant ensuite en une série de comparaisons de partitions d'une image à l'autre.

2. La deuxième expérimentation a pour but d'évaluer la pertinence des indices calculés. Elle s'inspire des applications de type *image retrieval* et a donc pour objectif de retrouver des textures similaires à une texture donnée en exemple. Les images testées appartiennent à une base constituée des dessins de traits. Chaque image possède une seule texture caractéristique (une lettrine, un crane, ...). Nous avons décidé de ne pas comparer deux images sur la base d'une étude de leurs histogrammes caractérisant la distribution de leurs caractéristiques. La raison de ce choix tient principalement au fait que nous désirons garder une information sur la localisation des indices textures calculés(chose que ne permet pas l'utilisation d'histogrammes). En effet, nous désirons mettre en place un indice de similarité permettant de savoir si localement l'information texture entre deux images est identique ou non (comparaison pixel à pixel). Nous utiliserons donc une mesure basée sur une étude locale des caractéristiques texture et permettant au final de caractériser globalement l'image analysée. L'objectif de cette expérimentation consiste principalement à savoir si les indices textures proposés permettent de différencier différents types d'éléments de contenus.

Notons enfin que pour le moment, les applications présentées dans la suite de ce chapitre opèrent de manière naïve en comparant séquentiellement l'ensemble des images de la base avec l'image rêquête.

4.2.2 Mesure de similarité post classification

Visuellement, une page est caractérisée par l'organisation spatiale des pixels de textes, dessins et fonds. Sur la base de cette définition, nous proposons l'utilisation d'outils de comparaison de partitions présentés dans [YS04]. Dans le cadre de notre travail, une partition est le résultat d'une classification de pixels réalisée sur la base des indices textures générés. La figure **Fig. 4.1** permet d'illustrer 3 exemples différents de partitions. Ainsi, si l'on souhaite savoir si deux partitions sont similaires ou bien si elles diffèrent significativement, on peut se référer aux indices définis dans [YS04] qui permettent de quantifier ces similitudes.

Il est, selon nous, possible d'utiliser ces outils mathématiques de comparaison de partitions, dans un but de comparaison de pages d'ouvrages. En effet, si on applique un algorithme de classification (par exemple Clara vu dans le chapitre précédent) sur un ouvrage complet, nous obtenons une partition pour chaque image. Le fait d'appliquer une classification sur l'ensemble d'un ouvrage permet de tirer partie des avantages cités dans le chapitre précédent. Ainsi, l'exemple **Fig. 4.1** simule un résultat de classification à 3 classes réalisée sur un ouvrage composé de 3 pages (une seule classification est donc opérée sur l'ensemble des pixels de toutes les pages). Les deux premières images comportent 3 classes, le fait que la classification ait été réalisée sur l'ouvrage entier a permis d'identifier qu'il n'y avait que deux classes sur la dernière page. De même, chaque pixel bleu (resp. jaune ou rouge) de la page 1 appartient effectivement à la même classe que les pixels bleus (resp. jaune ou rouge) de la page 2 ou 3.

Ci-dessous nous détaillons les différentes étapes que nous proposons de reprendre pour la comparaison de deux classifications de pixels.

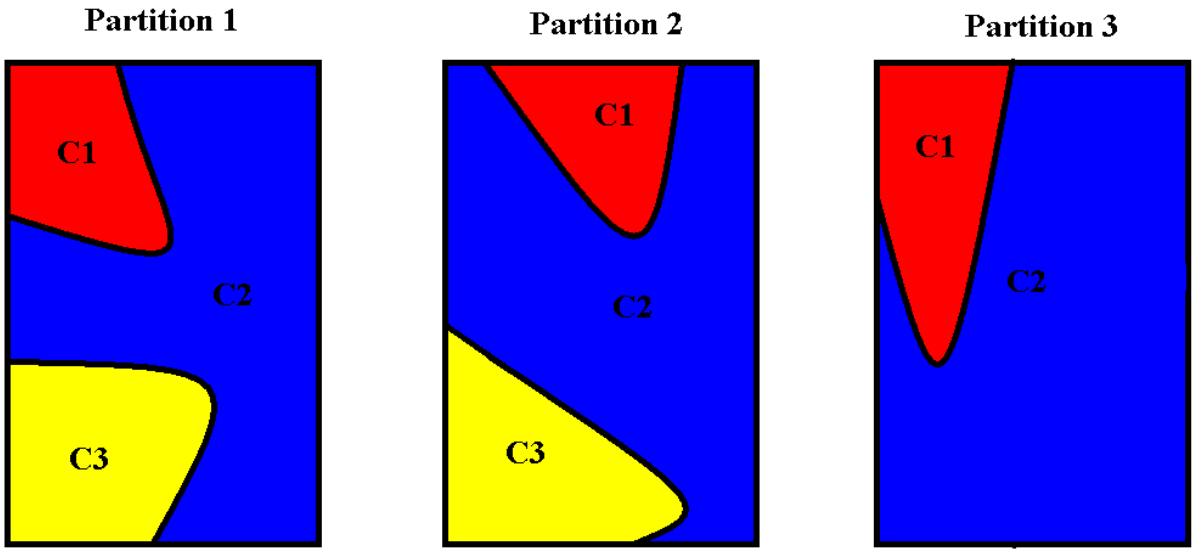


FIG. *4.1: Principe de la comparaison de partitions*

Une image α est représentée par un tableau relationnel R^{α}. Si l'image à n pixels, R^{α} est donc de taille nXn et sa construction est réalisée selon **Eq. 4.1** avec une complexité en $O(n^2)$.

$$\overset{\alpha}{R}_{ii'} = \begin{cases} 1 & \text{si } L^{\alpha}(i) = L^{\alpha}(i') \quad Avec \;\; L^{\alpha}(i) \;\; le \;\; label \;\; du \;\; pixel \;\; i \;\; de \;\; l'image \;\; \alpha \\ 0 & \text{sinon} \end{cases} \tag{4.1}$$

Dans certains cas il se peut qu'il ne soit pas commode de comparer des partitions en fonction des pixels classés (surtout pour des raisons de taille des données et de complexité algorithmique quand n est grand). En construisant le tableau de contingence $\overset{\alpha,\beta}{N}$ de deux images α, β (**Eq. 4.2**) il est possible de comparer des partitions dans un espace de données réduit ($\overset{\alpha,\beta}{N}_{uv}$ est de dimension pXq avec p le nombre de classes de l'image α et q le nombre de classes de l'image β) et une construction en O(n) (les deux images doivent donc avoir la même taille n).

$$\begin{aligned} \overset{\alpha,\beta}{N}_{uv \in p,q} &= \textstyle\sum_i X^i_{uv} \\ X^i_{uv} &= \begin{cases} 1 & \text{si} L^{\alpha}(i) = u \;\; et \;\; L^{\beta}(i) = v \\ 0 & \text{sinon} \end{cases} \end{aligned} \tag{4.2}$$

Dans [YS04], les auteurs montrent qu'il existe une relation linéaire entre la somme des R^{α} (nombre de paires de même classe dans une image) et la somme en ligne (ou colonne) des N_{uv} (**Eq. 4.3**).

$$\begin{aligned} \textstyle\sum_i \sum_{i'} R^{\alpha}_{ii'} &= \textstyle\sum_u N^2_{\alpha} \\ \textstyle\sum_i \sum_{i'} R^{\beta}_{ii'} &= \textstyle\sum_v N^2_{\beta} \end{aligned} \tag{4.3}$$

La comparaison de partitions se base sur le calcul de deux indices a et b où a est le nombre de paires de pixels ayant un même label dans la partition 1 et ayant toujours un label identique dans la partition 2 (**Eq. 4.4**).

$$a = \sum_{ii'} \Psi_{\alpha,\beta}^{ii'} \quad avec \quad \Psi_{\alpha,\beta}^{ii'} = \begin{cases} 1 & \text{si } L^\alpha(i) = L^\alpha(i') \text{ et } L^\beta(i) = L^\beta(i') \\ 0 & \text{sinon} \end{cases}$$

$$Soit \tag{4.4}$$

$$a = \sum_i \sum_{i'} R_{ii'}^\alpha R_{ii'}^\beta = \sum_u \sum_v N_{uv}^2$$

et b est le nombre de paires de pixels ayant un label différent dans la première partition et ayant également un label différent dans la deuxième partition (**Eq. 4.5**).

$$b = \sum_{ii'} \Omega_{\alpha,\beta}^{ii'} \quad avec \quad \Omega_{\alpha,\beta}^{ii'} = \begin{cases} 1 & \text{si } L^\alpha(i) \neq L^\alpha(i') \text{ et } L^\beta(i) \neq L^\beta(i') \\ 0 & \text{sinon} \end{cases}$$

$$Soit \tag{4.5}$$

$$b = \sum_i \sum_{i'} (1 - R_{ii'}^\alpha)(1 - R_{ii'}^\beta) = n^2 + \sum_u \sum_v N_{uv}^2 - \sum_u N_u^2 - \sum_v N_v^2$$

La figure **Fig. 4.2** illustre le principe du calcul des indices a et b. Nous avons choisi 5 points $(X_1, ..., X_5)$. Les indices a et b étudient l'évolution du label des paires de pixels d'une partition à une autre. Si l'on raisonne sur uniquement 5 points de la partition alors, les couples de pixels intervenant de le calcul de a sont : $\{(X_2, X_5), (X_3, X_4)\}$. Ceux intervenant dans le calcul de b sont : $\{(X_1, X_4), (X_1, X_3), (X_2, X_4), (X_2, X_3), (X_5, X_4), (X_5, X_3)\}$.

Ces indices permettent tous les deux de mesurer la stabilité entre 2 classifications. Nous insistons sur le fait que si une paire de pixels n'est pas dans la même classe sur la première image et qu'elle ne l'est pas non plus dans la deuxième, alors ceci renforce l'indice de stabilité (cf. le couple de pixels (X_2, X_5) de la figure **Fig. 4.2**)

Ainsi, selon nous, les indices calculés pour la comparaison de deux partitions et qui sont présentés dans [YS04] permettent de répondre à une requête de type "recherche de contenus similaires d'une page". En effet, l'originalité de la mesure que nous utilisons, est qu'elle évalue la stabilité d'association de pixels d'une classification à l'autre. Dans notre cas nous calculerons un pourcentage de "paires en accord" $(a + b)/n^2$ (n^2 étant le nombre maximum de paires) pour mesurer la similarité entre deux images par la formule **Eq. 4.6**. Deux images sont donc semblables, s'il y a une stabilité entre associations de pixels d'une image à l'autre.

$$R = (a + b)/n^2 = \frac{2 * \sum_u \sum_v N_{uv}^2 - \sum_u N_u^2 - \sum_v N_v^2 + n^2}{n^2} \tag{4.6}$$

Nous présentons les indices R^u car ils sont intuitivement plus simples à comprendre, mais en pratique nous les calculons dans N_{uv} en utilisant la relation linéaire qui les relie. C'est évidemment le grand nombre de pixels à traiter, qui nous a poussés à calculer les indices de similarité de cette manière.

Dans la version décrite dans [YS04], le nombre de paires en accord (**Eq. 4.6**) n'est pas lié aux labels des pixels après classification. Nous proposons donc une adaptation de **Eq. 4.6**, mais qui cette fois ci, tient compte du label des couples de pixels (**Eq. 4.7**).

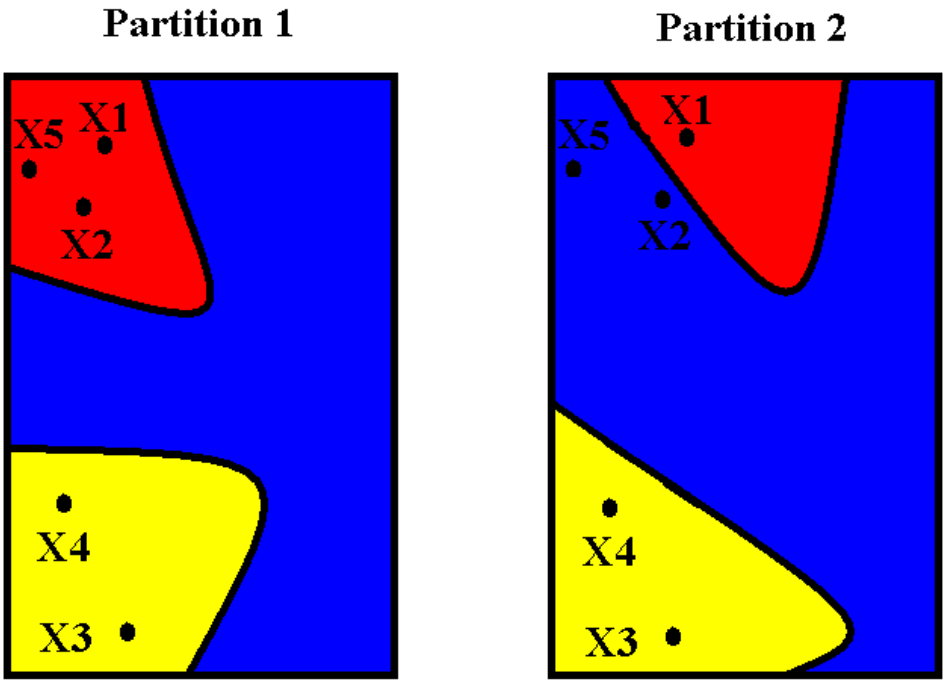


FIG. *4.2: Principe du calcul des indices a et b*

Si nous reprenons l'exemple de la figure **Fig.** 4.2, avec l'utilisation de l'indice **Eq.** 4.7 le couple de pixels $\{(X_2, X_5)\}$ n'entrera pas dans le calcul de a, puisque ce couple de pixels ne respecte pas le fait que les 4 pixels doivent être de même label.
De ce fait, nous décidons de prendre en compte dans cette adaptation de l'indice de stabilité, uniquement les paires de pixels qui ont le même label d'une image à l'autre. Ceci se traduit par une modification du calcul de l'indice a en a'. L'indice b n'est pas modifié et le nombre maximum de paires ne change pas, on normalise donc par n^2.

$$a' = \sum_{ii'} \Psi_{\alpha\beta}^{ii'} \quad avec \quad \Psi_{\alpha\beta}^{ii'} = \begin{cases} 1 & \text{si } L^\alpha(i) = L^\alpha(i') = L^\beta(i) = L^\beta(i') \\ 0 & \text{sinon} \end{cases} \tag{4.7}$$

Ce qui donne la mesure de similarité finale **Eq.** 4.8

$$R' = (a' + b)/n^2 = \frac{\sum_u \sum_v N_{uv}^2 + \sum_u N_{uu}^2 - \sum_u N_u^2 - \sum_v N_v^2 + n^2}{n^2} \tag{4.8}$$

L'utilisation de l'indice de comparaison de partitions nous permet donc de comparer des résultats issus de différentes classifications. Cette mesure est peu sensible à la position géographique des pixels de différents labels sur la page. Ce que nous mesurons ici est plutôt lié à la notion de proportions des labels présents dans les différentes versions de la classification.
Au travers d'une application permettant la comparaison de pages de documents, nous verrons que ce choix de mesure de similarité post-classification, représente une vraie alternative à une solution qui consisterait à comparer directement les attributs textures de chaque pixel.

4.2.3 Mesure de similarité pour la comparaison pixel à pixel d'indices texture

Comme nous l'avons énoncé dans l'introduction de cette section, nous souhaitons calculer une similarité entre deux images en fonction des similarités locales des textures qui la composent. Pour ce faire, nous proposons l'utilisation d'une métrique permettant de mesurer une similarité globale à l'aide d'une succession de mesures de similarités locales.

Soit Sim la fonction qui au couple de pages P_k, P_l associe le calcul d'une distance de similarité d.

$$Sim : (P_k, P_l) \longmapsto d(P_k, P_l)$$

Si l'on soustrait la matrice d'indices textures de l'image k à la matrice d'indices textures de l'image l, alors plus la différence terme à terme sera faible, plus les pixels auront une texture similaire. Le fait de multiplier la matrice par elle même, et de calculer la somme des éléments de la matrice, permet d'obtenir une similarité globale entre les deux matrices textures. Ainsi, plus la somme totale sera faible, plus les deux images seront considérées comme "ressemblantes".

$$d(P_k, P_l) \quad = \sqrt{trace((\overset{k}{C}_{i,j} - \overset{l}{C}_{i,j}).{}^t(\overset{k}{C}_{i,j} - \overset{l}{C}_{i,j}))}$$

4.3 Exemples d'exploitation des indices textures proposés

Nous rappelons ici, que nous ne prétendons pas réaliser d'applications complètes de Recherche d'Information par le Contenu (RIC), mais des tests de pertinence de nos caractéristiques de texture. Il faudrait, comme cela a été présenté en introduction, prendre en compte d'autres critères (couleurs, formes,...) et également réfléchir à la place de l'utilisateur dans le système. Certains auteurs préconisent son intervention pour corriger les résultats des requêtes et ainsi permettre d'affiner la qualité des réponses.

4.3.1 Comparaison de pages

4.3.1.1 Proposition d'un protocole d'évaluation

Notre objectif de comparaison de pages nous amène à faire face à certaines difficultés. Tout d'abord, il se pose légitimement la question de savoir comment on détermine si deux pages se ressemblent ? Évidemment, selon l'application, cette notion diffère. Après réflexion, nous nous sommes inspirés des travaux de [MMS06]. Dans leurs travaux relatifs à la conception d'un CBIR d'images de documents, les auteurs ont pris le parti de séparer les images de leur corpus en 8 classes différentes. Ces classes sont extraites à partir d'une combinaison des caractéristiques de mises en pages trouvées dans leurs images. Les auteurs recherchent ensuite des pages de titres, des pages avec des équations mathématiques, des pages sur deux colonnes de texte, des pages sur deux colonnes de textes et une image en haut, des pages sur deux colonnes de textes et une image en bas... Nous avons adopté ce type de classement pour évaluer la comparaison de pages d'ouvrages anciens.

Nous avons donc décidé de séparer les documents en 4 classes différentes :

– Les pages avec une cadre qui entoure complètement le contenu.

- Les pages composées uniquement de texte et justifiées à droite et à gauche
- Les pages composées uniquement de texte mais cette fois ci disposé sur deux colonnes
- Les pages composées d'une lettrine et le reste de la page composée uniquement de texte.
- Les pages composées entièrement de dessins

Les résultats montrés dans la suite de ce chapitre ont tous été réalisés sur la même base d'images. Nous avons ainsi choisi près de 200 pages de 9 ouvrages différents (cf. annexes pour voir des extraits de cette base). Chaque test est effectué sur la base entière selon le protocole suivant :

1. Application de l'algorithme de classification Clara pour 3 classes. Les pixels constituant l'échantillon sont choisis aléatoirement parmi l'ensemble des pixels des 200 images. A l'issue de cette étape, chaque pixel de chaque page est classé dans l'une des 3 classes. Pour chaque image nous obtenons donc une partition correspondant à la répartition texte/fond/dessin.
2. Chaque partition est comparée à toutes les autres à l'aide de l'indice **Eq. 4.7**.
3. L'ensemble des comparaisons permet de construire une matrice de similarité entre les images constituant la base étudiée.

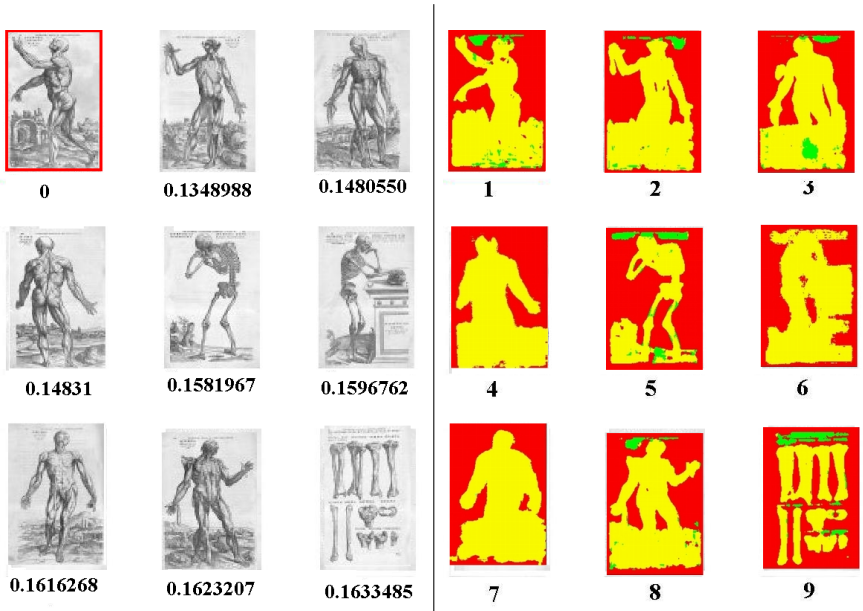
Il est à noter que les résultats de classifications donnés dans la section suivante l'ont été pour une classification à 3 classes. De ce fait, la comparaison se résume à l'étude du ratio des pixels de textes/fonds/dessins. Il n'est donc pas possible de différencier deux types de dessins différents.

4.3.1.2 Etude des résultats de comparaison de pages

Sur les 200 images ayant servi à ces tests, nous allons évaluer la capacité à retrouver, par exemple, toutes les pages parmi tous les ouvrages de notre base qui sont composées avec un cadre lorsque l'on donne en requête une image avec un cadre. Dans notre base, il n'y a pas que ces 5 catégories de pages. Nous avons volontairement introduit des images pouvant perturber les résultats obtenus par calcul de similarité : nous avons par exemple introduit des pages de documents contemporains.

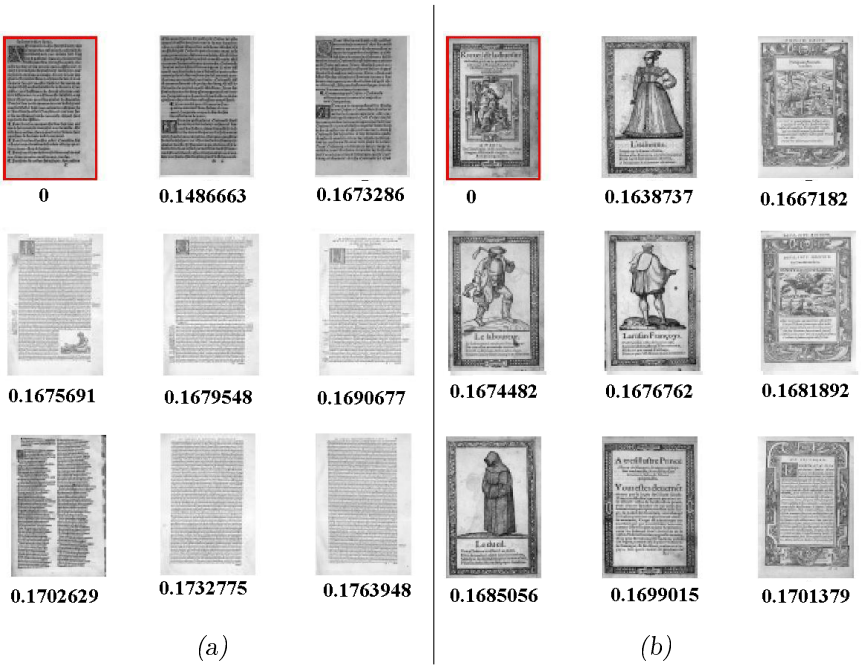
La figure **Fig. 4.3**.a illustre la capacité de l'indice utilisé à discerner des pages visuellement similaires. L'image requête utilisée dans la figure **Fig. 4.3**.a possède la caractéristique d'être composée en grande partie d'une illustration et d'une ou deux lignes de textes. La base comporte une dizaine d'images avec un être humain (ou un squelette). On remarque que seule la dernière réponse ne correspond pas à l'image requête. Cependant, au sens de l'indice R', il peut sembler logique que cette image composée d'os fasse partie des réponses, puisqu'elle est composée de pixels de textes et de dessins dans les mêmes proportions que l'image requête. La figure **Fig. 4.3**.b représente les partitions qui ont été étudiées afin de mesurer la similarité entre pages.

De l'ensemble des tests que nous avons réalisés, nous souhaitons présenter ceux correspondants aux figures **Fig. 4.4**a-b. La base est composée de pages provenant de plusieurs ouvrages. Certaines pages ont donc des caractéristiques communes (cadre, texte accompagné de lettrine, texte sur deux colonnes...). Les résultats présentés montrent, dans la figure **Fig. 4.4**.a, que les réponses 1, 2 et 3 proviennent du même ouvrage alors que les réponses 4,5,6,8,9 proviennent d'un autre. La page placée en septième position est issue d'un troisième ouvrage. Cette page possède une petite lettrine et une disposition sur deux colonnes, elle possède donc également des caractéristiques proches de l'image requête.
De même, les résultats de la figure **Fig. 4.4**b, montrent que pour une image composée d'une zone de dessin avec du texte au dessus et au dessous, le tout entouré d'un cadre ; les réponses sont issues de deux ouvrages (les pages 1,2,4,5,7 d'un coté, et 3,6,8,9 de l'autre).



(a) Images initiales (b) Classification obtenue

FIG. 4.3: Exemples de résultats de requêtes utilisant l'indice de comparaison de partitions modifié (R')



(a) (b)

FIG. 4.4: Exemples de résultats de requêtes utilisant l'indice de comparaison de partitions modifié (R')

L'intérêt de notre proposition d'une comparaison de pages en calculant un indice de similarité sur le résultat de la classification est double. En effet, non seulement ce choix évite de coûteux

calculs entre vecteurs de caractéristiques, mais cela permet également de fournir un indice de similarité peu sensible à la position des éléments dans une page. Sur ce dernier point précis, nous avons désiré comparer les résultats que nous obtenons en utilisant notre méthode, avec les résultats obtenus en calculant une mesure naïve qui consiste à étudier les niveaux de gris des pixels des deux images. Ainsi, nous avons effectué les tests sur la même base d'images, avec cette fois une mesure de similarité entre deux images P_k et P_l égale à la somme des différences des niveaux de gris : $d(P_k, P_l) = \sum_i (|P_k(i) - P_l(i)|)$. La figure **Fig.** 4.5 illustre deux résultats obtenus selon l'indice choisi. Ces résultats reflètent relativement bien ceux obtenus sur l'ensemble de la base. La qualité des réponses obtenues en comparant les indices de niveaux de gris, est significativement faible. En opérant de la sorte, il est impossible d'identifier des pages de contenu similaire. La raison principale de cette confusion vient du fait que les niveaux de gris correspondant aux pixels de dessins sont très proches de ceux correspondant aux illustrations.



(a) Similarité calculée après classification



(b) Similarité calculée en fonction des niveaux de gris



(c) Similarité calculée après classification



(d) Similarité calculée en fonction des niveaux de gris

FIG. *4.5: Comparaison des résultats obtenus de deux manières différentes*

Dans la suite de cette section, nous proposons d'évaluer la capacité de notre indice à différencier plusieurs types de pages. Le protocole d'évaluation que nous avons mis en place est lié

à la perspective d'une application que nous souhaitons développer dans un avenir proche. Cette application devra permettre d'aider un utilisateur à naviguer dans un ouvrage (ou ensemble d'ouvrages). L'un de ces outils d'aide à la navigation devra permettre à un utilisateur d'avoir accès aux pages dont le contenu ressemble à la page qu'il donne en requête.

Le tableau (**Table 4.1**) permet de résumer les taux de bonnes réponses obtenus pour 5 types de requêtes différents. Les résultats correspondent à des taux de précision pour un Top5, Top10 et Top 15. Le taux de précision, est usuellement associé avec le taux de rappel. Ces deux taux sont d'ailleurs fortement liés et dépendent du seuil choisi pour la similarité minimum pour qu'une image soit acceptée. En fonction des critères que nous venons de détailler, nous préférons calculer de manière très simple un taux de précision en divisant le nombre de bonnes réponses obtenues après requête par le nombre d'images considérées (taille du Top étudié) $Taux = \frac{Bonnes\ \ reponses}{Taille\ \ du\ \ Top}$. Sur les 200 images de la base, nous avons sélectionné un peu plus d'une centaine de pages correspondant à des caractéristiques visuelles précises (Cadres, Bi-colonnes, Dessin, Pleins texte, Pages avec lettrines). Pour chacune de ces images nous avons calculé le taux de précision pour trois tailles de Top différents. Une moyenne de ces taux a permis d'obtenir un taux global par classe de pages.

	Top5	*Top10*	*Top15*
Cadres	1	0.93	0.86
Bi-colonnes	0.93	0.76	0.78
Dessins	0.9	0.62	0.6
Pleins Texte	0.74	0.56	0.50
Lettrines	0.65	0.56	0.55

TAB. *4.1: Taux de précision obtenus pour 5 types de requêtes différents*

Dans les tests effectués, toutes les pages de tous les ouvrages sont mélangées. Comme énoncé précédemment, la mesure permet de discriminer des structures visuellement très différentes les unes des autres. Par exemple, pour une image avec un cadre, le taux de précision est de 100% pour un top 5, 93% pour un top 10 et 86% pour un top 15. Le cadre est un élément très discriminant. On pourrait faire la même remarque pous les images composées uniquement de dessins.
En revanche, pour des images composées de lettrines, le taux de précision se situe autour de 65% pour le top 5 et chute à 56% pour le top 10 et 55% pour le top 15. Ce mauvais résultat est dû au fait que ces images sont confondues avec les pages composées uniquement de texte. La mesure utilisée ne permet pas de faire des requêtes de cet ordre de finesse. Dès lors que les lettrines sont petites, le nombre de pixels de dessins n'est pas suffisament important pour influencer le calcul de similarité. Une solution consisterait à pondérer l'importance des différentes zones constituant la page. Par exemple, une pondération relative à la quantité de chaque classe dans une page, une pondération relative à la localisation ou la disperssion des éléments d'une classe dans une page, permettrait certainement d'améliorer les résultats. Nous reviendrons sur cette pondération dans les perspectives directes en fin de chapitre.
La classification à 3 classes opérée pour ces tests, fait qu'il n'est pas possible de différencier deux type de dessins. De ce fait, certaines pages comportant des illustrations de tailles équivalentes à une lettrine sont jugées comme similaires.

A travers des expériences menées sur la comparaison de pages, nous avons principalement validé la possibilité de décrire et de comparer des pages de documents à l'aide d'informations textures sans avoir à effectuer un processus de segmentation/rétro-conversion. Les premiers résultats sont encourageants et montrent la pertinence de l'utilisation de l'indice de comparaison de

partition après classification des pixels. En comparaison avec les approches utilisant la modélisation par graphes, nous proposons une solution alternative permettant de mesurer une similarité entre pages, sans avoir à mettre en place un processus complexe de création de graphes et des mesures d'appariement qui en découlent.

4.3.2 Comparaison d'images

4.3.2.1 Proposition d'un protocole d'évaluation

Les CIR cités en introduction de ce chapitre, permettent à un utilisateur de retrouver des images naturelles similaires à celle qu'il donne en exemple. Dans le système Kiwi ([Lou00]) il est par exemple possible d'effectuer des requêtes sur des images d'émission TV, des images naturelles, des photos de visages... Dans les travaux proposés par [Haf05], il est par exemple possible de rechercher des images de photos aériennes...
Pouvoir réaliser un recherche d'images par le contenu sur une base constituée d'images de traits de documents anciens, nous semblait intéressant à expérimenter. Ainsi, nous avons tout d'abord constitué une base de tests contenant plus de 400 images de traits. Ces images sont disponibles sur le site des bibliothèques virtuelles de Tours [12]. La figure **Fig. 4.6** illustre quelques uns de ces dessins. Les lettrines sont les illustrations les plus présentes dans les documents anciens. De ce fait, plus d'un tiers de la base en est constitué (toutefois les styles divergent d'un ouvrage à l'autre), le reste se divise en plusieurs catégories : blasons, personnages, emblèmes, crânes, éléments décoratifs divers...

Cet objectif de recherche d'images par le contenu nécessite de pouvoir comparer les textures composant les images de traits. Comme énoncé dans l'introduction de cette section, nous désirons comparer localement l'information texture entre deux images. Ce choix fait qu'il n'est pas pertinent de comparer des images constituées de plusieurs textures (par exemple les images naturelles). Les dessins de traits constituant la base ont la particularité d'être composés de textures homogènes. Par exemple, il existe des différences de styles entre les lettrines (**Fig. 4.6**) qu'il serait intéressant de pouvoir caractériser. Les crânes, les blasons ou les icônes qui constituent notre base se caractérisent elles aussi par des textures bien spécifiques. Ce choix induit qu'il ne sera pas abordé ici la comparaison des images constituées de plusieurs textures.

Pour réaliser cette expérience, nous utilisons une comparaison sans classification de pixels. En effet, si classer les pixels en 3 classes sur des images de documents a un sens (séparer le texte des illustrations et du fond), ce n'est pas le cas avec des images de traits. Il n'est, en effet pas possible de maîtriser le comportement de la classification (elle diffère selon le nombre de classes désirées et selon la nature de l'image). Une approche basée sur la comparaison de zones après classification des pixels ne nous semble donc pas pertinente. nous préférons de manière plus classique, comparer directement les attributs textures des pixels. Ce choix nous amène à poser le postulat que deux images sont identiques si elles possèdent les mêmes attributs d'orientations et de fréquences au même endroit.

4.3.2.2 Etude des résultats

Dans cette section, les tests réalisés correspondent à une recherche d'image par l'exemple. Ainsi, une image est donnée en requête (entourée en rouge dans les exemples qui suivent) et le

[12] Base constituée en coopération avec l'ACI MADONNE : http ://www.bvh.univ-tours.fr/madonne.asp

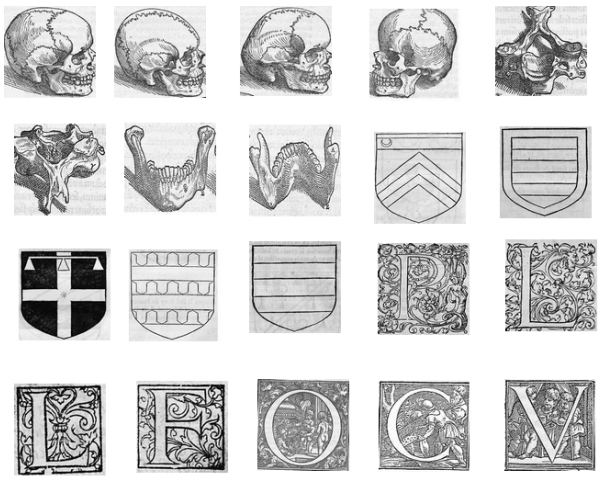
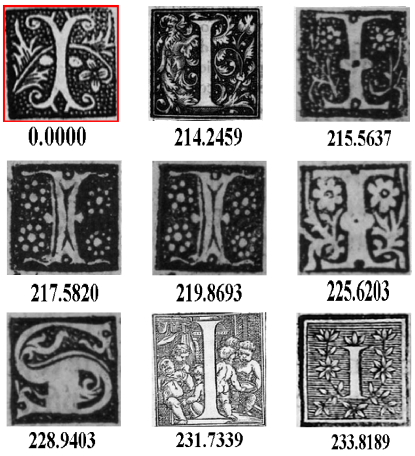


FIG. 4.6: *Extraits des illustrations composant la base d'images testée*

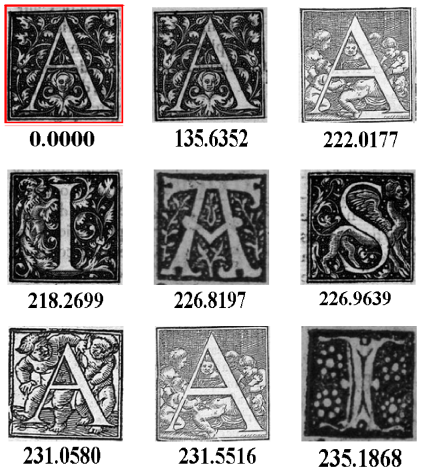
système fournit les images qui lui sont le plus similaire.
La figure **Fig.** 4.7.a-e illustre les bons résultats obtenus sur la base d'images de traits. Après étude des résultats, nous constatons que la discrimination des différentes catégories d'illustrations de la base est conforme aux attentes. Sur plus d'une centaine de lettrines testées, la majorité des réponses obtenues dans un top 20 sont des lettrines. En étudiant plus précisément les réponses, on constate même que les indices calculés sont sensibles à la lettre de la lettrine. On peut, par exemple, voir que sur la figure **Fig.** 4.7.b la lettrine requête est un I et que les réponses les plus proches sont également des I. On remarque le même type de classement obtenu sur une lettrine A (**Fig.** 4.7.c).Sur les figures **Fig.** 4.7.a-d-e, on peut constater la qualité des résultats obtenus pour d'autres illustrations que les lettrines. On remarquera la (mauvaise) présence d'une lettrine pour une requête où l'image est un blason. La figure **Fig.** 4.7.f illustre le type d'erreurs couramment rencontré. Notre base est composée d'icônes aux traits caractéristiques très proches de celles rencontrées dans certaines lettrines. On remarque que certaines d'entre elles sont présentes dans les réponses les plus proches. Ceci illustre bien toute la subjectivité de la notion de texture.



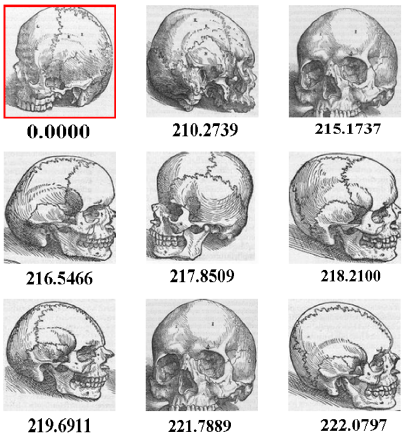
(a) *Recherche de portraits*



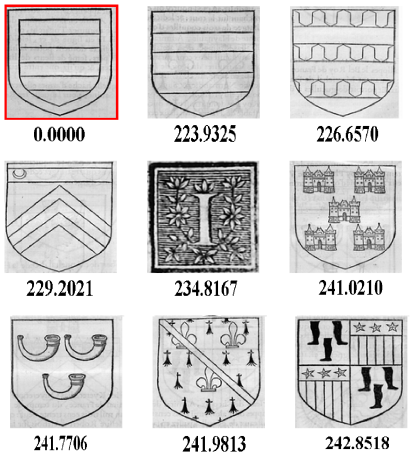
(b) *Recherche de lettrines*



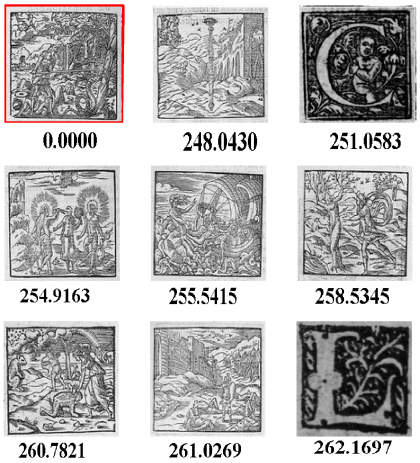
(c) *Recherche de lettrines*



(d) *Recherche de crânes*



(e) *Recherche de blasons*



(f) *Recherche d'icônes*

FIG. *4.7: Recherche dans un ouvrage*

Pour permettre une évaluation globale des requêtes effectuées, nous proposons de mettre en place le même protocole que celui utilisé pour la comparaison de pages. Ainsi nous calculons un taux de précision pour un top5, 10 et 15 pour 5 textures différentes de la base. Le tableau **Table 4.2** récapitule les taux moyens obtenus.

	Top5	*Top10*	*Top15*
Lettrines	0.95	0.92	0.90
Portraits	0.92	0.90	0.89
Cranes	0.91	0.86	0.79
Icones	0.90	0.87	0.78
Blasons	0.88	0.78	0.73

TAB. *4.2: Taux de Précisions pour des requêtes effectuées sur des images de traits*

Pour ce qui est des comparaisons réalisées sur des lettrines, les taux de précisions obtenus sont moins bons que ceux trouvés dans la littérature. Dans l'article [RP05] les auteurs arrivent à discriminer, parfois avec 100% de précision, différents styles de lettrines. De même, dans la référence [PVU+06], les auteurs mettent en place un système permettant la comparaison de lettrines avec un taux de bon classement de 94%. Cela souligne en particulier la nécessité d'introduire éventuellement d'autres fonctionnalités propres à un système de CBIR. Nous obtenons donc de moins bons résultats que ceux obtenus par des méthodes dédiées à l'analyse ce type d'images (et plus spécifiquement des lettrines). Si nous voulions mettre en place une application équivalente à celles que l'on trouve dans la littérature, il faudrait repenser le mode de comparaison et prendre en compte d'autres critères.

4.3.3 Conclusion de nos expérimentations

Nous avons réalisé des tests de faisabilité sur deux types d'applications : l'une dédiée à la recherche de documents dont la nature des contenus est similaire (similarité de pages) et l'autre réservée à la comparaison de textures dans les dessins de traits. Dans le premier cas, les résultats de la classification des pixels à partir des indices textures sont utilisés pour comparer les pages selon une méthode de comparaison de partitions. Dans le deuxième cas, nous comparons directement les indices textures de chaque pixel afin d'obtenir une similarité globale entre deux images.

A court terme, nous pensons apporter plusieurs améliorations à ces méthodes. Parmi celles possibles, il nous semble intéressant d'améliorer les temps de réponses après requête qui, dans la version que nos proposons, sont trop longs (plusieurs minutes de calculs). Une voie possible serait de mettre en place une solution évitant une recherche linéaire dans la base. Un autre point qu'il nous semble intéressant d'approfondir est lié aux conclusions données dans le chapitre précédent et disant qu'il est possible d'échantillonner nos données textures sans pour autant nuire à la qualité globale de l'information. Il est donc envisageable de ne calculer les indices de comparaison de partitions ou de distances entre matrices que sur une partie des pixels des images. Une autre amélioration, serait de pouvoir pondérer certaines caractéristiques entrant en jeu dans le calcul de similarité. Pour ce qui est de la comparaison de pages, dans la version actuelle que nous proposons, une faible représentation d'une classe dans une page (eg : un petit dessin) ne permet pas d'influer suffisamment sur le calcul global de similarité.

A plus long terme, ces premières expériences nous permettent d'envisager la mise en place

d'applications dédiées à l'indexation d'images de documents anciens (segmentation, spotting, outils d'aide à la navigation...)

4.4 Vers de véritables applications de recherche d'information par le contenu

Les expériences précédentes avaient surtout pour but de valider la pertinence des indices textures proposés pour la comparaison d'images de documents. Pour cela, elles n'utilisaient que l'information texture associée à des techniques de comparaisons assez simples à mettre en place. Pour finir ce mémoire, nous proposons de nous attarder sur ce que nous pensons être les points intéressants à travailler ou à étudier sur la base de nos propositions et pouvant déboucher à terme sur de réelles applications.
Dans ce mémoire, nous avons proposé plusieurs éléments de réponses relatifs au problème de l'analyse de masse de données d'images de documents anciens. Cependant, il reste encore certains points posant problème si l'on veut mettre en place de réelles applications. Ainsi, le nombre d'images constituant les bibliothèques numériques pose le problème du traitement de masses de données. Si l'on veut, par exemple, proposer des applications temps-réel à un utilisateur, le temps nécessaire au traitement de plusieurs centaines d'images ne doit pas être une contrainte. Le second point posant problème est celui de la segmentation. Pouvoir séparer les différents éléments constituant les pages de documents anciens, reste à ce jour problématique.

4.4.1 Vers la recherche d'éléments de contenu

Accéder au contenu d'images de documents reste un problème complexe à traiter. Dans ce chapitre nous avons vu qu'il était possible de comparer des blocs d'illustrations pré-segmentés sur la base de leur attributs textures. Cependant, transposer cet objectif sur des blocs non segmentés (ie : rechercher une information dans une image), nécessite de pouvoir appréhender le problème du traitement et de l'analyse de grosses masses de données.

Ainsi, rechercher un motif précis dans un ensemble conséquent de pages pose le problème du temps de calculs nécessaires à ce traitement. Ce problème (proche de celui du "word spotting") ne peut raisonnablement se résoudre à une comparaison exhaustive du motif recherché avec l'ensemble des pixels des images d'un ouvrage.

Dans l'optique de rendre exploitable un tel outil de recherche de motifs, nous proposons d'utiliser les informations textures obtenues lors de la classification de pages d'ouvrages que nous avons mis en place. Le schéma **Fig.** 4.8 détaille le fonctionnement de notre approche.

- Point A : Une classification à n classes est opérée sur l'ensemble des pages. Dans l'exemple donné, une classification à 3 classes est réalisée et chaque pixel est donc étiqueté rouge, vert ou jaune (Réalisation hors ligne).
- Point B : Pour chacune des classes, est calculé le centre de classe (Réalisation hors ligne). Les 20 indices textures représentatifs de chacune des classes de texture présentent dans l'image, sont stockées comme métadonnées de description synthétique.
- Point C : Un utilisateur utilise une image requête n'appartenant pas à la base. En temps réel, les indices textures sont calculés pour chaque pixel de l'image requête et un vecteur moyen caractérisant cette zone est déterminé.
- Pont D : Le vecteur caractérisant la zone requête est comparé aux vecteurs de centres de classes de chaque image indexée. Un classement de ces dernières peut être produit

en fonction des mesures de similarité obtenues par comparaison avec les caractéristiques textures de l'image requête.

– Point E : L'affichage des résultats peut permettre la navigation de l'utilisateur dans un ouvrage.

L'architecture de recherche que nous proposons possède l'avantage de ne pas avoir à recalculer une classification à chaque fois qu'un utilisateur effectue une requête (l'image qu'il donne n'est pas issue de la base). De plus, cette architecture permet de réduire considérablement les temps de calculs relatifs à la phase de comparaison avec l'image requête.

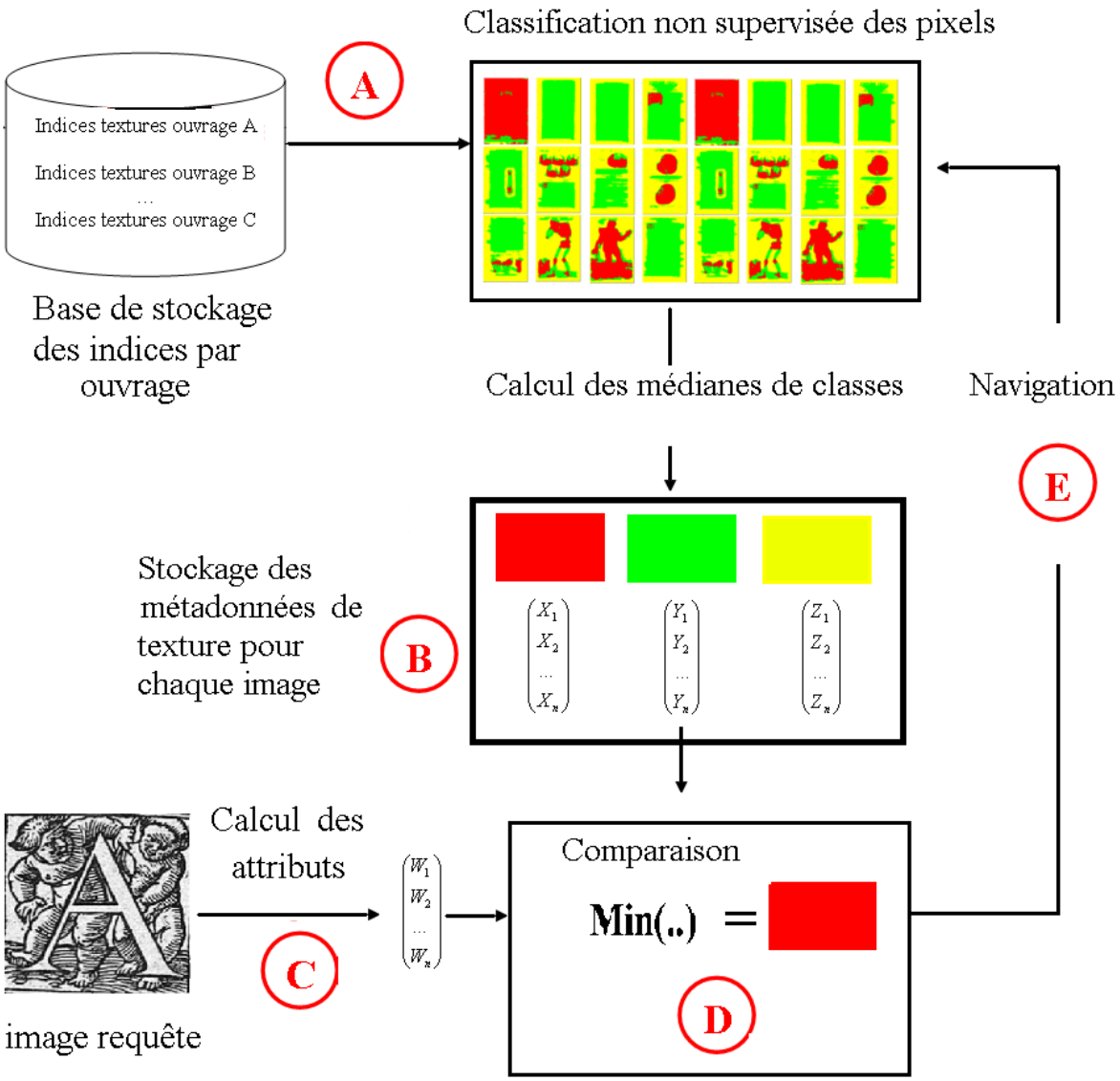


FIG. 4.8: *Principe de la recherche d'éléments de contenu*

4.4.2 Vers une meilleur segmentation des images de documents

Dans le deuxième chapitre, nous avons souligné toute la difficulté que représente la segmentation d'images de documents anciens. Le principal défaut des approches classiques (ascendantes, descendantes ou mixtes) se trouve être le manque de généricité. Nous avons également vu que Les approches par filtres de Gabor ont, quant à elles, des difficultés à traiter des images de traits. La classification de pixels que nous proposons est elle aussi insuffisante pour segmenter directement l'image en blocs de contenus homogènes (segmentation des illustrations, segmentation en titres ou paragraphes...). En effet, comme la plupart des approches textures, seul un marquage des pixels est réalisé. Une phase de post-traitements est donc indispensable pour permettre la réalisation d'une segmentation.

Il semble donc intéressant de coupler plusieurs approches pour pallier aux inconvénients de chacune. Ainsi, dans cette section, nous proposons une expérience de segmentation basée sur une étude des textures contenues dans les composantes connexes d'une image.

Afin de simplifier cette phase d'analyse des composantes, nous proposons d'étudier les attributs textures dans chaque composante de l'image afin d'en déterminer son label. Le schéma **Fig. 4.9** détaille le fonctionnement de notre proposition.

- Point A : Une classification à 3 classes est opérée sur l'ensemble des pages. Dans l'exemple donné, chaque pixel est donc étiqueté rouge, vert ou jaune (Réalisation hors ligne). Ceci permet d'identifier les différentes zones de dessins/textes/fonds.
- Point B : L'image que l'on souhaite segmenter est binarisée et les composantes connexes sont extraites.
- Point C : Une analyse des informations textures associées à chaque pixel d'une composante connexe permet de statuer sur son label. Ce processus est réalisé pour chaque composante.
- Point D : L'affectation d'un label est effectuée pour chaque composante (avec un taux de confiance associé)

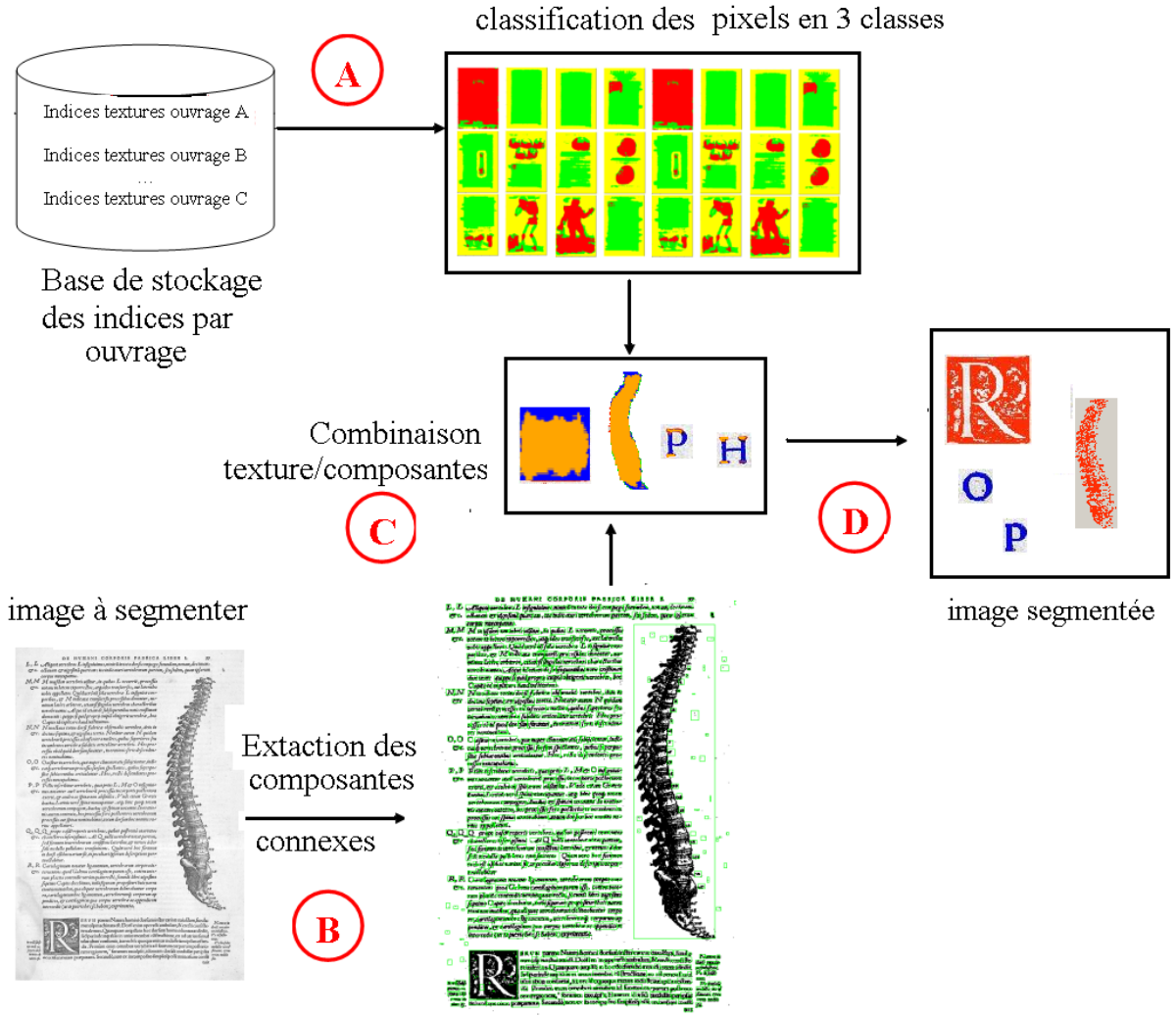


FIG. 4.9: *Segmentation combinant les informations composantes connexes et indices textures*

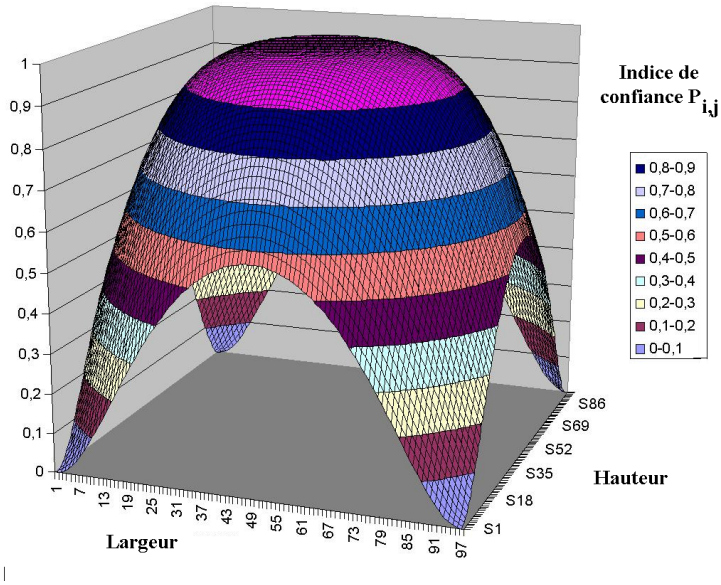
La majorité des erreurs de classification des pixels à partir des informations textures, se situe au niveau des illustrations (proximité avec du texte) et des titres (caractères trop grands qui sont assimilés à des dessins). La difficulté est donc de prendre la meilleure décision de labellisation possible pour la composante connexe étudiée car il y a de fortes chances que plusieurs classes de pixels s'y trouvent (3 au maximum). Nous avons donc décidé d'associer un taux de confiance à cette opération.

Cette opération d'étiquetage est rendue possible grâce à l'étude de la classe des pixels présents dans la composante analysée. Pour cela, nous calculons deux coefficients pour chaque composante.

On associe à chaque pixel de coordonnées (i, j) un poids $P_{(i,j)}$ relatif à la distance qui le sépare du centre de la composante. Par expérimentation, nous avons constaté que la majorité des erreurs de marquage se situent aux extrémités des composantes. Ceci est dû à la difficulté d'analyser les textures aux zones de transitions (fenêtre d'analyse à cheval entre du texte et une illustration).

Le premier indice de confiance s'inspire du modèle de Tuckey (**Eq. 4.9**) et permet de n'attribuer un taux de confiance faible aux pixels qui sont significativement éloignés du centre de gravité de la composante analysée. Les composantes extraites sont représentées par l'intermédiaire de leur rectangle englobant de largeur l, de hauteur h, de centre de gravité $C = \left(\left(\frac{l}{2}\right), \left(\frac{h}{2}\right)\right)$ et de distance maximale au centre de gravité $T^2 = \left(\frac{l}{2}\right)^2 + \left(\frac{h}{2}\right)^2$. L'indice est indépendant du résultat de la classification et dépend simplement de la forme de la composante. La figure à la droite de l'équation **Eq. 4.9** illustre le fait que les pixels au centre de la composante ont un indice de confiance $P_{(i,j)}$ proche de 1 alors que ceux qui sont éloignés du centre ont un indice $P_{(i,j)}$ proche de 0.

$$P_{ij} = 1 - \left(\frac{d^2\left(C, x_{i,j}\right)}{T^2}\right)^2 \qquad (4.9)$$



Le fait que les erreurs principales se situent sur les bords des composantes, fait que les pixels mal classés sont souvent isolés les uns des autres. Il n'y a pas de grandes plages de pixels mal classés mais plutôt des petits agrégats de pixels disséminés un peu partout dans la composante. Nous proposons donc de calculer un deuxième indice de confiance lié à ce phénomène. Ainsi, pour chaque pixel d'une composante, lui est associé un indice $V_{(i,j)}$ relatif au nombre de pixels voisins qui sont du même label que le sien. Donc, plus un pixel (i, j) possède de voisins avec même label que lui, plus le taux de confiance associé sera élevé. Ainsi, pour une étude de l'ensemble Ω des pixels voisins du pixel étudié (i, j) de label $L_{(i,j)} = \alpha$, on calcule la proportion de voisins de même label α :

$$V_{i,j} = \frac{\sum_{u,v \in \Omega} L_{(u,v)} = \alpha}{Card(\Omega)} \qquad (4.10)$$

Le taux de confiance final affecté au pixel (i, j) est égale à la moyenne des indices $P_{(i,j)}$ et $V_{(i,j)}$.

De même, un taux de confiance associé à la composante connexe a est égale à la moyenne des taux de confiance des pixels de cette zone.

Le label de la composante est choisi en faisant un simple vote majoritaire en tenant compte des indices de confiance associés à chaque pixel. La composante est donc du même label que le label le plus présent dans la composante à l'indice de confiance prêt. En accordant plus de confiance à certains pixels qu'à d'autres, on associe à cette décision le taux de confiance C_a. La

figure **Fig.** 4.10 illustre le résultat de labellisation de composantes connexes après analyse de leurs textures. Dans chacune de ces composantes la texture rouge (dessin) est majoritaire, elles sont donc toutes labellisées "dessin" (rouge). A chacune de ces labellisations, est associé un taux de confiance.

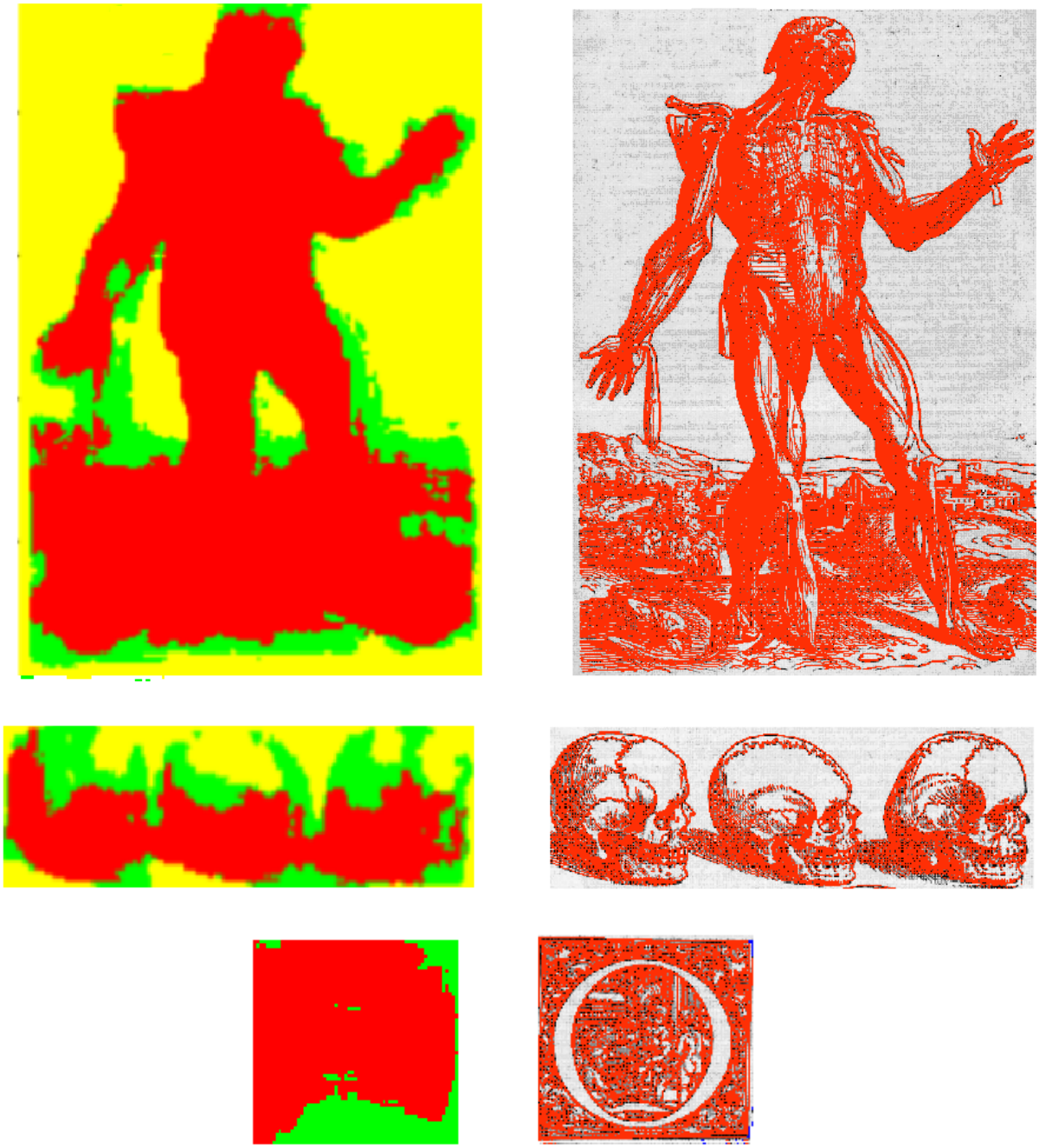


FIG. 4.10: *Exemple de labellisations de composantes de dessins de traits*

4.4.3 conclusion

Dans ce chapitre, nous avons proposé un ensemble d'expérimentations basées sur l'analyse des textures composant les images de documents anciens. Il ne s'agit ici que d'apporter des éléments de réponses plus concrets aux interrogations posées au début de ce chapitre. Ainsi, nous avons détaillé comment il était possible, à l'aide des indices des textures calculés, de caractériser le contenu d'une image de documents de différentes manières et avec différents objectifs (mise en page, illustrations, texte, ...) sans pour autant avoir segmenté ou retrouvé la structure logique précise ou sans fixer différents paramètres complexes à appréhender par des utilisateurs non experts en analyse d'images de documents.
En terme d'outils d'indexation d'images de documents anciens, dont nous avions fait ressortir le besoin, les expériences de comparaisons de pages, de recherche de motif ou de segmentation permettent d'entrevoir les nombreux avantages apportés par l'adjonction d'une approche par caractérisation n'utilisant aucun a priori sur la spécificité des images de documents traitées. De nombreuses questions posées tout au long de ce manuscrit restent encore à approfondir. Nous pensons plus particulièrement à la place que l'utilisateur peut prendre dans les systèmes d'indexation que ce soit en tant qu'expert permettant d'apporter ses connaissances ou en tant qu'utilisateur recherchant un ensemble d'informations.

Pour conclure, les expérimentations menées dans ce chapitre ont montré qu'il est tout à fait envisageable d'appréhender la problématique de l'analyse d'images de documents à contenu hétérogène autrement que par la mise en place d'un système de reconnaissance de structure.

Conclusion et perspectives

Les bibliothèques numériques soulèvent, depuis dix ans, de nombreuses questions, inquiétudes et attentes. En effet, si les campagnes de numérisations ont permis d'apporter une solution efficace au problème de la conservation du patrimoine, le souhait de rendre disponible ces documents numériques représente encore un enjeu scientifique important. La raison principale est que les outils informatiques permettant un accès généralisé à ces fonds restent encore, pour la plus part, à inventer.

Ce mémoire a permis de mettre en avant les différents axes de recherche actuellement en cours d'élaboration en analyse d'images de documents anciens. Si les travaux relatifs à la conception de logiciels de transcription de ces textes anciens en représentent une grande partie, d'autres travaux soulèvent tout autant d'attentes. En effet, en termes d'usages, l'accès au contenu du texte est loin d'être le seul besoin exprimé par les utilisateurs. L'importance d'une conception de logiciels permettant l'analyse de la structure des documents, la recherche ou la comparaison d'illustrations, l'indexation de la mise en page, montrent bien q'une partie de l'information recherchée par les utilisateurs n'est pas en relation directe avec le contenu textuel. Cette volonté de fournir de tels outils nécessite donc la mise en place de méthodes de traitement d'images permettant la caractérisation des images de documents anciens.

Sur ce problème précis, la littérature propose déjà plusieurs méthodes de caractérisation de contenu. Les tests effectués sur le corpus du Centre d'Etudes Supérieures de la Renaissance de Tours, ont montré toute la difficulté à mettre en place une méthode générique de caractérisation sur la base de ces outils dédiés initialement à l'analyse de documents contemporains (filtres de Gabor, ...). La raison principale de ce constat, est que la nature du corpus traité (contenu fortement hétérogène), ne permet pas de dégager un modèle unique pour l'ensemble des documents traités. Le point principal de notre contribution, consiste donc en la proposition et l'utilisation d'outils de traitements d'images, permettant une caractérisation fine de leur contenu. Quelque soit le type d'images, l'originalité de notre proposition tient tout particulièrement au fait que nous ne cherchons pas à segmenter ou extraire la structure des documents analysés. Ainsi, nous décrivons dans ce mémoire comment il est possible de caractériser le contenu d'images de documents en se basant sur des informations textures non paramétriques. Cette démarche se veut générique et adaptable à tout type d'ouvrage en s'appuyant sur l'homogénéité intraouvrage. En effet, s'il n'est pas possible de dégager un modèle du document ancien ou encore de connaître les caractéristiques physiques des images traitées, il existe néanmoins des similitudes d'une page à l'autre à l'intérieur d'un ouvrage : taille des caractères utilisée pour les titres différents de celle utilisée pour le corps de texte, Alphabet de lettrines, justification du texte, position des marges... La caractérisation du contenu est réalisée à l'aide d'une étude multirésolution des textures contenues dans les images de documents. Ainsi, en extrayant des signatures liées aux fréquences et aux orientations des différentes parties d'une page, ils est possible d'extraire, de comparer ou encore d'identifier des éléments de haut niveau sémantique (lettrines, illustrations, texte, mise en page...) sans émettre d'hypothèses sur la structure physique ou logique des documents analysés mais simplement en les comparant entre eux ou à une image requête. Cette approche très générique, est intéressante puisqu'elle laisse la possibilité d'ajouter des indices supplémentaires correspondant à des caractéristiques non extraites dans la version proposée dans ce manuscrit. Après l'explication du mode de calcul de ces nouveaux indices textures avec une approche multirésolution garantissant la généricité de notre proposition, nous effectuerons une analyse poussée des données générées, de leur complémentarité, de leur redondance, de leur robustesse et de l'influence des paramètres extérieurs.

Cette caractérisation du contenu n'est qu'une première partie de notre contribution. En effet, au travers de plusieurs expériences, nous avons montré qu'il était tout à fait envisageable de réaliser des outils d'aide à l'indexation ou d'aide à la navigation dédiés aux images de documents anciens. A l'aide d'une caractérisation des images et d'une étude de la répartition spatiale de ses différents attributs textures, nous avons montré qu'il était possible de comparer différentes formes de mises en pages (pages sur simple ou double colonnes, pages avec ou sans illustrations, ...). Cette comparaison est réalisée sans segmentation en zones.

Les indices textures proposés permettent également d'entrevoir d'autres formes d'indexation. La réalisation d'expérimentations sur des bases d'images de traits issues de documents anciens, a montré la pertinence de l'utilisation d'informations relatives aux orientations et aux fréquences présentes et qui peuvent être incluses de manière synthétique dans les métadonnées de description d'images. Nous montrons également que nos caractéristiques textures peuvent servir de base pour différentes classification des pixels des images afin d'obtenir les principales zones de texture homogène contenues dans les images de documents (séparation texte, fond, graphiques,...). La qualité des résultats donnés dans le dernier chapitre de ce manuscrit montre qu'il est aussi possible de différencier plusieurs catégories d'éléments de contenu.

Les propositions faites dans ce mémoire de thèse, permettent d'entrevoir de nombreuses perspectives en terme de réalisation d'outils d'aide à la navigation ou d'aide à l'indexation. Les expériences présentées dans ce manuscrit sont avant tout une vitrine illustrant ce qu'il est possible de réaliser. La présentation de ces expériences aux usagers potentiels (essentiellement des personnes des sciences humaines et sociales), permet d'établir un lien entre ce qu'ils désirent et ce que peut leur offrir la communauté informatique.

Dans ce travail de thèse, nous nous sommes focalisés sur l'exploitation d'informations texture pour l'indexation de documents. Si nous avons à plusieurs reprises montré l'intérêt de ce type d'information, il reste maintenant à étudier leur intégration dans des dispositifs d'indexation plus complets (systèmes de CBIR par exemple). La première des perspectives que nous nous fixons est donc de finaliser un système d'indexation capable de produire automatiquement les métadonnées descriptives des images de documents incluant nos indices textures mais aussi d'autres informations (liées aux couleurs, aux formes, à leurs positions,...). Nous pourrons ensuite poursuivre nos recherches sur les mesures de similarité afin de définir de nouvelles manières de comparer les images selon plusieurs critères simultanément.
Mettre en place un système d'apprentissage automatique grâce auquel l'usager montre au système plusieurs exemples de ce qu'il recherche afin que le système sélectionne lui-même les caractéristiques pertinentes à utiliser lors de la comparaison parait être une voie prometteuse. Nous désirons également poursuivre nos travaux, déjà entamés, sur l'exploitation des caractéristiques textures pour simplifier et améliorer la segmentation des images de documents en différents éléments de contenus de haut niveau sémantique. Pour cela, deux voies peuvent être explorées : utiliser les informations textures pour déterminer automatiquement certains paramètres ou seuils actuellement fixés manuellement ; coupler les informations textures avec d'autres types d'informations pour améliorer la segmentation.
Nous prévoyons de réaliser une partie de ces travaux dans le cadre d'une collaboration entre le Centre d'Etudes de la Renaissance de Tours et des travaux de recherche menés au laboratoire d'informatique de Tours, les avancées proposées dans ce mémoire vont permettre d'enrichir la plate-forme de logiciel de traitements d'images de documents anciens nommée AGORA. Cette plate-forme est actuellement utilisée dans le processus de création d'une bibliothèque virtuelle accessible sur internet.

Chapitre 5

Annexes



FIG. *5.1: Extrait des images de la base utilisée pour la comparaison de pages*

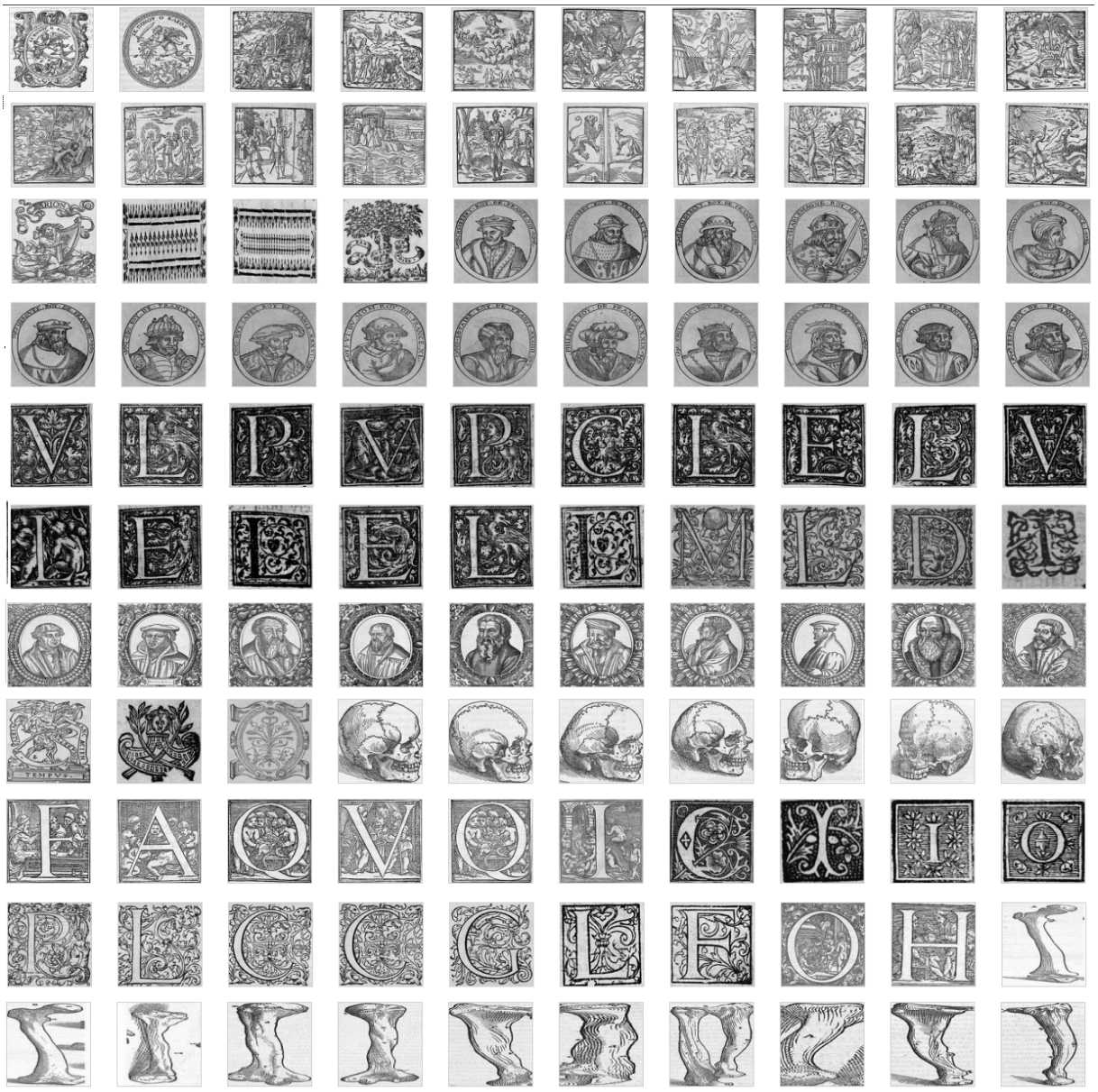


FIG. 5.2: *Extrait des images de la base utilisée pour la comparaison de pages*



FIG. *5.3: Classification à 4 classes sur des images de documents contemporains*

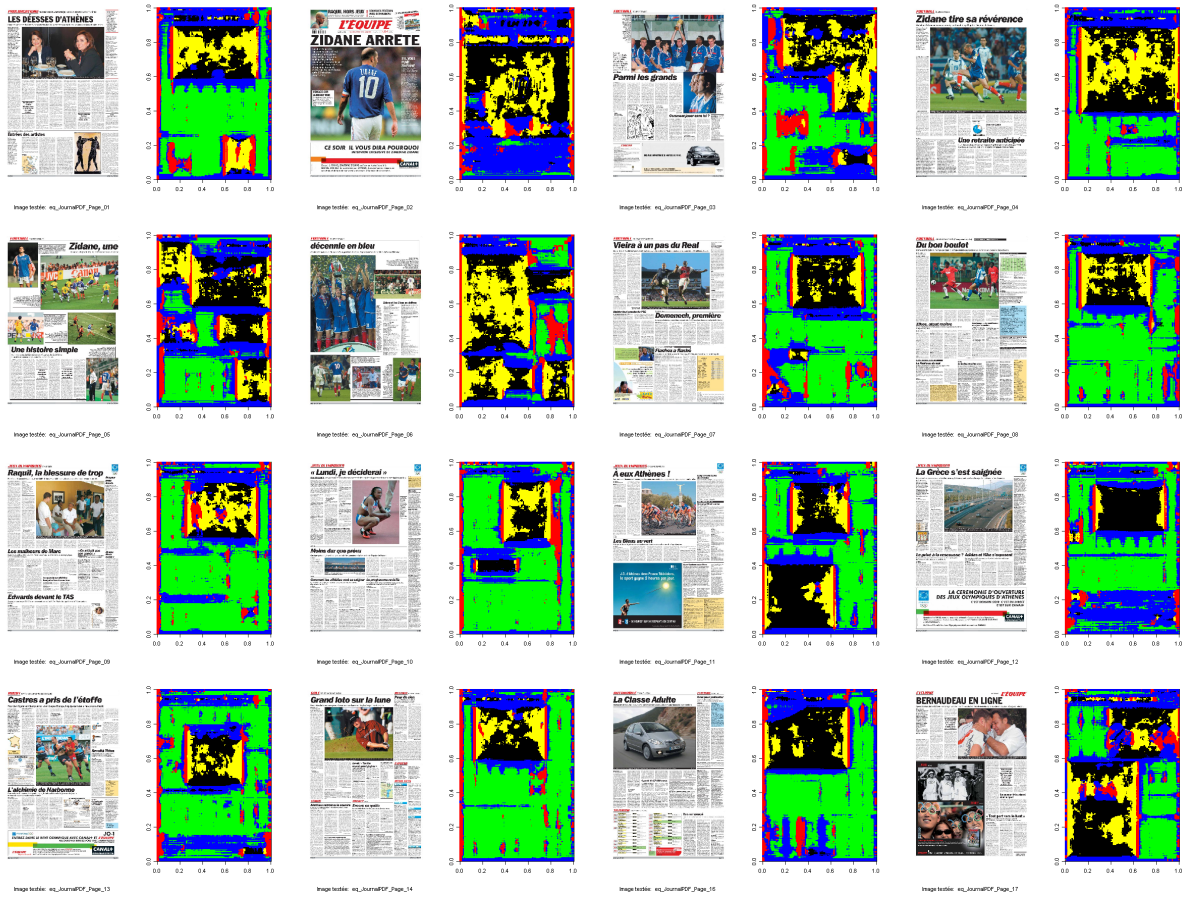


FIG. 5.4: *Classification à 5 classes sur des images de documents contemporains*



FIG. 5.5: *Classification à 3 classes sur des images de documents anciens*

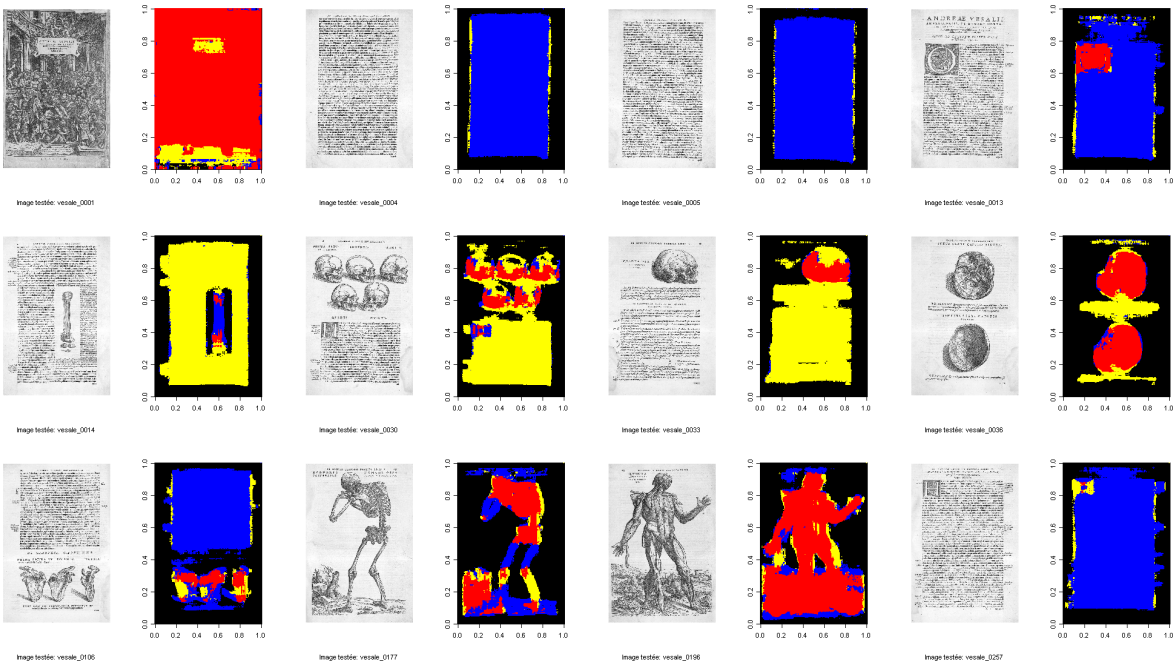


FIG. 5.6: *Classification à 4 classes sur des images de documents anciens*

Bibliographie

[ABCI04] Accary, Bénel, Calabretto, and Iacovella. Confrontation de points de vue sur des corpus documentaires : Le cas de la modélisation du temps archéologique. In *"Actes du 14ème Congrès Francophone AFRIF-AFIA de Reconnaissance des Formes et Intelligence Artificielle"*, pages 197–205, 2004.

[ABG+03] Nicolo G. Alessi, Sebastiano Battiato, Giovanni Gallo, Massimo Mancuso, and Filippo Stanco. Automatic discrimination of text images. volume 5017, pages 351–359. SPIE, 2003.

[AE03] Allier and Emptoz. Font type extraction and character prototyping using gabor filters. *ICDAR*, 02 :799, 2003.

[All04] B Allier. *Contribution à la Numérisation des Collections : Apports des Contours Actifs*. PhD thesis, LIRIS, université de Lyon, 2004.

[AMT02] Marco Aiello, Christof Monz, and Leon Todoran. Document understanding for a broad class of documents. *IJDAR*, 5(1) :1–16, 2002.

[Ant98] Apostolos Antonacopoulos. Page segmentation using the description of the background. *Comput. Vis. Image Underst.*, 70(3) :350–369, 1998.

[AOC+01] ADAM, OGIER, CARIOU, MULLOT, GARDES, and LECOURTIER. Utilisation de la transformée de fourier-mellin pour la reconnaissance de formes multi-orientées et multi-échelles : application à l'analyse automatique de documents techniques. *Traitement du signal*, 18(1) :17–33, 2001.

[AZ03] Tim Andersen and Wei Zhang. Features for neural net based region identification of newspaper documents. In *ICDAR 03 : Proceedings of the Seventh International Conference on Document Analysis and Recognition*, pages 403–407, Washington, DC, USA, 2003. IEEE Computer Society.

[Bag04] Andrew D. Bagdanov. *Style Characterization of Machine Printed Texts*. PhD thesis, Faculteit der Natuurwetenschappen, Wiskunde, Informatica Kruislaan, 2004.

[BBM96] J. Bigun, S. K. Bhattacharjee, and S. Michel. Orientation radiograms for image retrieval. In *International conf. on pattern recognition, ICPR-96*, volume C, pages 346–350. IEEE Computer Society, 1996.

[BC97] Andrea Bozzi and Sylvie Calabretto. The digital library and computational philology : The bambi project. In *ECDL '97 : Proceedings of the First European Conference on Research and Advanced Technology for Digital Libraries*, pages 269–285, London, UK, 1997. Springer-Verlag.

[BELM00] Bouché, Emptoz, Lebourgeois, and Metzger. Debora projet européen. Technical report, LIRIS,université de Lyon, 2000.

[bET+01] Le bourgeois, Emptoz, Trinh, Muge, Pinto, and 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. Technical report, LIRIS,université de Lyon, 2001.

[BFH+04] Bitlis, Feng, Harris, Pollak, Bouman, Harper, and Allebach. A hierarchical document description and comparison method. 2004.

[BGPR02] Barbey, Guillemain, Péoc'h, and Ract. La renaissance du livre ancien : bilan du projet debora et perspectives d'avenir. Technical report, Ecole nationale supérieure des Sciences de l'Information et des Bibliothèques ., 2002.

[BH01] Belisle and Hembise. Usages du livre numérique : Besoins liés aux pratiques et proposition d'un environnement de travail debora. Technical report, Rapport pour la commission européenne, 2001.

[BHL+00] Léon Bottou, Patrick Haffner, Yann LeCun, Paul Howard, Pascal Vincent, and Bill Riemers. Djvu : Un système de compression d'images pour la distribution réticulaire de documents numérisés. *CIFED '2000 : colloque international francophone sur l'écrit et le document*, pages 453–462, 2000.

[BKKS04] Faisal Bashir, Shashank Khanvilkar, Ashfaq Khokhar, and Dan Schonfeld. *Multimedia Systems : Content Based Indexing and Retrieval (book)*. Academic Press, 2004.

[BLEP01] Souad Bensafi, Franck Lebourgeois, Hubert Emptoz, and Marc Parizeau. La relaxation probabiliste pour étiquetage logique des documents : Application aux tables des matières. June 7-9 2001.

[BMN+04] Étienne Baudrier, Gilles Millon, Frédéric Nicolier, Ralph Seulin, and Su Ruan. Hausdorff distance based multiresolution maps applied to an image similarity measure. In *Proc of the The Topical Meeting on Optical Sensing and Artificial Vision (OSAV'2004) (SPIE)*, St Petersbourg, Russie, 18-21 oct 2004.

[BMS03] Stefano Baldi, Simone Marinai, and Giovanni Soda. Using tree-grammars for training set expansion in page classification. In *ICDAR* [DBL03], pages 829–833.

[Bén04] Bénel. Expression du point de vue des lecteurs dans les bibliothèques numériques spécialisées. In *"Actes du Colloque International sur le Document Numérique, "Approches sémantiques sur le document numérique"*, 2004.

[Bor01] Borghi. Tresy - a search engine on xml documents. Technical report, 2001.

[Bre94] Bres. *Contributions à la quantification des critères de transparence et d'anisotropie par une approche globale*. PhD thesis, LIRIS, université de Lyon, 1994.

[Bre02] Thomas M. Breuel. Two geometric algorithms for layout analysis. In *DAS '02 : Proceedings of the 5th International Workshop on Document Analysis Systems V*, pages 188–199, London, UK, 2002. Springer-Verlag.

[Bre03] T. M. Breuel. High performance document layout analysis. In *Symposium on Document Image Understanding Technology (SDIUT '03), Greenbelt, Maryland*, page (accepted for publication), 2003.

[BS78] Bouroche and Saporta. *L'analyse des données Collection Que Sais-Je ?* Presses Universitaires de France, 1978.

[BSN04] Basa, Sabari, and Nishikanta. Gabor filters for document analysis in indian bilingual documents. *Proceedings International Conference on Intelligent Sensing and Information Processing*, pages 123–126, 2004.

[BUR91] Gilles BUREL. *Réseaux de neurones en traitement d'images : Des Modèles Théoriques aux Applications Industrielles*. PhD thesis, Université de Bretagne Occidentale, 1991.

[BUS06] BUSSON. Applications du logiciel d'analyse d'images agora sur les ouvrages imprimés de la renaissance le projet des bibliothèques virtuelles humanistes. Technical report, CESR report - Université François Rabelais de Tours, 2006.

[CC01] W. Chan and G. Coghill. Text analysis using local energy. *Pattern Recognition*, 34(12) :2523–2532, December 2001.

[CCMM98] R. Cattoni, T. Coianiz, S. Messelodi, and C. Moden. Geometric layout analysis techniques for document image understanding : a review. Technical report, IRST, Trento, Italy, 1998.

[CCMV03] Yves Caron, Harold Charpentier, Pascal Makris, and Nicole Vincent. Power law dependencies to detect regions of interest. *Lecture Notes in Computer Science*, 2886/2003 :495–503, November 2003.

[CLKH96] D. Chetverikov, J. Liang, J. Komuves, and R. M. Haralick. Zone classification using texture features. In *ICPR '96 : Proceedings of the International Conference on Pattern Recognition (ICPR '96) Volume III-Volume 7276*, page 676, Washington, DC, USA, 1996. IEEE Computer Society.

[CLM98] L. Cinque, L. Lombardi, and G. Manzini. A multiresolution approach for page segmentation. *Pattern Recogn. Lett.*, 19(2) :217–225, 1998.

[CM91a] CHASSERY and MELKEMI. Diagramme de voronoï appliqué à la segmentation d'images et à la détection d'évènements en imageris multi-sources. *Traitement du Signal*, 8(3) :155–164, 1991.

[CM91b] J.M. Chassery and A. Montanvert. *GEOMETRIE DISCRETE en analyse d'images*. HERMES, 1991.

[Com98] Comeau. Encoded archival description (ead) et la création d'instruments de recherche électronique. *Flash Réseau (spécial TEI)*, Volume 58, 1998.

[CR03] Couasnon and Rapp. Accès par le contenu aux documents manuscrits d'archives numérisées. *Document numérique*, 7 :61–84, 2003.

[CV00] Cooper and Venters. A review of content-based image retrieval systems. *JISC Technology Applications Programme*, 2000.

[CWS03] Zheru Chi, Qing Wang, and Wan-Chi Siu. Hierarchical content classification and script determination for automatic document image processing. *Pattern Recognition*, 36(11) :2483–2500, 2003.

[DA02] Duygulu and Atalay. A hierarchical representation of form documents for identification and retrieval. *International Journal on Document Analysis and Recognition*, 5(1) :17–27, 2002.

[DBL03] *7th International Conference on Document Analysis and Recognition (ICDAR 2003), 2-Volume Set, 3-6 August 2003, Edinburgh, Scotland, UK*. IEEE Computer Society, 2003.

[DCES01] Jean Duong, Myriam C&#244 ;te, Hubert Emptoz, and Ching Y. Suen. Extraction of text areas in printed document images. In *DocEng '01 : Proceedings of the 2001 ACM Symposium on Document engineering*, pages 157–165, New York, NY, USA, 2001. ACM Press.

[DeB04] DeBoisdeffre. Culture & recherche. *Journal du ministère de la culture et de la communication*, 103, 2004.

[Doe98a] David Doermann. The indexing and retrieval of document images : A survey. *Computer Vision and Image Understanding : CVIU*, 70(3) :287–298, 1998.

[Doe98b] David Doermann. The indexing and retrieval of document images : a survey. *Comput. Vis. Image Underst.*, 70(3) :287–298, 1998.

[EDC97] Kamran Etemad, David Doermann, and Rama Chellappa. Multiscale segmentation of unstructured document pages using soft decision integration. *IEEE Trans. Pattern Anal. Mach. Intell.*, 19(1) :92–96, 1997.

[Egl98] V Eglin. *Contribution à la structuration fonctionnelle des documents imprimés.* PhD thesis, LIRIS, 1998.

[ELEL03] Emptoz, Lebourgeois, Eglin, and Leydier. La reconnaissance dans les images numérisées, ocr et transcription, reconnaissance des structures fonctionnelles et des méta-données. *Dans : La numérisation des textes et des images : techniques et réalisations*, pages 105–129, 2003.

[FB05] Farhat and Bénel. Du virtuel au manipulable : Interfaces homme/machine pour l'appropriation des documents dans les bibliothèques numériques. In *actes de la Conférence Hypertexte et Hypermédia, "Créer, jouer, échanger : Expériences de réseaux"*, pages 295–306. Hermès - Paris, 2005.

[FHA90] Fisher, Hinds, and Amato. A rule-based system for document image segmentation. *Pattern Recognition*, 1 :567–572, 1990.

[FSN+95] Myron Flickner, Harpreet Sawhney, Wayne Niblack, Jonathan Ashley, Qian Huang, Byron Dom, Monika Gorkani, Jim Hafner, Denis Lee, Dragutin Petkovic, David Steele, and Peter Yanker. Query by image and video content : The qbic system. *Computer*, 28(9) :23–32, 1995.

[FWT98] Kuo-Chin Fan, Liang-Shen Wang, and Yin-Tien Tu. Classification of machine-printed and handwritten texts using character block layout variance. *Pattern Recognition*, 31(9) :1275–1284, 1998.

[GJ96] G.L. Gimel'Farb and A.K. Jain. On retrieving textured images from an image database. *Pattern Recognition*, 29(9) :1461–1483, September 1996.

[GVCJ00] P. Gupta, N. Vohra, S. Chaudhury, and S. Joshi. Wavelet based page segmentation. *ICVGIP*, pages 51–56, 2000.

[Haf05] Adel Hafiane. *Caractérisation de textures et segmentation pour la recherche d'images par le contenu.* PhD thesis, Université de PARUS-SUD XI - faculté des sciences d'Orsay, 2005.

[HB00] Mryka Hall-Beyer. Glcm texture : A tutorial, 2000.

[HBS05] Hatem HAMZA, Abdel BELAID, and Eddie SMIGIEL. Neural based binarization techniques. *icdar*, 0 :317–321, 2005.

[HC97] Jonathan J. Hull and John Cullen. Document image similarity and equivalence detection. In *ICDAR '97 : Proceedings of the 4th International Conference on Document Analysis and Recognition*, volume 00, pages 308–313, Washington, DC, USA, 1997. IEEE Computer Society.

[HDD+05] May Huang, Daniel DeMenthon, David Doermann, Lynn Golebiowski, and Booz Allen Hamilton. Document ranking by layout relevance. *icdar*, 0 :362–366, 2005.

[HDDK05] J. He, Q. D. M. Do, A. C. Downton, and J. H. Kim. A comparison of binarization methods for historical archive documents. *icdar*, 0 :538–542, 2005.

[HER01] HEROUX. *contribution au problème de la rétro-conversion des documents structurés.* PhD thesis, UNIVERSITE DE Rouen, 2001.

[HI03] Karim Hadjar and Rolf Ingold. Arabic newspaper page segmentation. *icdar*, 02 :895–900, 2003.

[HKW99] Jianying Hu, Ramanujan Kashi, and Gordon Wilfong. Document classification using layout analysis. In *DEXA '99 : Proceedings of the 10th International Workshop on Database & Expert Systems Applications*, page 556, Washington, DC, USA, 1999. IEEE Computer Society.

[HOP+95] David Harwood, Timo Ojala, Matti Pietik, Shalom Kelman, and Larry Davis. Texture classification by center-symmetric auto-correlation, using kullback discrimination of distributions. *Pattern Recogn. Lett.*, 16(1) :1–10, 1995.

[HSD73] R.M. Haralick, K. Shanmugam, and I. Dinstein. Textural features for image classification. *SMC*, 3(6) :610–621, November 1973.

[IBCH05] Iacovella, Bénel, Calabretto, and Helly. Assistance à l'interprétation dans les bibliothèques numériques pour les sciences historiques. In *Actes du colloque de bilan du programme interdisciplinaire "Société de l'information"*, pages 167–179. J.-L. Lebrave (Ed.), 2005.

[IK00] L. Itti and C. Koch. A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research*, 40(10-12) :1489–1506, May 2000.

[IKK05] Ilonen, J. Kamarainen, and J.-K. Kälviäinen. Efficient computation of gabor features. Technical report, Lappeenranta University of Technology, Department of Information Technology, 2005.

[IV96] IDE and VÉRONIS. Présentation de la tei :text encoding initiative. *Cahiers GUTenberg (spécial TEI*, 24 :4–10, 1996.

[IW98] INGLIS and WITTEN. Document zone classification using machine learning. 1998.

[Jea05] Jeanneney. Quand google défie l'europe, plaidoyer pour un sursaut. edition mille et une nuit - paris - 150 pages. pages 105–129, 2005.

[JKJ04] K. Jung, K.I. Kim, and A.K. Jain. Text information extraction in images and video : a survey. *Pattern Recognition*, 37(5) :977–997, May 2004.

[JMER06] N. Journet, R. Mullot, V. Eglin, and J.Y Ramel. Dedicated texture based tools for characterisation of old books. *dial*, 0 :60–69, 2006.

[JMRE05] Nicholas Journet, Rémy Mullot, Jean-Yves Ramel, and Véronique Eglin. Ancient printed documents indexation : A new approach. In Singh et al. [SSAP05], pages 580–589.

[Jou04] Jourdy. Culture & recherche. Technical report, Journal du ministère de la culture et de la communication. Volume : 100, 2004.

[Jou06] N. Journet. *Analyse d'images de documents anciens : une approche texture.* PhD thesis, L3I, université de La Rochelle, 2006.

[JS95] J.Cocquerez and S.Philipp. *Analyse d'images :Filtrage et segmentation.* Masson, 1995.

[Kal00] Kalldremxhiu. Les logiciels de numérisation des livres anciens. Technical report, Université Claude Bernard -Lyon1., 2000.

[Kau99] Hannu Kauniskangas. Document image retrieval with improvements in database quality. Technical report, 1999.

[KDK00] Stefan Klink, Andreas Dengel, and Thomas Kieninger. Document structure analysis based on layout and textual features. In *Proceedings of the 4th IAPR International Workshop on Document Analysis Systems (DAS 2000)*, pages 99–111, Rio de Janeiro, Brazil, December 2000.

[KIM99] K. Kise, M. Iwata, and K. Matsumoto. On the application of voronoi diagrams to page segmentation. *Proc. of the Workshop on Document Layout Interpretation and Its Applications*, (IV-C) :1–4, September 1999.

[KKS04] Soo-Hyung Kim, Hee K. Kwag, and Ching Y. Suen. Word-level optical font recognition using typographical features. *IJPRAI*, 18(4) :541–561, 2004.

[KR90] Kaufman and Rousseeuw. *Finding Groups in Data.* John Wiley Sons New York, Chichester, Brisbane : book, 1990.

[KRSG03] Swapnil Khedekar, Vemulapati Ramanaprasad, Srirangaraj Setlur, and Venugopal Govindaraju. Text - image separation in devanagari documents. In *ICDAR '03 : Proceedings of the Seventh International Conference on Document Analysis and Recognition*, volume 2, page 1265, Washington, DC, USA, 2003. IEEE Computer Society.

[KS92] F. Kimura and M. Shridhar. Segmentation-recognition algorithm for zip code field recognition. *Mach. Vision Appl.*, 5(3) :199–210, 1992.

[KS06] E. Kavallieratou and E. Stamatatos. Improving the quality of degraded document images. *DIAL06*, 0 :340–349, 2006.

[KSI98] Koichi Kise, Akinori Sato, and Motoi Iwata. Segmentation of page images using the area voronoi diagram. *Comput. Vis. Image Underst.*, 70(3) :370–382, 1998.

[KSM97] Koichi Kise, Akinori Sato, and Keinosuke Matsumoto. Document image segmentation as selection of voronoi edges. *DIA '97 : Proceedings of the 1997 Workshop on Document Image Analysis*, 00 :32, 1997.

[KSP97] M. Koivusaari, J. Sauvola, and M. Pietikaeinen. Automated document content characterization for a multimedia document retrieval system. In C.-C. J. Kuo, S. F. Chang, and V. N. Gudivada, editors, *Proc. SPIE Vol. 3229, p. 148-159, Multimedia Storage and Archiving Systems II, C.-C. J. Kuo ; Shih Fu Chang ; Venkat N. Gudivada ; Eds.*, pages 148–159, October 1997.

[KYT96] K. Kise, O. Yanagida, and S. Takamatsu. Page segmentation based on thinning of background. *icpr*, 03 :788, 1996.

[Law80] K. I. Laws. Rapid texture identification. In *Image processing for missile guidance ; Proceedings of the Seminar, San Diego, CA, July 29-August 1, 1980. (A81-39326 18-04) Bellingham, WA, Society of Photo-Optical Instrumentation Engineers, 1980, p. 376-380.*, pages 376–380, 1980.

[LBJ99] E. Loupias, S. Bres, and J. M. Jolion. From methods to images. *Computer Science*, 62(4) :265–275, 1999.

[LC06] Sylvie Lainé-Cruzel. Appropriation, mutualisation, expérimentations des technologies de l'information scientifique et technique. *Ametist*, 0, 2006.

[LET03] Frank Lebourgeois, Hubert Emptoz, and Eric Trinh. Compression et accessibilité aux images de documents numérisés : application au projet debora. *Document Numérique*, 7(3-4) :103–125, 2003.

[LG00] J. Li and R.M. Gray. Context-based multiscale classification of document images using wavelet coefficient distributions. 9(9) :1604–1616, September 2000.

[LKPJ06] Carl Lagozei, Dean B. Kraffti, Sandy Payettei, and Susan Jesurogaii. Qu'est-ce qu'une bibliothèque numérique, au juste ? *Ametist*, 0, 2006.

[LLC93] Lethelier, Leroux, and Couthouis. Automatic processing of numerical amounts on postal cheques. *Proc. of 1st European Conf on Postal Technologies (JetPoste '93)*, 2 :697–704, 1993.

[Lou00] Etienne Loupias. *Indexation d'images : aide au télé-enseignement et similarités préattentives.* PhD thesis, LIRIS, 2000.

[LWT04] Yue Lu, Zhe Wang, and Chew Lim Tan. Word grouping in document images based on voronoi tessellation. *Lecture Notes in Computer Science*, 3163 :147 – 157, 2004.

[MD05] H. Ma and D. S. Doermann. Font identification using the grating cell texture operator. 5676 :148–156, 2005.

[MEA02] Donato Malerba, Floriana Esposito, and Oronzo Altamura. Adaptive layout analysis of document images. In *ISMIS '02 : Proceedings of the 13th International Symposium on Foundations of Intelligent Systems*, volume 2366, pages 526–534, London, UK, 2002. Springer-Verlag.

[Mic00] Michard. *Finding Groups in Data.* Eyrolles, 2000.

[MM85] W.Y. Ma and B.S. Manjunath. Image indexing using a texture dictionary. (260) : :288–296, October 1985.

[MM96a] W. Y. Ma and B. S. Manjunath. Texture features and learning similarity. *cvpr*, 00 :425, 1996.

[MM96b] B. S. Manjunath and W. Y. Ma. Texture features for browsing and retrieval of image data. *IEEE Trans. Pattern Anal. Mach. Intell.*, 18(8) :837–842, 1996.

[MM99] Wei-Ying Ma and B. S. Manjunath. Netra : A toolbox for navigating large image databases. *Multimedia Systems*, 7(3) :184–198, 1999.

[MMS05] Simone Marinai, Emanuele Marino, and Giovanni Soda. Layout based document image retrieval by means of xy tree reduction. 1 :432–436, 2005.

[MMS06] Simone Marinai, Emanuele Marino, and Giovanni Soda. Tree clustering for layout-based document image retrieval. In *DIAL '06 : Proceedings of the Second International Conference on Document Image Analysis for Libraries (DIAL'06)*, pages 243–253, Washington, DC, USA, 2006. IEEE Computer Society.

[MRK03] Mao, Rosenfeld, and Kanungo. Document structure analysis algorithms : A literature survey. *SPIE*, 5010 :197–207, 2003.

[MSB97a] Gerd Maderlechner, Peter Suda, and Thomas Brückner. Classification of documents by form and content. *Pattern Recogn. Lett.*, 18(11-13) :1225–1231, 1997.

[MSB97b] Gerd Maderlechner, Peter Suda, and Thomas Bruckner. Classification of documents by form and content. Siemens AG, Corporate Research and DeÍelopment, Otto-Hahn-Ring 6, D-81730 Munchen, Germany, 1997.

[MSM02] P. Musé, F. Sur, and J.-M. Morel. Recherche dans les grandes bases de formes. Technical report, CMLA report, 2002.

[MY01] Phillip E Mitchell and Hong Yan. Newspaper document analysis featuring connected line segmentation. *icdar*, 00 :1181, 2001.

[Nag00] George Nagy. Twenty years of document image analysis in pami. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(1) :38–62, 2000.

[NKK+88] George Nagy, Junichi Kanai, Mukkai Krishnamoorthy, Mathews Thomas, and Mahesh Viswanathan. Two complementary techniques for digitized document analysis. In *DOCPROCS '88 : Proceedings of the ACM conference on Document processing systems*, pages 169–176, New York, NY, USA, 1988. ACM Press.

[NKPH06] Nicolas, Kessentini, Paquet, and Heutte. Handwritten document segmentation using hidden markov random fields. *ICDAR*, 1 :212–216, August 2006.

[NS96] D. Niyogi and S. N. Srihari. Using domain knowledge to derive the logical structure of documents. In L. M. Vincent and J. J. Hull, editors, *Proc. SPIE Vol. 2660, p. 114-125, Document Recognition III, Luc M. Vincent; Jonathan J. Hull; Eds.*, pages 114–125, March 1996.

[O'G93] L. O'Gorman. The document spectrum for page layout analysis. *PAMI*, 15(11) :1162–1173, November 1993.

[OP00] Okun and Pietikäinen. A survey of texture-based methods for document layout analysis. *Texture Analysis in Machine Vision*, 40 :165–177, 2000.

[OR03] Oulahal and Raphaël. Accès unique à des ressources numériques distribuées. *JRES*, 2003.

[PA02] CORNU Philippe and SMOLARZ André. Caractérisation d'images par textures associées. *Traitement du signal (Trait. signal)*, 19(1) :29–35, 2002.

[PC01] U. Pal and B. B. Chaudhuri. Machine-printed and hand-written text lines identification. *Pattern Recognition Letters*, 22(3/4) :431–441, 2001.

[Pet02] E. Petrakis. Design and evaluation of spatial similarity approaches for image retrieval. *Image and Vision Comp.*, 20(1) :59–76, 2002.

[PLCS01] H. Peng, F. Long, Z. Chi, and W.C. Siu. Document image template matching based on component block list. 22(9) :1033–1042, July 2001.

[PM05] Pigeard-Micault. Les sciences et l'histoire des sciences dans gallica. *HEP Libraries Webzine*, 11, 2005.

[PPS96] A. Pentland, R. W. Picard, and S. Sclaroff. Photobook : content-based manipulation of image databases. *Int. J. Comput. Vision*, 18(3) :233–254, 1996.

[Pra78] W.K. Pratt. *Digital Image Processing (Book : First Edition)*. Wiley, 1978.

[PRTB99] Jan Puzicha, Yossi Rubner, Carlo Tomasi, and Joachim M. Buhmann. Empirical evaluation of dissimilarity measures for color and texture. In *ICCV (2)*, pages 1165–1172, 1999.

[PVU+06] Rudolf Pareti, Nicole Vincent, Surapong Uttama, Jean-Marc Ogier, Jean-Pierre Salmon, Salvatore Tabbone, Laurent Wendling, and Sebastien Adam. On defining signatures for the retrieval and the classification of graphical drop caps. *dial*, 0 :220–231, 2006.

[PZ91] Pavlidis and Zhou. Page segmentation by white streams. *icdar*, 2 :945–953, 1991.

[RBD06] J.Y. Ramel, S. Busson, and M.L. Demonet. Agora : the interactive document image analysis tool of the bvh project. *dial*, 0 :145–155, 2006.

[ROB01a] ROBADEY. *2(CREM) Une méthode de reconnaissance structurelle de documents complexes basée sur des patterns bidimensionnels.* PhD thesis, UNIVERSITE DE Fribourg, 2001.

[Rob01b] L Robadey. *2(crem) : Une méthode de reconnaissance structurelle de documents complexes basée sur des patterns bidimensionnels.* PhD thesis, Institut d'Informatique de l'Université de Fribourg, (Suisse), 2001.

[Ros99] C Rosenberg. *Mise en oeuvre d'un système adaptatif de segmentation d'images.* PhD thesis, Laboratoire d'analyse des systèmes de traitement de l'information, ENSSAT, 1999.

[RP05] N. Vincent R. Pareti. Global discrimination of graphics styles. *CVPR '96 : Proceedings of the 1996 Conference on Computer Vision and Pattern Recognition (CVPR '96)*, pages 120–128, 2005.

[RPR05] S.S. Raju, P.B. Pati, and A.G. Ramakrishnan. Text localization and extraction from complex color images. *ISVC05*, pages 486–493, 2005.

[Sal05] Salaün. Bibliothèques numériques et google-print. *Regard sur l'actualité. La documentation Française - Montréal*, 2005.

[SBK05] GUILLAS S and OGIER J BERTET K. Les treillis de galois : un outil pour la sélection de primitives ? *Traitement du Signal*, 22(3) :273–291, 2005.

[SD99a] C. Shin and D. Doermann. Classification of document page images. *Document Image Understanding Technology*, pages 166–175, 1999.

[SD99b] C. K. Shin and D. S. Doermann. Classification of document page images based on visual similarity of layout structures. In D. P. Lopresti and J. Zhou, editors, *Proc. SPIE Vol. 3967, p. 182-190, Document Recognition and Retrieval VII, Daniel P. Lopresti ; Jiangying Zhou ; Eds.*, pages 182–190, December 1999.

[SG04] Shi and Govindaraju. Dynamic local connectivity and its application to page segmentation. *Proceedings of the 1st ACM workshop on Hardcopy document processing*, 2004.

[SG05] Zhixin Shi and Venu Govindaraju. Multi-scale techniques for document page segmentation. *icdar*, 0 :1020–1024, 2005.

[SK05] Vassilis Athitsos Jonathan Alon Stan Sclaroff and George Kollios. Filtering methods for similarity-based multimedia retrieval. *Seventh International Workshop of the EU Network of Excellence DELOS on Audio-Visual Content and Information Visualization in Digital Libraries (AVIVDiLib)*, 2005.

[SKB06] Faisal Shafait, Daniel Keysers, and Thomas M. Breuel. Performance comparison of six algorithms for page segmentation. 3872 :368–379, Feb 2006.

[SLYcC02] Von-Wun Soo, Chen-Yu Lee, Jaw Jium Yeh, and Ching chih Chen. Using sharable ontology to retrieve historical images. In *JCDL '02 : Proceedings of the 2nd ACM/IEEE-CS joint conference on Digital libraries*, pages 197–198, New York, NY, USA, 2002. ACM Press.

[SS01] Chaman L. Sabharwal and S. R. Subramanya. Indexing image databases using wavelet and discrete fourier transform. In *SAC '01 : Proceedings of the 2001 ACM symposium on Applied computing*, pages 434–439, New York, NY, USA, 2001. ACM Press.

[SSAP05] Sameer Singh, Maneesha Singh, Chidanand Apté, and Petra Perner, editors. *Pattern Recognition and Data Mining, Third International Conference on Advances in Pattern Recognition, ICAPR 2005, Bath, UK, August 22-25, 2005, Proceedings, Part I*, volume 3686 of *Lecture Notes in Computer Science*. Springer, 2005.

[TCL+99] Y. Tang, M. Cheriet, J. Liu, J. Said, and C. Suen. Document analysis and recognition by computers. *Handbook of Pattern Recognition and Computer Vision, World Scientific Publishing Company (1999) 2nd Edition, World Scientific Publishing*, pages 679–712, 1999.

[TFMB04] A. Trémeau, C. Fernandez-Maloigne, and P. Bonton. *Image numérique couleur. De l'acquisition au traitement.* Dunod, 2004. Parmi les auteurs et contributeurs : S. Philipp-Foliguet, M. Cord.

[TJ90] M. Tuceryan and Anil K. Jain. Texture segmentation using voronoi polygons. *IEEE Trans. Pattern Anal. Mach. Intell.*, 12(2) :211–216, 1990.

[TJ98] M. Tuceryan and A. K. Jain. Texture analysis. *In The Handbook of Pattern Recognition and Computer Vision (2nd Edition)*, pages 207–248, 1998.

[TLS96] Yuan Yan Tang, Seong-Whan Lee, and Ching Y. Suen. Automatic document processing : A survey. *Pattern Recognition*, 29(12) :1931–1952, 1996.

[Tri03] Trinh. *De la numérisation à la consultation de documents anciens.* PhD thesis, Université De Lyon, 2003.

[Tru05] Trupin. La reconnaissance d'images de documents : Un panorama. *Traitement du Signal*, 22(3) :159–1892, 2005.

[Tuc94] M. Tuceryan. Moment-based texture segmentation. *PRL*, 15(7) :659–668, July 1994.

[TZ00] Chew Lim Tan and Zheng Zhang. Text block segmentation using pyramid structure. *Document Recognition and Retrieval VIII*, 4307(1) :297–306, 2000.

[UOL05] Uttama, J Ogier, and P Loonis. Top-down segmentation of ancient graphical drop caps. *GREC*, pages 87–95, 2005.

[Var04] Varshney. Block-segmentation and classification of grayscale postal images. *Report in school of electrical and computer Engineering*, 2004.

[WCW82] WONG, CASEY, and WAHL. Document analysis system, ibm journal of research and development. *IBM Journal of Research and Development*, 26 :647–656, 1982.

[WMR97] Victor Wu, R. Manmatha, and Edward M. Riseman. Finding text in images. pages 3–12, 1997.

[WS89] Dacheng Wang and Sargur N. Srihari. Classification of newspaper image blocks using texture analysis. *Computer Vision, Graphics, and Image Processing*, 47(3) :327–352, 1989.

[XHW02] J. Xi, J. Hu, and L. Wu. Page segmentation of chinese newspapers. 35(12) :2695–2704, December 2002.

[Yac96] El Yacoubi. *Modélisation markovienne de l'écriture manuscrite. Application à la reconnaissance des adresses postales.* PhD thesis, UNIVERSITE DE RENNES, 1996.

[YS04] Youness and Saporta. Une méthodologie pour la comparaison de partitions. *Revue de Statistique Appliquée*, 52 :97–120, 2004.

[YUA01] Text extraction from gray scale document images using edge information. In *ICDAR '01 : Proceedings of the Sixth International Conference on Document Analysis and Recognition*, page 302, Washington, DC, USA, 2001. IEEE Computer Society.

[ZTB04] Nicolas Zlatoff, Bruno Tellez, and Atilla Baskurt. Image understanding using domain knowledge. In *RIAO (Recherche d'information Assistée par Ordinateur)*, pages 277–290, Avignon, France, 2004.

[ZTW01] Y. Zhu, T. Tan, and Y. Wang. Font recognition based on global texture analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(10) :1192–1200, 2001.

Résumé

Mes travaux de thèse sont liés à la problématique de l'indexation de grosses quantités d'images de documents anciens. Ils visent à traiter un corpus de documents dont le contenu (texte et image) ainsi que la mise en page se trouvent être fortement variables (l'ensemble des ouvrages d'une bibliothèque par exemple). Cette grande hétérogénéité implique qu'il est impossible de dégager un modèle de l'ensemble des documents traités. Ainsi, il n'est donc pas envisageable de travailler sur ce corpus tel qu'on le fait habituellement avec les documents contemporains. En effet, les premiers tests que nous avons effectués sur le corpus du Centre d'Etudes Supérieures de la Renaissance de Tours, avec qui nous travaillons, ont confirmé que les approches classiques (guidée par le modèle) sont peu efficaces car il est impossible d'émettre des hypothèses sur la structure physique ou logique des documents traités.. En parallèle de ces faiblesses techniques, il faut aussi souligner le manque d'outils d'aide à l'indexation de documents anciens disponibles. Si l'ensemble des travaux réalisés à ce jour se concentre essentiellement sur la reconnaissance de caractères (techniques qui se heurtent à un problème de robustesse sur les documents anciens), il paraît important de proposer d'autres types d'outils pour effectuer des requêtes sur des grandes bases de données documentaires, véritable pivot dans la constitution et l'utilisation de bibliothèques numériques. Face à ce double constat, ces travaux de thèse nous ont amené à proposer une méthode de caractérisation du contenu d'images. Cette caractérisation est réalisée à l'aide d'une étude multirésolution des textures contenues dans les images de document. Ainsi, en extrayant des signatures liées aux fréquences et aux orientations des différentes parties d'une page, il est possible d'extraire, de comparer ou encore d'identifier des éléments de haut niveau sémantique (lettrines, illustrations, texte, mise en page...) sans émettre d'hypothèses sur la structure physique ou logique des documents analysés. Ces indices textures ont permis la création d'outils d'aide à l'indexation de grosses bases d'images de documents anciens. Des expérimentations montrent la faisabilité de la réalisation d'outils d'aide à la navigation ou d'aide à l'indexation. Au travers de ces expérimentations, nous mettons en avant la pertinence de ces indices et les avancées qu'ils représentent en terme de caractérisation de contenu d'un corpus fortement hétérogène. L'ensemble de ces expérimentations permet de manipuler les images de documents anciens par requête " image " sans connaissance du modèle, ce qui est l'objectif de ce travail.

Mots-clés:

Analyse d'images de documents, Texture, Indexation, Bibliothèques numériques, Aide à la navigation, Recherche d'informations

Abstract

My phd thesis subject is related to the topic of old documents images indexation. The corpus of old documents has specific characteristics. The content (text and image) as well as the layout information are strongly variable. Thus, it is not possible to work on this corpus such as it usually done with contemporary documents. Indeed, the first tests which we realised on the corpus of the "Centre d'Etude de la Renaissance", with which we work, confirmed that the

traditional approaches (driven-model approaches) are not very efficient because it's impossible to put assumptions on the physical or logical structure of the old documents. We also noted the lack of tools allowing the indexing of large old documents images databases. In this phd work, we propose a new generic method which permits characterization of the contents of old documents images. This characterization is carried out using a multirésolution study of the textures contained in the images of documents. By constructing signatures related with the frequencies and the orientations of the various parts of a page it is possible to extract, compare or to identify different kind of semantic elements (reference letters, illustrations, text, layout...) without making any assumptions about the physical or logical structure of the analyzed documents. These textures information are at the origin of creation of indexing tools for large databases of old documents images.

Keywords:

Document image analysis, Texture, Indexation, Digital libraries, navigation, information retrieval