Analyse d'Images de Documents Anciens: une Approche Texture

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

Analyse d'Images de Documents Anciens : une Approche Texture

Old Document Image Analysis : A Texture Approach

Nicholas Journet [1], Jean-Yves Ramel [1], Véronique Eglin [2], Rémy Mullot [3]

[1]Laboratoire d'Informatique, 64 avenue Jean Portalis, 37200 Tours
njournet(jean-yves.ramel)@univ-tours.fr
[2]LIRIS-UMR 5205, INSA de Lyon, Bâtiment Jules Verne, 69621 Villeurbanne cedex
veronique.eglin.insa-lyon.fr
[3] L3I, Université de La Rochelle, Pôle Sciences et Technologie, 17042 La Rochelle cedex 1
rmullot@univ-lr.fr

Manuscrit reçu le

Résumé et mots clés

Dans cet article, nous proposons une méthode de caractérisation d'images d'ouvrages anciens basée sur une approche texture. Cette caractérisation est réalisée à l'aide d'une étude multirésolution des textures contenues dans les images de documents. Ainsi, en extrayant cinq indices liés aux fréquences et aux orientations dans les différentes parties d'une page, il est possible d'extraire et de comparer des éléments de haut niveau sémantique sans émettre d'hypothèses sur la structure physique ou logique des documents analysés. 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 mettrons en avant la pertinence de ces indices et les avancées qu'ils représentent en terme de caractérisation de contenu d'un corpus fortement hétérogène.

Analyse d'images de documents, indices texture, multirésolution, indexation, bibliothèque numérique.

Abstract and key words

In this article, we propose a method of characterization of images of old documents based on a texture approach. This characterization is carried out with the help of a multi-resolution study of the textures contained in the images of the document. Thus, by extracting five features linked to the frequencies and to the orientations in the different areas of a page, it is possible to extract and compare elements of high semantic level without expressing any hypothesis about the physical or logical structure of the analysed documents. Experimentations demonstrate the performance of our propositions and the advances that they represent in terms of characterization of content of a deeply heterogeneous corpus.

Document image analysis, Texture features, Multiresolution, digital libraries, indexation.

1. Introduction

La numérisation massive de milliers de pages de documents, mais surtout le désir de les diffuser, 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 (optical character recognition) 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, le cadre nouveau que représente l'analyse d'images de documents anciens ne permet pas la simple transposition de ces outils initialement dédiés aux documents contemporains. L'explication se trouve principalement dans la nature même des corpus d'images traitées. En effet, l'hétérogénéité des pages composant ce corpus d'images de documents anciens, 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 témoigne également de la nécessité d'une nouvelle réflexion autour de la création de nouveaux outils de traitements d'images dédiés aux documents anciens.

Cet article répond à une problématique fondamentale qu'est la caractérisation de contenu d'images de documents anciens vue comme une alternative aux méthodes d'analyses de documents contemporains qui jusqu'alors ont principalement été basées sur une segmentation des pages et une interprétation de leur structure.

Cet article s'articule autour de trois parties :

Dans un premier temps nous proposons une étude des méthodes de caractérisation d'images de documents. Nous verrons notamment que certains outils (initialement dédiés aux documents contemporains) sont inappropriés sur des documents dont les contenus sont à la fois riches et irréguliers.

La deuxième partie de cet article, détaille notre proposition de caractérisation de contenu d'images de documents anciens. À l'aide du calcul d'indices de textures à différentes résolutions, nous montrons qu'il est possible de caractériser le contenu des images sans émettre d'hypothèses, ni sur la structure ni sur les caractéristiques des images traitées.

Dans la dernière partie, nous montrons comment la caractérisation de textures peut être exploitée à des fins d'indexation par le contenu.

2. Les méthodes de caractérisation de contenu d'images de documents

Dans [Jou06], nous avons réalisé une synthèse sur le fonctionnement et les usages relatifs aux bibliothèques numériques actuellement en ligne. Ce tour d'horizon a permis d'établir que l'indexation des ouvrages a, dans la plupart des cas, été réalisée manuellement. Pour chaque nouvel ouvrage mis en ligne, une personne a associé à ce dernier, des mots-clefs permettant à un moteur de recherche de référencer chaque ouvrage.

Si une bibliothèque numérique souhaite donner accès à l'information contenue dans un ouvrage (recherche sur le texte, rechercher des illustrations...), une indexation manuelle consistant à associer des mots-clefs à chaque page (ou portion de page) n'est pas réalisable. La masse de travail que représenterait la saisie de mots-clefs est une limite évidente. L'autre difficulté est liée au choix des mots-clefs à employer pour décrire des informations non explicites. En effet, comment associer les mots-clefs judicieux pour décrire une illustration ou un élément de structure ?

Des travaux d'indexation automatique par analyse de contenu ont donc été mis en place dans l'optique de pouvoir indexer rapidement et précisément un grand nombre d'images de documents anciens. Parmi ceux-ci, on peut citer les OCR dédiés aux documents anciens [BELM00, BC97], mais également les logiciels permettant d'indexer la structure des documents [RBD06, CR03] ou encore ceux permettant d'indexer les illustrations des documents anciens [PVU+06, UOL05].

Après une rapide présentation des caractéristiques des images de documents anciens, nous présentons un état de l'art sur les méthodes de caractérisation de contenu et d'indexation d'images de documents.

2.1. Caractéristiques des images de documents anciens

Dans le cadre d'une collaboration avec le Centre d'étude Supérieur de la Renaissance de Tours, nous avons eu accès à plus d'une centaine d'ouvrages numérisés datant du XV^{e} et XVI^{e} siècle. Une des caractéristiques fortes de ces documents anciens porte sur l'hétérogénéité des ouvrages disponibles. La rudimentarité des techniques et du matériel utilisé, la dégradation des documents et la variété des règles éditoriales sont quelques unes des raisons expliquant la diversité des images de documents anciens. Les images de documents auxquelles nous avons eu accès recouvrent 3 siècles d'imprimerie et d'histoire. Les mises en pages complexes (plusieurs colonnes de taille irrégulière), l'utilisation de fontes spécifiques (plus utilisées de nos

Figure 1. Extrait de notre corpus d'images de documents anciens.

jours), l'utilisation fréquente d'ornements (enluminures, lettrines, cadres...), les faibles espacements inter-lignes, ou encore la présence d'espacements non constants entre les caractères et les mots, la superposition de couches d'information (bruit, notes manuscrites...), sont autant de spécificités des documents anciens dont il faudra tenir compte pour une caractérisation du contenu. La figure 1 illustre cette diversité des documents traités.

2.2. Analyse d'images de documents : les approches utilisant un fort *a priori*

2.2.1. Définitions

D'une manière générale, les méthodes d'analyse de structures utilisent de forts *a priori* sur la répétitivité supposée de cette structure des documents dans un corpus. Elles se basent sur l'étude de caractéristiques « physiques » *via*, par exemple, des techniques de regroupement ou de fusion de pixels noirs, de découpage de zones d'images, d'étude d'alignements de pixels blancs... Ces méthodes d'analyse de documents ont pour objectif principal de segmenter en blocs les divers éléments de contenu d'une page. Les états de l'art les plus connus ([Doe98, Tru05, MRK03]) décomposent ces méthodes selon qu'elles soient pilotées par les données ou par le modèle. Les approches pilotées pas les données sont caractérisées par une analyse très locale convergeant vers une émergence de blocs ou de zones homogènes. Les approches pilotées par le modèle procèdent à l'inverse par une représentation de plus en plus locale des contenues partant de l'image dans sa globalité. Des travaux de comparaison de méthodes de segmentation d'images de documents sont proposés dans [SKB06]. Ils ont fait apparaître que les méthodes pilotées par les modèles sont efficaces dès lors que la variabilité dans la structure des images est réduite. Les approches pilotées par les données permettent d'appréhender la structure des documents avec une plus grande généricité, au risque d'atteindre certaines incohérences dans la phase de reconnaissance.

2.2.2. Méthodes guidées par le modèle ou les données

L'objectif d'indexation du contenu d'images de documents anciens, nécessite de pouvoir décomposer une page en éléments homogènes pour les catégoriser et ainsi permettre leur indexation. Dans cette optique, nous avons tout d'abord porté notre étude sur les méthodes se basant sur l'extraction et l'analyse des composantes connexes d'un document. L'extraction des composantes connexes, permet de localiser les blocs de pixels connexes présents dans une image. Appliqué sur des images de documents, cela permet, la plupart du temps, de localiser les caractères et les blocs d'illustrations. Dans ce cadre précis, cet outil génère un grand nombre de composantes dont le tri et l'analyse sont généralement laissés à un algorithme étudiant la taille des composantes, les espaces blancs entre elles, leurs positions... Les travaux respectifs de [Bre02, RDB06] illustrent comment il est possible d'extraire la structure d'un document sur la base de ce type d'analyse.

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 d'une image. Ainsi, les auteurs de [Hl03] extraient les lignes d'une image de documents à l'aide de RLSA. L'application de cet algorithme sur les composantes connexes permet de fusionner les lettres en lignes et les lignes en paragraphes.

Une méthode de segmentation est proposée par [Ant98, PZ91]. Elle est particulièrement bien adaptée aux documents dont les zones sont clairement délimitées 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. Ce type d'analyse a été également employé par les auteurs de [CWS03] pour permettre une séparation des dessins et du texte en paragraphes.

Autre méthode guidée par le modèle, l'algorithme de découpage en XY (XY-CUT) est largement employé dans les travaux de segmentation de documents. Proposé par [NKK+88] dans les années 80, son principe se base sur un découpage consistant à appliquer récursivement le même algorithme sur une zone, et ceci jusqu'à ce qu'une condition portant sur la séparation des objets (en paragraphes ou en lignes) soit satisfaite. Le découpage peut, par exemple, être une division en quatre parties égales. Le découpage XY se trouve être 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.
Plusieurs méthodes ont été proposées afin de permettre une caractérisation de document dont la structure présente de fortes variations. Parmi ceux-ci on peut citer les méthodes de segmentation utilisant une décomposition par diagramme de Voronoï. Appliqué aux documents comportant une forte variabilité, le partitionnement par pavage de Voronoï permet de découper une page de manière plus fine que ne le permet le XY-CUT. Ainsi, au lieu d'un découpage rectangulaire il est possible de découper les zones selon leurs contours et de permettre ainsi 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 leur article [KIM99, LWT04], les auteurs présentent différentes méthodes de segmentation texte/dessin (avec les textes séparés en paragraphes). Pour cela, ils extraient les composantes connexes et effectuent le choix d'un échantillon dans les points des contours des composantes, ce qui permet de calculer le pavage de Voronoï sur l'image du document. La fusion des pavés se base sur une étude statistique des distances entre composantes. La densité de chacune est calculée afin d'émettre des hypothèses sur les espaces inter-lettres, inter-mots, inter-lignes et du fusionner les pavés en conséquence.
Enfin, les approches multirésolution pour la segmentation d'images de documents permettent également de traiter des corpus de documents au contenu très variable. Dans leurs articles respectifs, les auteurs de [CLM98, TZ00, SG05] utilisent une pyramide de plusieurs niveaux de résolution pour permettre la reconnaissance de la structure physique des documents qu'ils traitent. Pour chaque résolution de l'image, les auteurs extraient diverses informations liées aux composantes connexes (taille, position...). Une analyse des données obtenues à ces différentes résolutions permet de définir le label des composantes et la manière de les fusionner entre elles pour obtenir une segmentation texte/dessin.

2.2.3. Bilan

Après comparaison de 6 méthodes de segmentation (XY-CUT, RLSA, recherche d'espaces blancs [KIM99], Voronoï, une méthode mixte [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'espaces 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. Dans notre contexte, il serait trop complexe d'éditer des règles de classification ou d'effectuer un apprentissage étant donné la variabilité du contenu en passant d'un ouvrage à l'autre. Ces outils sont donc bien adaptés dans des documents à structure formatée comme le sont les documents contemporains.

2.3. Analyse du contenu d'images de documents : les approches textures

2.3.1. Approches basées sur une étude spatiale des niveaux de gris

Véritables alternatives aux méthodes décrites précédemment, certains outils de traitements des images 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. Parmi ceux-ci les outils d'analyse des textures sont très souvent utilisés sur les images de documents. Il est bien entendu impossible de lister l'ensemble des outils permettant de caractériser les textures. Il existe un grand nombre d'états de l'art sur ce sujet ([Ros99, Lou00, TJ98]). On trouvera également de très bons états de l'art sur le sujet spécifique de la segmentation d'images de documents par approche texture dans [OP00, All04].
Dans [TJ98] l'auteur présente 4 « familles » d'outils de caractérisation de texture. On distingue parmi elles les méthodes statistiques, les méthodes géométriques, les méthodes à base de modèles probabilistes et les méthodes fréquentielles.
Parmi les grands classiques des méthodes statistiques, 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 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 $(d = (d_x, d_y))$. Des attributs calculés sur la GLMC permettent de caractériser la régularité, la répétitivité et le contraste des textures.
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,...).
Pour compléter ce court listing, on peut citer la matrice des longueurs de plages qui se construit en recherchant des successions (plages) de pixels selon un niveau de gris et un angle précis ([Ros99]). Un peu à la manière de la GLMC, on peut calculer des attributs sur cette matrice (importance des plages courtes, répartition des plages...). On peut également citer la fonction d'autocorrélation qui est un outil permettant d'obtenir des informations sur les caractéristiques d'une texture. Ainsi, si la texture est « grossière » (motifs larges) alors la fonction baisse lente-

ment en augmentant la distance d'analyse. Au contraire, si la texture est plus fine (petits motifs peu espacés) alors la fonction décroît rapidement ([UOL05]).
Ces méthodes statistiques de description de textures sont à l'origine de nombreux travaux. Ainsi, [Ros99] utilise certains de ces indices statistiques, pour mettre en place une méthode de segmentation d'images naturelles. Dans [HOP+95] l'auteur propose une méthode d'analyse de textures basée, entre autres, sur le calcul de l'autocorrélation. 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.
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. Cependant on notera 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 des variations 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.
Les méthodes géométriques correspondent à une caractérisation des formes et de leurs relations spatiales composant une texture. Dans [Tuc94], les auteurs montrent qu'il est possible de segmenter des textures à l'aide du calcul des moments géométriques. Il est également possible d'utiliser une méthode géométrique pour segmenter des images de documents. Ainsi, les auteurs de [KRSG03] proposent une méthode de séparation texte/dessin de documents hébreux basée sur la construction d'histogrammes horizontaux. 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 textures extraits sont issus d'une analyse des résultats de projections de pixels selon différents angles.
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. La dimension fractale est, quant à elle, utilisée pour mesurer la rugosité d'une texture et la répétitivité (spatiale ou à différentes résolutions) d'un motif.
Dans [CCMV03], les auteurs utilisent la loi puissance (loi de Zipf) pour identifier des zones d'intérêts dans une image naturelle. Dans [NKPH06], les auteurs utilisent les champs de Markov pour segmenter des images de documents manuscrits en zones d'intérêts labellisées (lignes de texte, rayures, notes de marge...). Ceci leur permet de segmenter les notes manuscrites de Flaubert qui ont la particularité de contenir de nombreuses hachures et ratures rendant les approches classiques peu performantes.

2.3.2. Approches basées sur une étude des texture dans le domaine des fréquences

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. Ces outils fonctionnent dans le domaine fréquentiel. Les transformées de Fourier, Gabor ou ondelettes sont largement utilisées dans les travaux portant sur l'indexation et segmentation d'images naturelles. Dans [MM96b, MM96a] les auteurs utilisent 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 se compose de 4 résolutions et 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. 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 méthodes de segmentation d'images de documents, exploitent généralement le fait que les zones de texte, du fait du grand nombre de transitions encre/papier, sont caractérisées par de hautes fréquences, alors que les images sont généralement constituées de zones homogènes plus étendues et donc associées à des fréquences faibles. Les auteurs de [EDC97, LG00, CC01, RPR05] utilisent Gabor ou les ondelettes pour segmenter leurs images de documents.

2.3.3. Bilan

Selon nous, le principal avantage d'une utilisation d'outils textures 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 guidées exclusivement par les données ou le modèle. Parmi les autres avantages, on peut citer le fait que dans la plupart 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 si ces outils textures permettent de caractériser le contenu des images, elles ne permettent pas d'obtenir une segmentation en blocs (paragraphes, illustrations, titres...) ; cet objectif n'est réalisable qu'au terme de post-traitements.

3. Notre approche texture pour la caractérisation du contenu

La réalisation d'outils de traitements d'images permettant la caractérisation du contenu de documents anciens (et plus globalement de documents à structure variable) soulève un double problème. Le premier est lié a celui de la généricité du processus de caractérisation. Le deuxième problème est lié au traitement de grandes masses de données. 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 nous a amenés à proposer une nouvelle approche de caractérisation de contenu d'images.

3.1. Présentation de notre approche

À 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'informations issues d'une analyse des textures qui composent l'image, sans rechercher ni tenir compte de connaissances *a priori* sur la structure des pages. Face aux caractéristiques des images de notre corpus, nous avons donc décidé de porter notre choix sur l'extraction d'informations bas niveau. Étant donné que l'usage de l'outil est orienté vers des non-spécialistes en traitement d'images, notre méthode 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 d'images de documents. Cette caractérisation des pixels de l'image, sera la base de la création ultérieure d'outils permettant d'analyser le contenu des images.
Le fonctionnement de notre système est décrit dans la figure 2. 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. Ainsi, lorsqu'un nouvel ouvrage est numérisé, il est automatiquement mis en entrée de notre système d'analyse et pour chaque page, cinq indices textures sont calculés. Les trois premiers sont relatifs aux orientations, les deux autres sont relatifs aux informations de fréquences de transitions. Ces indices sont calculés localement 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, les dimensions du dessin étant limitées par la main de l'artiste et celles du texte par les caractéristiques physiques de la presse et les choix de mise en page. À l'aide d'une analyse par fenêtre glissante (dont la taille est l'unique paramètre de notre méthode), il est possible d'associer à chaque pixel de l'image, des métadonnées correspondant aux résultats d'extraction d'attributs texture. Cette analyse est réalisée à quatre résolutions différentes, donnant au final 20 valeurs numériques décrivant chaque pixel. Une fois l'ensemble des pages de l'ouvrage analysé, les informations extraites sont stockées dans une base de métadonnées. La deuxième partie du schéma correspond à l'ensemble des applications potentiellement réalisables grâce aux descripteurs textures stockés dans la base.

3.2. Caractérisation du contenu des images

3.2.1. Caractérisation des orientations par la fonction d'autocorrélation

L'orientation est l'une des principales caractéristiques visuelles impliquée dans la vision préattentive. Nous nous sommes intéressés à cette caractérisation afin de proposer trois indices textures liés aux informations d'orientation. Nous avons choisi d'utiliser un outil non paramétrique basé sur la fonction d'autocorrélation : la rose des directions (proposée par Bres dans [Bre94]).

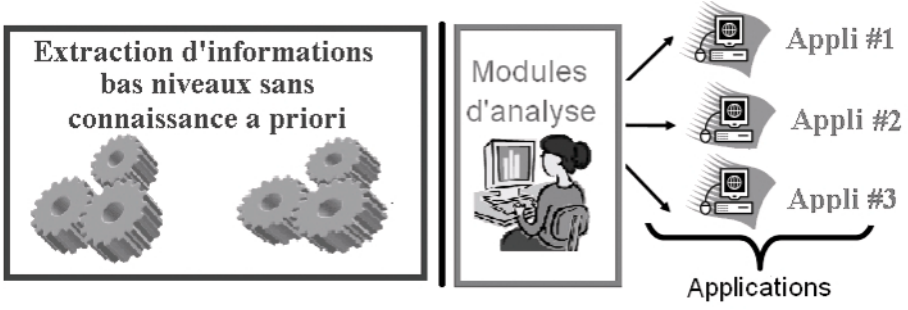


Figure 2. Présentation de notre approche.

La définition de la fonction d'autocorrélation pour un signal bidimensionel est définie par l'équation 1 :

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

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. Cette fonction a déjà été utilisée dans [Pra78, HOP+95] afin de caractériser des textures naturelles. 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 3 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 différente.

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 3.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 (équation 2).

(a) Images testées

(b) Autocorrélations correspondantes

Figure 3. Exemples d'autocorrélation.

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

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

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

Nous allons maintenant étudier le comportement de la rose des directions sur des images de documents. Comme le montre la figure 4, 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 dans la rose. Pour le texte, la forme de la rose est tributaire du nombre de lignes, de la taille des caractères, de l'orientation du texte (4.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 directions (4.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 permettant d'effectuer une caractérisation fine du contenu.

La figure 5 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. Le fait que la rose soit calculée en niveau de gris permet de ne pas être sensible à ce type de bruit. On voit sur la figure 5.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.

Dans le domaine de l'analyse d'images de documents, la multirésolution permet de percevoir des structures de tailles différentes. Dans notre méthode, nous réduisons par 2 les dimensions de l'image à chaque changement de résolution. La figure 6.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 (128×128, 64×64 et 32×32 pixels) utilisées génèrent 3 roses de formes différentes. On remarquera néanmoins que l'information de l'horizontalité ne varie pas. La figure 6.b illustre le même principe, mais cette fois-ci lorsque la rose est calculée sur une illustration. À l'inverse du calcul sur du texte, la rose présente de fortes variations dès lors que la fenêtre est de taille différente.

Une étude complète portant sur la variance de l'allure de la rose dans des conditions d'exploitations sur des images de documents anciens en niveaux de gris a été proposée dans [Jou06]. Cet outil s'est montré adapté et robuste aux bruits présents dans les documents anciens. Nous avons ainsi défini des caractéristiques pertinentes pour décrire le contenu des documents à partir de cet outil.

Nous avons donc décidé d'extraire 3 indices permettant de caractériser les informations relatives aux orientations. Le premier indice extrait, est l'angle correspondant à l'orientation principale de la rose des directions (équation 4). Pour ne pas avoir à manipuler de données circulaires, cet angle est normalisé en fonction de l'écart à l'angle horizontal. Ainsi, pour une résolution k, on calcule pour chaque pixel (i,j) d'une image l'attribut texture 1.

$$Indice\,1^k(i,j) = |180 - ArgMax(R'_{(i,j)}(\theta))| \quad Avec \quad \theta \in [0,\pi] \tag{4}$$

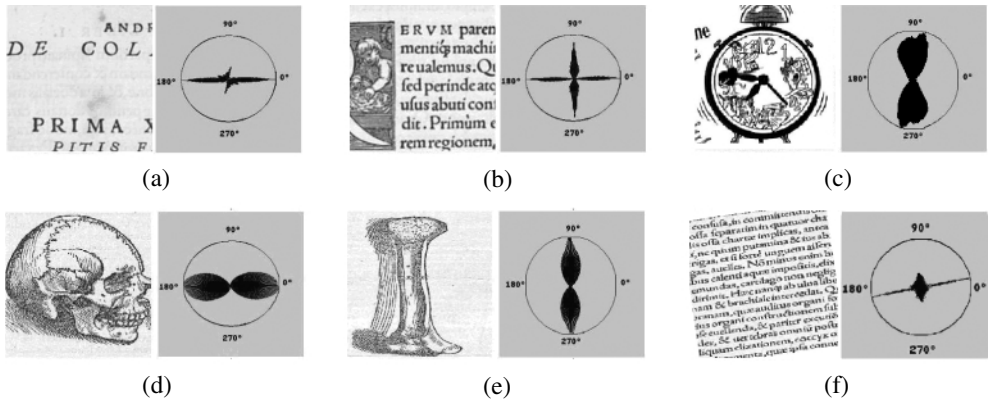
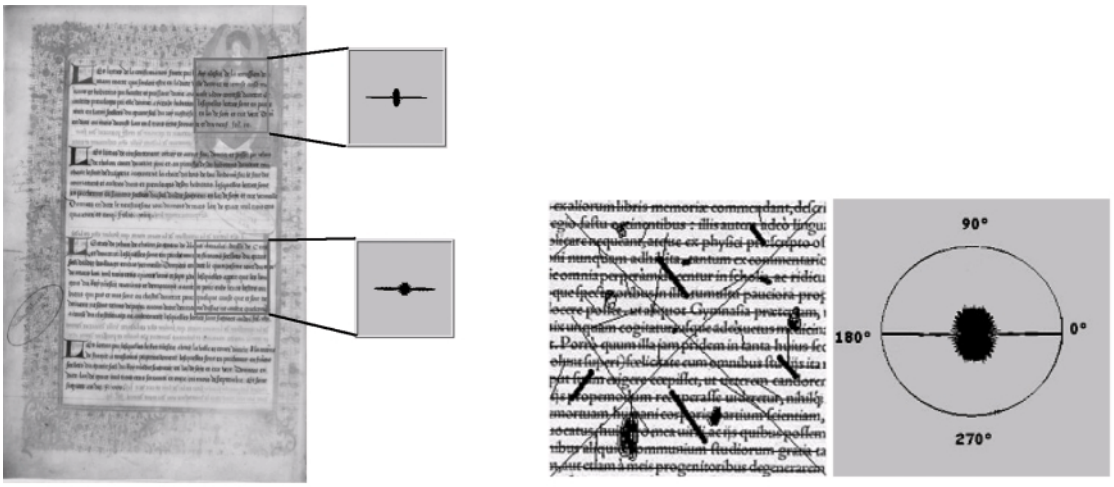


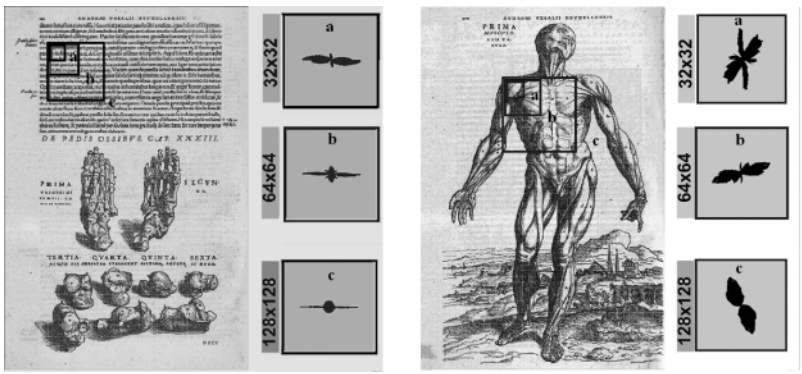
Figure 4. Exemple de roses des directions.



(a) Rose des directions sur une image bruitée

(b) Rose des directions sur du texte artificiellement bruité

Figure 5. Exemples du comportement de la rose sur une image de document bruitée.



(a) Multirésolution sur du texte

(b) Multirésolution sur une illustration

Figure 6. Importance d'un calcul à différentes résolutions.

Une autre caractéristique riche en information, est l'intensité non normée de la fonction d'autocorrélation. 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 calculée par l'équation 2 est plus importante quand l'image présente une forte anisotropie. Cette caractéristique est évaluée en fonction de l'intensité de la fonction d'autocorrélation. Ainsi, pour l'orientation principale trouvée par 4, chaque pixel (i,j) sera caractérisé à l'aide de l'équation 5. Ce calcul est effectué sur la valeur non normée de la fonction d'autocorrélation.

$$Indice\,2^k(i,j) = R(ArgMax(R'_{(i,j)}(\theta))) \quad Avec \quad \theta \in [0,\pi] \tag{5}$$

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. Si la variance est forte, cela signifie que la rose est difforme et que plusieurs orientations sont présentes dans des proportions diverses.

$$Indice\,3^k(i,j) = STD(R'_{(i,j)}(\theta)) \quad Avec \quad \theta \in [0,\pi]\backslash ArgMax(R'_{(i,j)}(\theta)) \tag{6}$$

3.2.2. Caractérisation de la fréquence des transitions encre/fond

En complément des informations liées aux orientations, nous allons extraire en supplément des indices liés aux fréquences. La notion de « fréquence » sur des images de documents, est liée à la fréquence de transition entre le papier et l'encre.
Afin de caractériser les fréquences des 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.
Notre apport se situe à deux niveaux différents. Tout d'abord, les travaux cités ci-dessus exposent des méthodes de caractérisation des fréquences sur des blocs pré-segmentés. Dans notre cas, il faut 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 fonctionnent 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. Nous avons préféré calculer des indices utilisant les niveaux de gris de l'image. 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 niveaux 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 (formule 7).

$$Indice\,4^k(i;j) = Avg(\sum_{i\in I'}(\sum_{j\in J'}(p_{ij} - p_{ij+1})) \tag{7}$$

Avec I' et J' la taille de la fenêtre d'analyse et p_{ij} le niveau de gris du pixel de coordonnées (i,j).

Le dernier indice texture calculé s'inspire de [RBD06] qui propose un algorithme de caractérisation des plages blanches séparant les composantes connexes. Nous recherchons ainsi un moyen d'obtenir des informations sur l'étendue des diverses zones de fond qui jalonnent les pages. Nous avons adopté une approche récursive consistant à calculer 4 itérations d'un algorithme XY-cut récursif. À chaque itération on coupe en quatre zones de taille identique celle qui vient d'être analysée et on calcule pour chacune d'entre elles l'indice de la formule 8. Cet indice est, pour chaque pixel, égal à la moyenne de la somme des niveaux de gris en colonne et en ligne.

$$Indice\,5^k = \frac{\sum_{l\in J'} p^k_{il} + \sum_{h\in I'} p^k_{il}}{2} \tag{8}$$

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

Chacun des attributs présentés dans cette section exprime une caractéristique liée aux fréquences ou aux orientations des motifs présents dans les images. Ils répondent à des informations visibles liées à la distribution des traits sur une page. Ils traduisent ainsi une régularité (dans la distribution des transitions dans les zones de texte) ou au contraire un plus grand désordre (plus généralement dans les zones graphiques). D'un point de vue complexité algorithimique, le calcul de l'autocorrélation, qui est calculée à l'aide d'une transformée de Fourier, est en $O(n\ \log(n))$ avec n le nombre de pixels de la fenêtre d'analyse. La complexité des traitements liés à l'extraction des indices liés aux transitions est en $O(n)$.

3.3. Discussion sur notre proposition

Dans cette section, la qualité de la catégorisation du contenu est évalué au travers d'une classification des pixels sur la base des 20 indices de textures proposés. Classifier les éléments de contenu des ouvrages permet, 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. Chaque pixel de l'image dispose de 20 valeurs issues des 5 indices calculés à 4 résolutions différentes. Notre objectif est de regrouper les pixels de l'image correspondant à des zones homogènes, ce qui revient à regrouper des vecteurs

caractéristiques 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. Nous avons utilisé un algorithme de classification de type centre mobile où seul est indiqué le nombre de classes que l'on souhaite obtenir.

3.3.1. Analyse qualitative et quantitative des résultats

La figure 7 montre le type de résultats que l'on obtient lorsque l'on effectue une classification sur un ouvrage complet. Pour l'ensemble des tests réalisés, la taille de la fenêtre a été fixée à 128x128 pixels (quelque soit la taille de l'image analysée). La taille de l'image est réduite par 2 à chaque changement de résolution. La taille de la fenêtre constitue donc un paramètre de notre proposition.

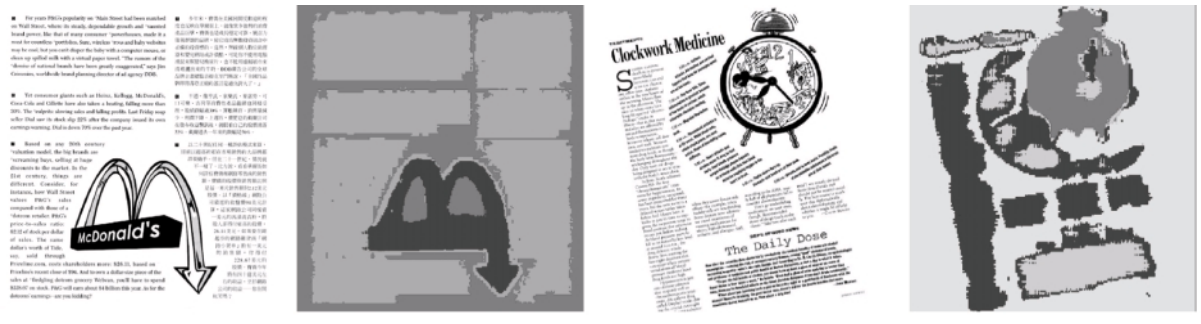
D'un point de vue mise en oeuvre, un ouvrage est considéré comme une seule image où toutes les pages seraient « collées » les unes à la suite des autres. De ce fait, si deux pixels ont la même couleur, cela signifie qu'ils appartiennent à la même classe. Cette classification a un réel sens. En effet, elle permet d'obtenir un point de vue global des textures semblables dans un ouvrage complet. Cela permet donc de fixer un nombre de classes pour un ouvrage, et non pas page par page. Lorsque la classification est réalisée page par page, il est alors complexe de déterminer, *a priori*, le nombre de classes présentes dans chaque image analysée. Par exemple, effectuer une classification à 3 classes sur l'ouvrage complet permet donc, très simplement, d'obtenir une séparation texte/fond/dessin sur l'ouvrage. Ainsi, on notera que sur la figure 7 certaines pages comportent 3 classes et d'autres uniquement 2 car il n'y a pas d'illustrations présentes.

Figure 7. Classification de pixels d'un ouvrage.

Nous avons également testé cette classification de pixels sur des images de documents contemporains. La figure 8 montre des exemples de résultats obtenus sur des images extraites de [MD05, KIM99]. Les résultats de la figure 8.a-b montrent qu'une classification à trois classes a permis de séparer correctement les pixels de textes des pixels de fonds et de dessins. Les résultats de la figure 8.c-d montrent qu'avec une classification à 4 classes, les deux types de textes (gras et non gras) font partie de deux classes différentes.

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 du marquage proposé, 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 la figure 7) ou encore les dessins qui sont très proches d'une zone de texte ne sont pas clairement marqués (par exemple une lettrine dans la figure 7).

Afin de mesurer la pertinence de la méthode, nous proposons une évaluation simple des capacités de nos données plutôt que de donner une grande quantité de résultats visuels, nous avons donc décidé d'évaluer la capacité de 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 zones de dessins et des zones de texte (vérité ter-



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

Figure 8. Classification de pixels de documents contemporains (images issues de [MD05, KIM99]).

rain). Un fichier est créé afin de stocker cette vérité terrain pour être finalement comparée à la classification calculée. Nos tests ont été réalisés sur 400 pages de documents anciens, extraites de 9 ouvrages différents. Étant donnée que nous voulions avoir une idée de la pertinence des indices extraits, nous avons composé un corpus d'images de test au contenu le plus varié possible.
Les résultats référencés dans le tableau 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 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.
Exceptés quelques zones de transitions difficiles à analyser et qui font chuter les taux de bonne classification, les indices textures utilisés permettent une bonne séparation des couches d'information des documents.

Tableau 1. Taux de bonne classification des pixels de dessin et de texte.

Dessin	Texte
83 %	92 %

3.3.2. Analyse comparative avec une approche par bancs de filtres de Gabor

Afin de juger de la pertinence de nos indices textures, nous avons également comparé nos résultats de classification à ceux obtenus après catégorisation du contenu par application de filtres de Gabor. Ces filtres permettent de caractériser les fréquences et orientations des textures présentes dans les images. Nous avons choisi de nous inspirer des algorithmes présentés dans [BSN04, CC01, RPR05] et qui permettent de segmenter les images de documents en détectant les zones de l'images possédant des caractéristiques d'orientations et de fréquences spécifiques.
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 les articles de référence, nous ont amenés à construire un filtre composé de 5 orientations $\theta_l = \{0, 30, 60, 90, 120\}$ 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 au même algorithme de classification utilisé pour les tests précédents.
La figure 9 montre quelques résultats obtenus sur des documents contemporains. Les résultats sont conformes à ceux attendus, avec notamment une bonne détection du texte quelque soit son orientation.
Nous avons appliqué des bancs de filtres sur des images de documents anciens. La figure 10 montre deux résultats de classifications résumant la qualité des résultats obtenus. Le problème récurrent est 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 deuxième image de la figure 10, 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.

3.3.3. Vers une réduction de l'espace des caractéristiques

Compte tenu de la quantité de données générées, une analyse de nos données est une étape obligatoire. Le fait de calculer des

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

Figure 9. Segmentation texte/dessin avec Gabor.

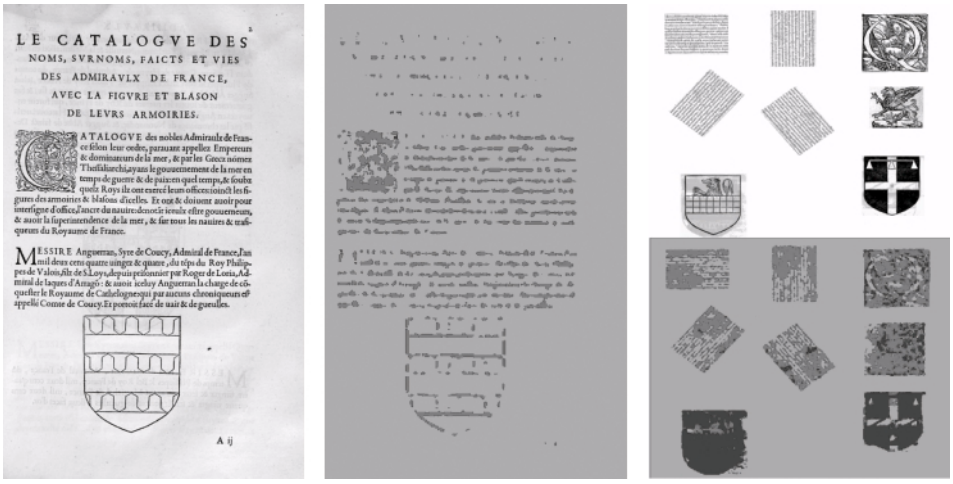


Figure 10. Segmentation de documents anciens avec Gabor.

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.

Une analyse factorielle (ACP) des données a permis de faire ressortir plusieurs informations intéressantes. La première est que l'inertie portée par les 4 premiers axes est toujours très bonne. Il est donc possible dans tous les cas, de réduire nos données d'une dimension 20 à une dimension 4 tout en gardant environ 78 % de l'information. La deuxième information ressortant de cette analyse est que les indices liés aux transitions encre/papier (équation 7) et à l'intensité de la corrélation pour l'orientation principale(équation 5) sont fortement corrélés. Ceci s'explique, par le fait que le texte est toujours horizontal et qu'il est très fortement présent dans nos images. L'indice lié aux transitions est sensible aux fortes transitions noir/blanc correspondant aux lignes de texte. Il se trouve que l'indice lié à l'intensité de l'autocorrélation est lui aussi sensible aux orientations privilégiées (ici horizontales). Il est donc tout a fait possible de ne calculer que 4 des 5 indices proposés dans cet article. Enfin, la dernière information pertinente, est que nous n'avons pas réussi à faire apparaître des combinaisons 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 variables synthétiques. Une combinaison linéaire est directement liée à la composition intrinsèque de la page.

La figure 11 illustre, entre autres, les relations existantes entre les variables lorsqu'on les projette sur le premier plan factoriel. Par mesure de clarté, les indices projetés sont symbolisés par une lettre (A : longueurs de plages, B : Transitions Encre/papier, C : Variance de la rose, D : Orientation principale, E : valeur de la fonction d'autocorrélation), chaque chiffre indique la résolution à laquelle a été calculé l'indice.

D'autre part, 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 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.

L'ensemble des tests réalisés a montré que l'inertie des 4 premiers axes ne varie pas et reste égale à 78 %. La question est donc de savoir si prendre un échantillon de pixels va dénaturer les relations entre variables.

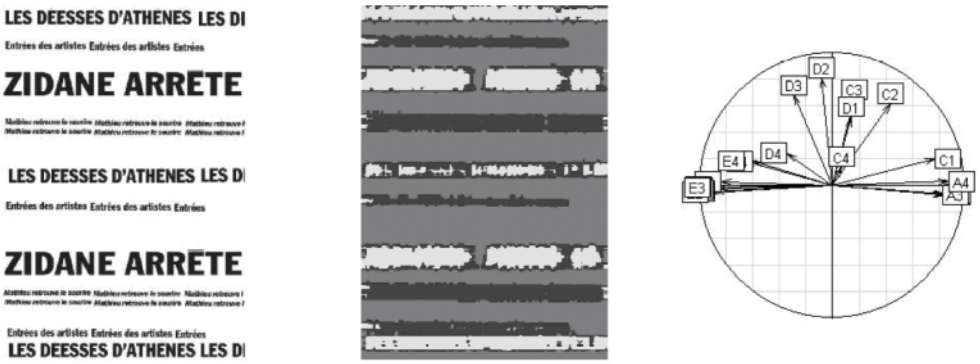


Figure 11. Corrélation entre variables.

Sur les essais réalisés sur des images de 600×889, il faut descendre en dessous de 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.

4. Vers de nouvelles applications de recherche d'information par le contenu

Si l'on se réfère au schéma 2, il est possible de réaliser des applications comme des modules indépendants venant utiliser les indices extraits dans la première phase de notre méthode. Les expérimentations que nous réalisons dans cette dernière section, illustrent la faisabilité de nouvelles applications qui s'appuient sur les indices extraits de façon systématique.

4.1. Comparaison de pages

La première expérimentation que nous souhaitons réaliser consiste en une comparaison de pages d'images de documents anciens. Cette expérimentation permettra d'étudier s'il est possible, sans segmenter et sans identifier la structure des pages, de comparer leurs contenus à partir d'informations textures. Nous avons choisi de caractériser les pages par l'organisation spatiale des blocs de textes, d'images et du fond. Sur la base de cette définition, nous proposons l'utilisation d'outils de comparaison de partitions présentés dans [YS041]. Dans le cadre de notre travail, une partition est le résultat d'une classification de pixels réalisée sur la base des indices textures générés.

Ci-dessous nous détaillons les différentes étapes permettant la comparaison de deux classifications de pixels. Soient deux images α et β pour lesquelles une classification de leurs pixels a été réalisée. Il est alors possible de construire le tableau de contingence $\overset{\alpha,\beta}{N}$ de ces deux images α,β (9). Ce tableau permet de comparer deux 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).

$$N^{\alpha,\beta}_{uv\in p,q} = \sum_i X^i_{uv}$$
$$X^i_{uv} = \begin{cases} 1 & \text{si } L^\alpha(i) = u \quad et \quad L^\beta(i) = v \\ 0 & \text{sinon} \end{cases} \tag{9}$$

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} (10).

$$\begin{aligned} \textstyle\sum_i \sum_{i'} R^\alpha_{ii'} &= \textstyle\sum_u N^2_\alpha \\ \textstyle\sum_i \sum_{i'} R^\beta_{ii'} &= \textstyle\sum_v N^2_\beta \end{aligned} \tag{10}$$

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 (11).

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

Soit

$$a = \sum_u N^2_{uu} \tag{11}$$

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 (12).

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

Soit

$$b = n^2 + \sum_u \sum_v N^2_{uv} - \sum_u N^2_u - \sum_v N^2_v \tag{12}$$

Ainsi, les indices calculés pour la comparaison de deux partitions qui sont présentés dans [YS04] permettent de répondre à une requête de type « recherche de contenus similaires d'une page ». 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 calculons 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 13. Deux images sont donc semblables, s'il y a une stabilité entre associations de pixels d'une image à l'autre.

$$\begin{aligned} R &= (a+b)/n^2 \\ &= \frac{\sum_u \sum_v N^2_{uv} + \sum_u N^2_{uu} - \sum_u N^2_u - \sum_v N^2_v + n^2}{n^2} \end{aligned} \tag{13}$$

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.

Pour évaluer la qualité de la comparaison d'images de documents, nous nous sommes inspirés des travaux de [MMS06]. Nous avons décidé de séparer les documents en 5 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 disposées sur deux colonnes, les pages composées d'une lettrine et le reste de la page composé uniquement de texte et enfin les pages composées entièrement de dessins.

Les résultats montrés dans la suite de cet article ont tous été réalisés sur la même base d'images. Nous avons ainsi choisi près de 400 pages de 9 ouvrages différents. Chaque test débute avec l'application de l'algorithme de classification pour 3 classes à partir des vingt indices que nous proposons. Chaque partition est comparée à toutes les autres à l'aide de l'indice 13. Enfin, l'ensemble des comparaisons permet de construire une matrice de similarité entre les images constituant la base étudiée.

Sur les 400 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 comportant un cadre lorsque l'image en requête en comporte un.

La figure 12.a illustre la capacité de l'indice utilisé à discerner des pages visuellement similaires. L'image requête utilisée dans la figure 12.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. La figure 12.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 correspondant aux figures 13.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 13.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. De même, les résultats de la figure 13.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).

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. Le tableau 2 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 Top15. 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. Nous

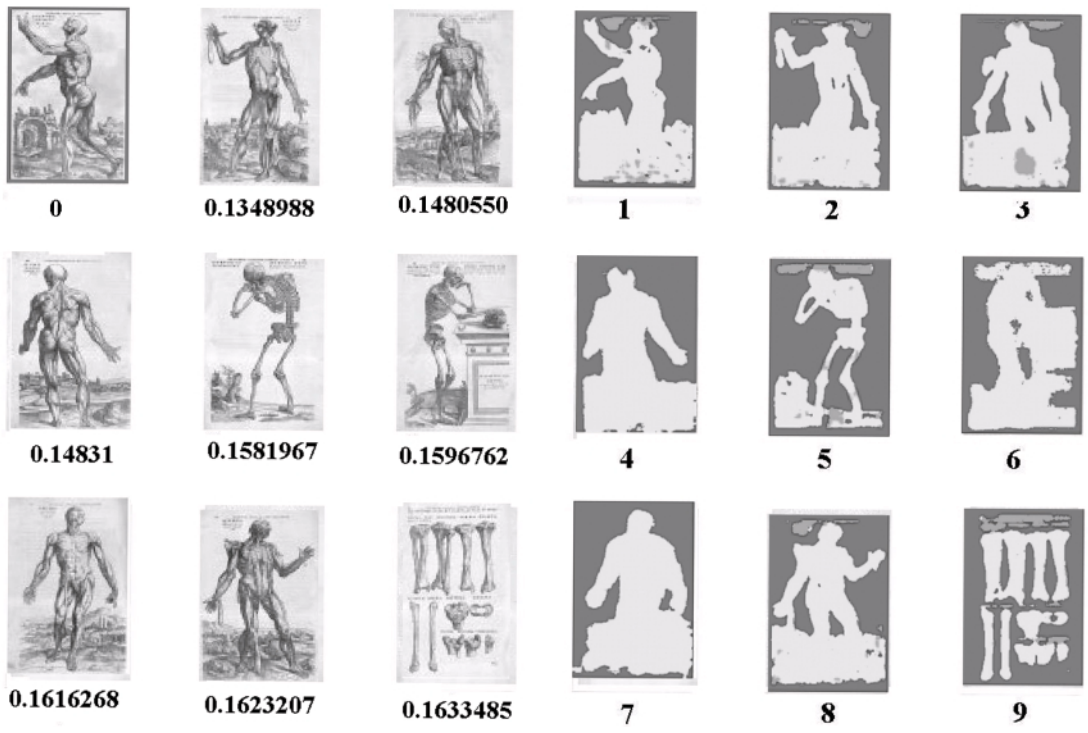


Figure 12. Exemples de résultats de requêtes utilisant l'indice de comparaison de partitions modifié (R′).

Figure 13. Exemples de résultats de requêtes utilisant l'indice de comparaison de partitions modifié (R').

préférons calculer de manière plus 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) $Taux = \frac{Bonnes\ réponses}{Taille\ du\ Top}$.

Tableau 2. Taux de précision obtenus pour 5 types de requêtes différents.

	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

Dans les tests effectués, toutes les pages de tous les ouvrages sont mélangées. La mesure utilisée 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 pour 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, 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 suffisamment 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 dispersion des éléments d'une classe dans une page, permettraient certainement d'améliorer les résultats.

À travers cette expérience menée 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. 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 de mesures d'appariement qui en découlent.

4.2. Comparaison d'illustrations texturées

La deuxième expérimentation consiste à réaliser un recherche d'images par le contenu sur une base constituée d'images de traits de documents anciens. 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 humanistes de la région Centre[1]. Plus d'un tiers de la base est constitué de lettrines, le reste se divise en plusieurs catégories : blasons, personnages, emblèmes, crânes, éléments décoratifs divers... Nous souhaitons calculer une similarité entre deux images en fonction de leurs caractéristiques textures qui les composent. Pour cela, nous proposons l'utilisation d'une métrique permettant de mesurer une similarité entre deux matrices d'indices textures. 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 est faible, plus les pixels sont considérés comme proche relativement aux indices de texture. 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 « semblables ».

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

Avec $d(P_k, P_l)$ la mesure de similarité entre deux images k,l et $\overset{k}{C}$ la matrice décrivant les textures des images k et l. 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 système fournit les images qui lui sont le plus similaire.

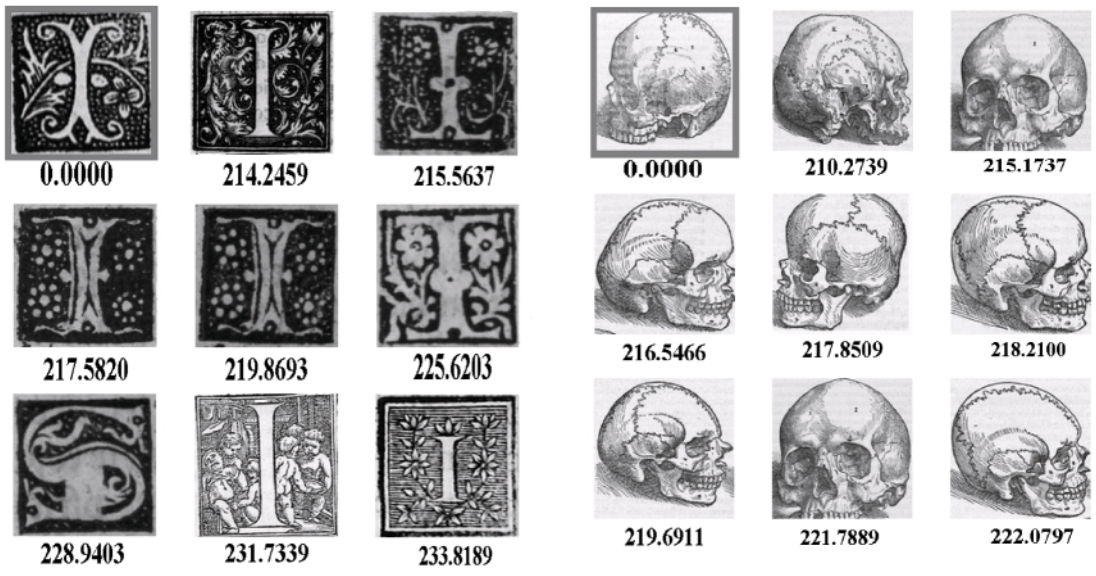
La figure 14 illustre les bons résultats obtenus sur la base d'images de traits. L'image requête est entourée en rouge, au dessous de chaque image réponse est indiqué la mesure de similarité (non normé) entre cette image et l'image requête. Après étude des résultats, nous constatons que la discrimination 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.

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 top 5, 10 et 15 pour 5 textures différentes de la base. Le tableau 3 récapitule les taux moyens obtenus.

Tableau 3. Taux de Précisions pour des requêtes effectuées sur des images de traits.

	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

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 [PVU+06] 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 %. Nous obtenons



(a) Recherche de lettrines

(b) Recherche de crânes

Figure 14. Recherche dans un ouvrage.

[1] Base constituée par le CESR de Tours : http://www.bvh.univ-tours.fr/madonne.asp

donc de moins bons résultats que ceux obtenus par des méthodes dédiées à l'analyse de ce type d'images (et plus spécifiquement des lettrines).

5. Conclusion et perspectives

Cet article présente notre proposition d'outils de traitements d'images pour une caractérisation des images de documents sans connaissance *a priori*. 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 comment il est possible de caractériser le contenu d'images de documents en se basant sur des informations textures non paramétriques et une approche multirésolutions. Cette démarche se veut générique et adaptable à tout type d'ouvrage en s'appuyant sur l'homogénéité intraouvrage. En extrayant des signatures liées aux fréquences et aux orientations des différentes parties d'une page, il est possible d'extraire et de comparer des éléments de contenu sans émettre d'hypothèses sur la structure physique ou logique des documents analysés. Cette approche, est intéressante puisqu'elle laisse la possibilité d'ajouter des indices supplémentaires correspondant à des caractéristiques non extraites dans la version proposée dans cet article.

Notre proposition permet d'entrevoir de nombreuses perspectives en terme de réalisation d'outils d'aide à la navigation ou d'aide à l'indexation. Il reste maintenant à étudier leur intégration dans des dispositifs d'indexation plus complets (*ie* : systèmes de recherche d'images par l'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.

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 article permettront 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[2].

[2] http://www.bvh.univ-tours.fr/

Références

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

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

[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.

[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.

[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.

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

[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.

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

[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.

[Doe98] 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.

[HB00] Mryka HALL-BEYER. Glcm texture : A tutorial. Technical report, 2000.

[HI03] Karim HADJAR and Rolf INGOLD. Arabic newspaper page segmentation. *icdar*, 02:895-900, 2003.

[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.

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

[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.

[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.

[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.

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

[Lou00] Etienne LOUPIAS. *Indexation d'images : aide au télé-enseignement et similarités pré-attentives*. 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.

[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.

[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.

[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.

[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.

[PA02] CORNU Philippe and SMOLARZ André. Caractérisation d'images par textures associées. *Traitement du signal (Trait. signal)*, 19(1):29-35, 2002.

[Pra78] W.K. PRATT. *Digital Image Processing (Book : First Edition)*. Wiley, 1978.

[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.

[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.

[RPR05] S.S. RAJU, P.B. PATI, and A.G. RAMAKRISHNAN. Text localization and extraction from complex color images. *ISVC05*, pages 486-493, 2005.

[SG05] Zhixin SHI and Venu GOVINDARAJU. Multi-scale techniques for document page segmentation. *ICDAR*, 0:1020-1024, 2005.

[SKB06] Faisal SHAFAIT, Daniel KEYSERS, and Thomas M. BREUEL. Performance comparison of six algorithms for page segmentation. 3872:368-379, Feb 2006.

[TJ98] M. TUCERYAN and A. K. JAIN. Texture analysis. *In The Handbook of Pattern Recognition and Computer Vision (2nd Edition)*, pages 207-248, 1998.

[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.

[WCW82] WONG, CASEY, and WAHL. Document analysis system, ibm journal of research and development. *IBM Journal of Research and Development*, 26:647-656, 1982.

[YS04] YOUNESS and SAPORTA. Une méthodologie pour la comparaison de partitions. *Revue de Statistique Appliquée*, 52:97-120, 2004.

Nicholas **Journet**

Nicholas Journet travaille actuellement dans l'équipe Reconnaissance des Formes et Analyse d'Images du Laboratoire d'Informatique de Tours. Il a soutenu sa thèse de doctorat en 2006. Sa thématique de recherche est l'analyse et l'indexation d'images de documents. Il travaille plus particulièrement sur la mise en place de méthodes a base d'extraction d'indices texture permettant la rétro-conversion d'images de documents anciens.

Jean-Yves **Ramel**

Jean-Yves Ramel a obtenu sa thèse de doctorat en Informatique en 1996. Maître de conférences à l'INSA de Lyon puis maintenant à l'Ecole Polytechnique de l'Université François Rabelais de Tours, ses activités de recherche actuelles concernent principalement le domaine de l'écrit et du document.

Ces travaux portent plus particulièrement sur les méthodes structurelles de reconnaissance des formes et sur les architectures et stratégies d'analyse d'images permettant la mise en place d'algorithmes coopératifs d'extraction d'information dans les images de documents anciens et graphiques.

Véronique **Eglin**

Titulaire en 1998 du doctorat en informatique, Véronique Eglin travaille depuis septembre 2000 comme maître de conférences à l'INSA de Lyon et est rattachée depuis 2003 au laboratoire LIRIS UMR 5205. Elle a participé depuis 2003 à différents projets de numérisation et de valorisation des grandes masses de données écrites issus de corpus littéraires des sciences humaines du IX^{ème} au XVIII^{ème} siècle. Ses thèmes de recherches sont centrés autour de la caractérisation et la classification des écritures manuscrites anciennes, l'identification des scripteurs, l'analyse de texture pour la caractérisation des textes, des typographies et de la mise en page et la recherche d'informations dans les documents.

Rémy **Mullot**

Rémy Mullot a une formation en Informatique Industrielle et Automatique qui lui a permis de soutenir sa thèse de doctorat de l'Université de Rouen en Janvier 1991 et son habilitation à diriger des recherches en 2000 en au sein du laboratoire L3i, dans le domaine de l'interprétation de documents techniques et cartographiques. Actuellement directeur du laboratoire d'Informatique Images et Interactions de La Rochelle (L3I), Rémy Mullot travaille activement sur l'analyse et l'indexation d'images de documents.