Contribution à la numérisation de documents patrimoniaux

Trinh Eric, LeBourgeois Frank, Emptoz Hubert

Laboratoire de Reconnaissance de Formes et Vision
I.N.S.A. de LYON - Bât 403
20 Av. A. Einstein 69621 Villeurbanne Cedex France
Tél : (33) (0)4 72 43 60 54 Fax : (33) (0)4 72 43 80 97
Emails : [trinh,flebourg,emptoz]@rfv.insa-lyon.fr

***Résumé*—Les progrès de l'informatique de ces dernières années permettent la conversion totale d'un document papier vers un document électronique. La numérisation est une phase critique dont dépend toute la suite de la dématérialisation d'un document. Les images numérisées doivent permettre de travailler sur ce patrimoine documentaire sans qu'il soit nécessaire de revenir à l'original. C'est la raison pour laquelle il faut définir un protocole qui permette, dans l'immédiat d'assurer un travail de qualité à partir de ce support et à terme, de revaloriser ce capital grâce à la pérennité des informations extraites par l'anticipation des besoins futurs. Pour illustrer notre propos, nous nous appuierons sur le cas du projet européen DEBORA (Digital accEss to BOoks of the RenAissance), ce projet propose une solution de consultation de livres numérisés du XVIe siècle.**

***Mots clés*— document, numérisation, compression, traitement**

I. Introduction

Avec les récents progrès et la démocratisation des moyens informatiques, la conversion des documents papiers en documents numériques est l'un des grands enjeux technologiques actuels. Ces dernières années, nous avons vu fleurir un grand nombre de vastes projets de numérisation de fonds de bibliothèques ou de musées. La numérisation doit être considérée comme une opération plus large que la simple digitalisation et doit être redéfinie comme l'ensemble des traitements qui participent à la construction de l'image finale. La numérisation s'effectue en plusieurs étapes qui commence par la digitalisation à l'aide de scanners, suivie des prétraitements (corrections géométriques, binarisation, restauration). Mais elle est aussi marquée tout le long de son processus par plusieurs phases de stockage. Les nombreux projets de numérisation commencés souvent dans l'urgence, par plusieurs bibliothèques et centres de documentation, sans une réelle étude préalable des besoins présents et futurs ni une analyse des problèmes technologiques, peuvent conduire à l'échec avec des conséquences financières désastreuses. Actuellement, nous rencontrons encore beaucoup de projets qui se sont satisfaits d'une prise d'images tout juste suffisante pour une interprétation humaine de leur contenu. Or, avec la croissance du nombre de documents à analyser, leur traitement manuel devient de plus en plus difficile et onéreux. Le traitement automatique devient une solution incontournable pour à la fois retranscrire le texte grâce à la Reconnaissance Optique de Caractère (OCR), et la reconnaissance de la structure des documents avec des logiciels de Reconnaissance Automatique de Documents (ARD). La qualité des images, qui dépend à la fois des conditions de la numérisation, des prétraitements appliqués aux images et du type de compression utilisé, a une grande influence sur les résultats des logiciels de reconnaissance. Ainsi un échec de la numérisation condamne définitivement l'utilisation des logiciels de reconnaissance. Notre étude portera sur les trois principaux points sensibles liés à la numérisation de documents. Dans un premier temps, nous présenterons la problématique globale liée à cette numérisation en nous appuyant sur le projet DEBORA. Ensuite, nous étudierons les problèmes de stockage de l'information précédemment acquise avec notamment l'utilisation de la norme de compression JPEG mais aussi du nouveau format JPEG2000. Nous terminerons par une étude sur les effets des outils de restauration sur les documents qui sont utilisés à des fins d'exploitation.

II. Problematique

L'objectif du projet DEBORA [4] (Digital accEss to BOoks of the RenAissance) est de développer des outils permettant l'accès, à partir de postes de consultation distants, à des

collections de documents du XVIe siècle par la numérisation des ouvrages. Nous avons pour cela dû mettre en place une plate-forme de numérisation complète de la prise des images à leur restauration. Lors de cette expérience, nous avons défini les conditions minimales requises pour que les images des documents puissent être interprétées automatiquement. Pour cela, il faut éviter au maximum, les pertes d'information lors de l'acquisition, du traitement et de la sauvegarde de chacune des images extraites.

Nous avons eu l'occasion de participer à des projets de revalorisation de plans de numérisation en développant des outils de traitement automatique. Malheureusement, beaucoup d'entre eux présentaient des images d'une qualité toute juste suffisante pour la perception humaine, où le seul critère de qualité est la capacité d'un utilisateur à lire le contenu. Ce concept macroscopique n'est cependant pas suffisant pour une analyse automatique, en effet une image lisible pour l'homme n'est pas nécessairement interprétable par une machine et réciproquement. A titre d'exemple, il n'est pas rare de rencontrer des ouvrages qui ont été numérisés à partir de photocopies, ni même de trouver des images sous-échantillonnées, mal binarisées et déformées par certains algorithmes de compression très largement utilisés comme JPEG.

La numérisation devrait permettre de pérenniser des documents rares et importants pour la culture et aussi de les rendre accessibles au grand public. Mais il faut équilibrer la quantité d'information à conserver et la qualité de l'image pour une utilisation ultérieure ; pour cela il est nécessaire d'anticiper :

- l'augmentation croissante des performances du matériel informatique,
- l'évolution des capacités des futurs réseaux (autoroutes de l'information),
- les fonctionnalités présentes et futurs des logiciels de reconnaissance.

La capture de l'image est l'étape primordiale qui déterminera ses qualités intrinsèques.

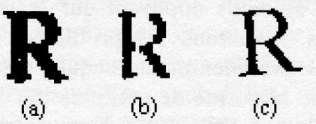


Figure 1: Exemple de dégradations: (a) modification du contour, (b) altération de la connexité, (c) inclinaison.

La principale cause de dégradation que nous avons observée est directement liée à une manipulation inadaptée lors de l'acquisition des images. En effet, il arrive souvent que l'ouvrage papier devant être numérisé existe déjà sous une autre forme. Ainsi, il est fréquent de rencontrer des numérisations effectuées à partir d'une image sur microfiche voire même issue d'une photocopieuse. Le passage par ces étapes intermédiaires implique inéluctablement l'ajout de défauts et d'approximations au résultat final.

Pour le projet DEBORA, les livres ont été digitalisés dans chaque bibliothèque propriétaire des ouvrages, notamment à Rome, à Lyon et à Coïmbra (Portugal). Nous avons recommandé de raccourcir le plus possible la chaîne de numérisation en évitant les manipulations ou les traitements superflus. Les livres anciens ont donc été directement digitalisés et les images brutes ont été conservées sans aucun traitement en prévision de traitements automatisés. Nous avons conseillé de numériser les documents avec la résolution optique maximale permise par le matériel d'acquisition et un nombre de bits par pixels supérieur ou égal à 8, c'est à dire au minimum 256 niveaux de gris. Les images brutes ont été stockées sans compression avec perte ni avec un format de stockage destructif. Les performances du scanner utilisé doivent être suffisantes pour que la résolution minimale du corps de texte de la plus petite police utilisée dans le document puisse être reconnu par les logiciels OCR actuels ; une meilleure résolution permet d'améliorer rapidement les performances de la reconnaissance et au travail de mémoire des historiens..

III. Stockage

Le problème du stockage des données est important dans tout projet de numérisation car il est associé directement à son coût. On notera ainsi qu'en moyenne, un livre d'environ 300 pages numérisées en mode image représentera une quantité d'information de plus de 3Go soit près de 5 CDs. Pour réduire cet encombrement, il est souvent tentant de faire appel à des algorithmes de compression dont notamment ceux avec pertes (JPEG, JPEG2000…) car ils permettent une diminution certaine du volume des données. Malheureusement, nous démontrerons que ces méthodes basées sur des modèles psycho-visuels détruisent les informations nécessaires à la reconnaissance automatique.

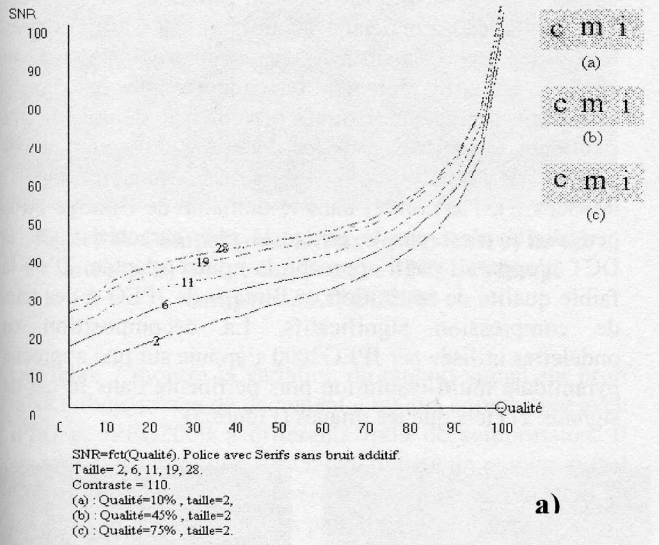
A. Etude de la compression JPEG

La compression des images couleurs ou en niveaux de gris n'est pas efficace avec les méthodes de compression sans pertes. C'est la raison pour laquelle ce sont essentiellement les méthodes de compression avec pertes qui sont utilisées pour comprimer les images de documents et notamment la compression à base de transformée en cosinus discrète comme JPEG ou plus récemment à base d'ondelettes. Nous avons effectué une étude des effets de la compression JPEG [1][11] sur les images de documents du point de vue de la reconnaissance automatique et de la lisibilité. Les facteurs impliqués dans l'influence de la compression et que nous avons étudié sont :

- le facteur qualité de compression JPEG,
- le contraste,
- la taille de la police des caractères,
- la quantité de bruit additif.

Nous avons négligé d'autres facteurs moins importants dans notre étude :

- la police de caractères
- la déclinaison de la police (gras, italique)
- les lettres utilisées.



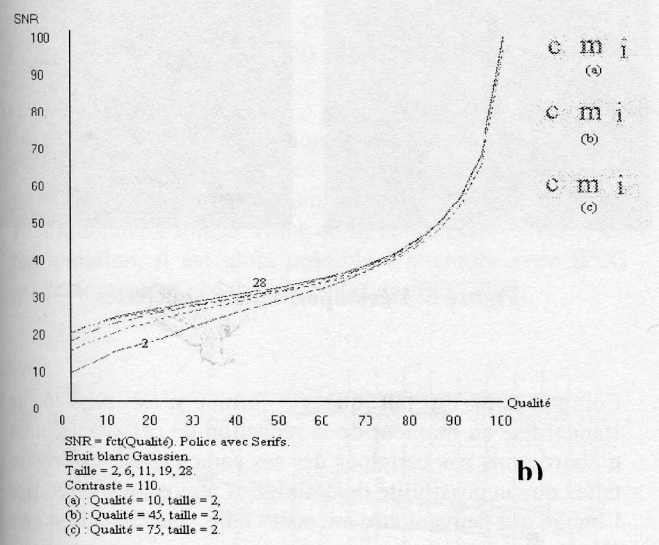


Figure 2: SNR en fonction de la qualité de la compression JPEG et de la taille des caractères : (a) sans bruit , (b) avec un bruit additif .

Les courbes montrent le rapport signal sur bruit (SNR) en fonction du facteur de qualité de la compression JPEG sur des images de synthèse sans bruit additif (Figure 2a) et avec un bruit additif (Figure 2b) en fonction de la taille des caractères. Les courbes SNR montrent que la compression JPEG modifie plus de 50% de la qualité de l'image dès que l'on diminue le facteur de qualité de 100% à 90% seulement, ceci signifie que la compression JPEG dégrade rapidement le signal, même avec un fort coefficient de qualité. Il est donc utopique de croire qu'une compression avec un fort taux de qualité préserve l'image des documents. En utilisant le taux moyen de compression JPEG fixé, par défaut, dans tous les logiciels à 70%, la dégradation du signal est supérieure à 60%. La taille de la police influence les résultats ; la dégradation de l'image est d'autant plus sensible que la taille de la police est petite. Cela s'explique par l'augmentation de la complexité des frontières des lettres à l'intérieur des blocs 8x8 que le filtrage fréquentiel ne peut plus restituer. Avec un bruit additif gaussien, la différence entre l'image d'origine et l'image comprimée est encore plus significative. Le bruit introduit des fréquences aléatoires qui peuvent être confondues avec des fréquences qui concernent la forme des caractères. Les fréquences correspondant au bruit peuvent être préservées aux détriments des fréquences plus importantes. Des courbes similaires ont été obtenues en mesurant le rapport signal sur bruit en fonction de la qualité pour différentes valeurs de contraste entre le caractère et l'arrière-plan ; cette observation montre que le contraste constitue un paramètre important. La compression JPEG préserve mieux les images si les caractères sont fortement contrastés. La présence de bruits dans l'image d'origine affecte plus le résultat de la compression que la modification de taille des caractères. Par conséquent, après le facteur de qualité de la compression JPEG, le bruit est le second facteur le plus important, il est suivi par le paramètre de contraste.

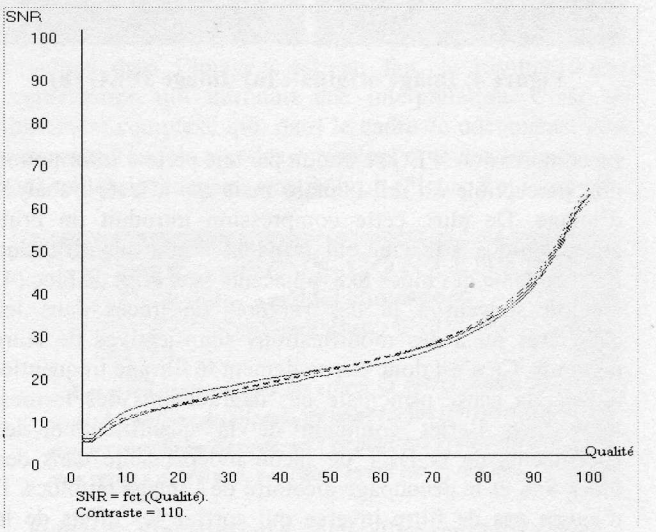


Figure 3: SNR en fonction de la qualité de la compression JPEG sur des images réelles.

Nous avons reproduit sur la Figure 3 les mêmes tests sur des images réelles de document du XVI^{ème} siècle ; elles présentent des bruits naturels qui ne sont pas réellement reproductibles avec les générateurs aléatoires programmés. Malgré un comportement assez similaire des courbes, nous constatons une aggravation de la perte d'information en présence des bruits réels que contiennent ces images. Ce

résultat s'explique par le fait que le bruit naturel des images est beaucoup plus complexe que le bruit synthétisé car c'est l'accumulation de plusieurs bruits provenant d'origines différentes dont la principale est probablement le bruit qui est lié aux formes des traits des caractères. Ce bruit présent dans le voisinage des contours des traits n'est pas reproduit dans les expériences précédentes et ce type de bruit contribue plus à augmenter les différences entre l'image compressée et l'image originale

Nous avons mesuré l'impact de la compression JPEG sur la reconnaissance de caractères ; les premiers évaluations du taux de reconnaissance, en fonction de la qualité JPEG sur des mots synthétisés grâce à des polices vectorielles, montrent qu'avec une facteur de qualité JPEG en dessous de 30%, les résultats de la reconnaissance deviennent nuls. Pour les mêmes raisons décrites précédemment, l'impact de la compression JPEG sur la reconnaissance de caractères deviendra encore plus important en fonction de la diminution de la taille des caractères, du contraste de l'image et du niveau de bruit.

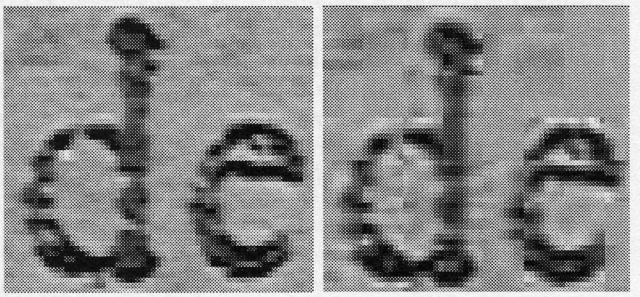


Figure 4: Image originale (a) Image JPEG (b)

La compression JPEG se traduit par une perte d'information non perceptible à l'œil humain mais qui affecte l'analyse d'image. De plus, cette compression introduit un bruit algorithmique artificiel qui provient d'une quantification indépendante des blocs 8x8 adjacents ; cet effet de bloc (4) conduit souvent à la des ruptures de tracés dans les caractères ou à des modifications significatives de leurs contours. Ce n'est donc pas seulement le filtrage fréquentiel qui est la cause principale des dégradations des formes, mais bien l'effet conjugué de la quantification des coefficients de la DCT de façon indépendante dans des blocs 8x8 et le découpage aléatoire de l'image en blocs. Il n'existe pas de filtre inverse qui corrige les effets de la compression JPEG. Même si certains artifices comme le filtrage des contours permettait de les améliorer, nous ne pourrions jamais retrouver l'information perdue d'origine. Ainsi, les fonds documentaires déjà numérisés et stockés sous ce format resteront difficilement interprétables par ordinateur. En constatant qu'il n'est pas possible aujourd'hui de comprimer des images de façon significative sans perdre d'information, il faut donc trouver le moyens de perdre l'information la moins utile possible pour la restitution et l'interprétation par ordinateur. La solution passe par des moyens de compression adaptés aux images des documents [5], [6] et à la conservation des informations essentielles pour la lecture et le traitement informatique.

B. Etude de la compression JPEG2000

L'avenir de la compression d'images semble maintenant appartenir au nouveau standard du Joint Picture Expert Group, JPEG2000 [2]. Ce nouveau format se base sur une transformation en ondelettes, contrairement à JPEG qui est basée sur une transformée en cosinus discrets. Cette dernière est particulièrement bien adaptée pour les signaux présentant des caractéristiques périodiques, ce qui lui vaut d'ailleurs son succès dans les domaines du traitement du signal 1D (analyse de la parole, transmission de données...). Par contre, dans le domaine de l'image où la périodicité n'est pas le critère le plus caractéristique, la DCT n'apparaît pas l'approche la mieux adaptée. D'où la faible qualité de restitution de l'image de JPEG à des taux de compression significatifs. La décomposition en ondelettes utilisée par JPEG2000 s'appuie sur une approche pyramidale multi-résolution plus pertinente dans le cas de signaux 2D tels que les images (Figure 5).

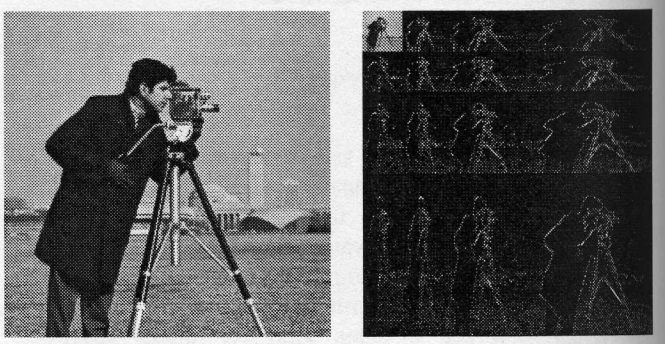


Figure 5: Décomposition en ondelettes

Compte tenu du fait que ce format n'est pas encore standardisé au moment de la rédaction de cet article, nous n'aborderons pas certaines des ses caractéristiques propres telles que la possibilité de définir des régions d'intérêt dans l'image qui peuvent être encodées dans une résolution plus élevée ou compressées sans pertes.

A la différence de la compression JPEG qui subit l'effet de bloc, la compression JPEG2000 donne des résultats locaux bien meilleur ; ainsi, la délocalisation des pixels due à JPEG n'apparaît plus avec JPEG2000 ; les contours des caractères s'en trouvent alors conservés, ce qui est particulièrement important pour traiter des documents textuels.

Par contre les images compressées grâce à JPEG2000 souffrent aussi d'un effet de flou sur l'image particulièrement accentué si le taux de compression est élevé. Cela se traduit alors par une perte de la dynamique du contraste de l'image (Figure 6). Ce phénomène rend alors plus difficile l'application de certains traitements tel que la binarisation.

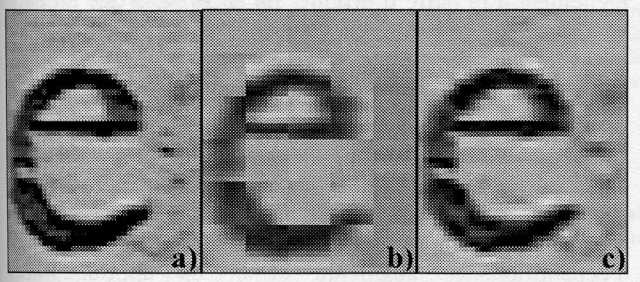


Figure 6: Compression à taux équivalent, a) image initiale, b) image compressée JPEG, c) image compressée JPEG2000

La Figure 7 montre la variation du rapport signal sur bruit (SNR) d'une image de document à 600dpi compressée avec JPEG et JPEG2000 à différents ratio de compression. Il apparaît clairement que JPEG2000 obtient un rapport signal sur bruit très nettement supérieur quel que soit le ratio de compression souhaité et plus particulièrement lorsque celui-ci est élevé. Ainsi, à qualité équivalente, il est possible avec JPEG2000 et à fort taux de compression d'obtenir un fichier pouvant être deux fois plus petit, voire même plus, qu'avec JPEG.

Globalement, la compression par DCT souffre du fait qu'elle a du mal à traiter des images hautes résolutions et que dans ce cas là, il est souvent plus efficace de sous-échantillonner l'image avant de la compresser. Avec cette manipulation, il est alors possible d'atteindre avec JPEG des performances proches de celles de JPEG2000.

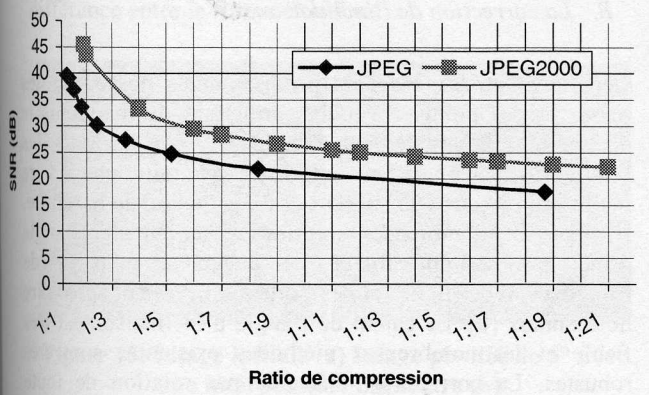


Figure 7:Comparaison du rapport signal sur bruit de JPEG et JPEG2000

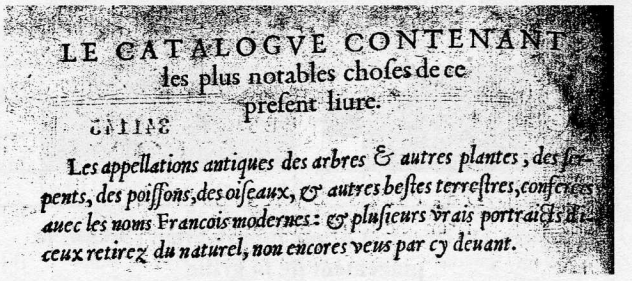
IV. LES TRAITEMENTS

Pour certaines applications telles que la réédition des ouvrages, la qualité de l'image sous sa forme brute n'est pas toujours satisfaisante. Les imperfections de celle-ci ne permettent pas d'obtenir la qualité souhaitée par l'utilisateur final.
Ces défauts peuvent être classés en deux grandes familles:
1- les défauts intrinsèques à l'ouvrage: avec le temps, la qualité du papier, l'acidité de l'encre, l'humidité du lieu de stockage, des tâches peuvent apparaître sur le papier.
2- les défauts liés à la numérisation: ce sont tous les défauts liés à la variation d'éclairage en particulier, à l'échantillonnage en pixels ou à la courbure de la page.

Nous avons cherché à réduire ces phénomènes en créant des algorithmes spécifiques. Nous allons faire une présentation de deux traitements différents parmi les plus employés, la binarisation et la correction d'inclinaison ; nous avons choisis de les présenter car ils ont des effets sur les documents qui doivent d'être maîtrisés.

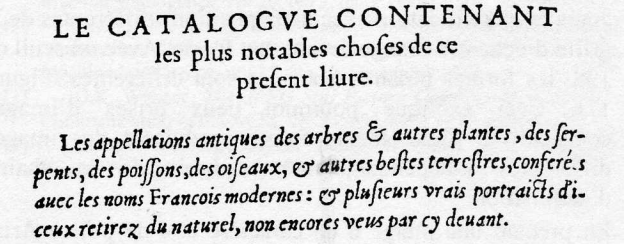
A. La binarisation

La binarisation est le passage d'une image en couleur ou définie par plusieurs niveaux de gris en image bitonale qui permet une classification entre le fond (image du support papier en blanc) et la forme (traits des gravures et des caractères en noir). La binarisation est un traitement irréversible qui détruit une grande part d'information contenue dans l'image ;c'est, en fait, le résultat d'une segmentation qui introduit une interprétation. C'est un traitement complexe qui, dans le cadre de documents, doit conserver à la fois tous les caractères et toutes les gravures sans toutefois récupérer trop de bruit (Figure 8 et Figure 9).


LE CATALOGVE CONTENANT
les plus notables choſes de ce
preſent liure.

Les appellations antiques des arbres & autres plantes, des ſer-
pents, des poiſſons, des oiſeaux, & autres beſtes terreſtres, conferé-s
auec les noms Francois modernes: & pluſieurs vrais portraicts d'i-
ceux retirez du naturel, non encores veus par cy deuant.

Figure 8: Binarisation à partir d'un seuillage global


LE CATALOGVE CONTENANT
les plus notables choſes de ce
preſent liure.

Les appellations antiques des arbres & autres plantes, des ſer-
pents, des poiſſons, des oiſeaux, & autres beſtes terreſtres, conferé-s
auec les noms Francois modernes: & pluſieurs vrais portraicts d'i-
ceux retirez du naturel, non encores veus par cy deuant.

Figure 9: Binarisation à partir d'un seuillage local

On peut trouver dans la littérature de très nombreux travaux concernant la binarisation d'un document [7][10]. Les plus simples utilisent l'histogramme de l'image (ou d'une portion) pour déterminer un seuil qui lui sera appliqué. Cette approche a l'avantage d'être extrêmement rapide mais la variation d'éclairage sur le document fait chuter la qualité de la binarisation (Figure 8). D'autres, telles que la maintenant classique méthode introduite par Niblack [7], utilisent une approche plus locale aux pixels afin de déterminer une valeur de seuil qui sera propre à chacun d'eux. Cette approche permet d'obtenir un résultat faiblement dépendant des variations de luminosité sur la page (Figure 9).

La binarisation peut être effectuée sur les images pour lesquelles cette segmentation est réalisable. Nous mesurons la qualité de la binarisation en fonction du nombre de caractères correctement isolés (pas de contacts entre caractères) et correctement segmentés (pas de rupture de traits). Mais même avec une segmentation parfaite, toute binarisation enlève une information qui peut être nécessaire à la reconnaissance. Par exemple la binarisation accentue l'effet du placement aléatoire de la grille d'échantillonnage qui provoque une quantification différente des valeurs des pixels suivant la position de la grille de digitalisation devant une forme [9]. Ceci affecte les contours des caractères de façon aléatoire et provoque une perte de la topologie des formes de caractères pour les images sous-échantillonnées.

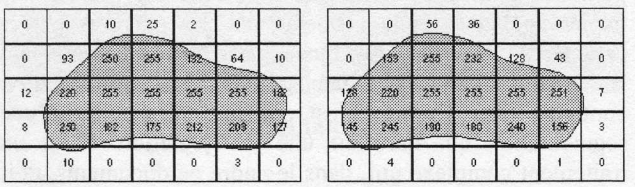


Figure 10: Quantification des pixels suivant le placement aléatoire de la grille

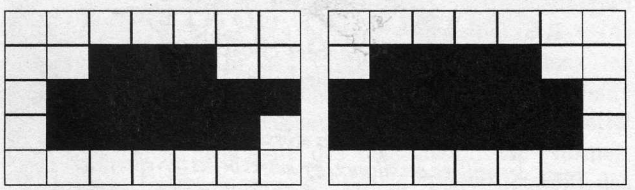


Figure 11: Résultat d'une binarisation suivant le placement de la grille

La Figure 10 montre les images à niveaux de gris obtenues après la quantification avec deux positions différentes de la grille d'échantillonnage devant une forme. Avec un seuil de 128, les formes binaires obtenues sont différentes (Figure 11). Ceci explique pourquoi deux prises d'images consécutives d'une même forme produisent des images différentes indépendamment du bruit de la chaîne d'acquisition.

En prenant une image d'un caractère «e» de police Arial issue d'une fonte vectorielle parfaite (Figure 12), numérisée sur une grille 10x10 pixels par simulation informatique reproduisant le fonctionnement d'un scanner, on obtient 139 formes binaires différentes de la même lettre. Sur ces 139 formes, nous avons comptabilisé 39 formes qui s'apparentent plus au chiffre «8», 25 formes qui s'apparentent à la lettre «C» et 14 à la lettre «B». Par conséquent moins de 44% des formes binaires ressemblent réellement à la lettre d'origine. Le chiffre marqué en haut de chaque image indique la fréquence d'apparition de cette forme binaire lors du déplacement de la grille. Ce qui indique que la forme de caractère «e» apparaît la plus fréquemment.

En conservant uniquement l'image binarisée, il est difficile de retrouver la forme d'origine. La numérisation pose le problème de la conservation des formes en passant du continu au discret. Cet effet devient plus important si le pas d'échantillonnage augmente et si on applique une binarisation. Il y a deux solutions simples pour ce problème, la première consiste à augmenter la résolution de façon à ce que chaque pixel élémentaire constitue une surface plus petite de la forme. La seconde solution passe par la conservation de l'image en niveaux de gris.

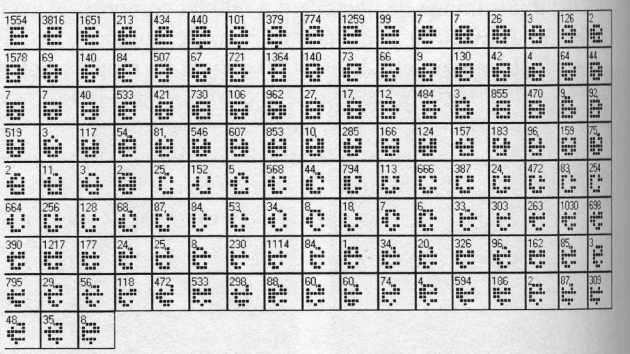
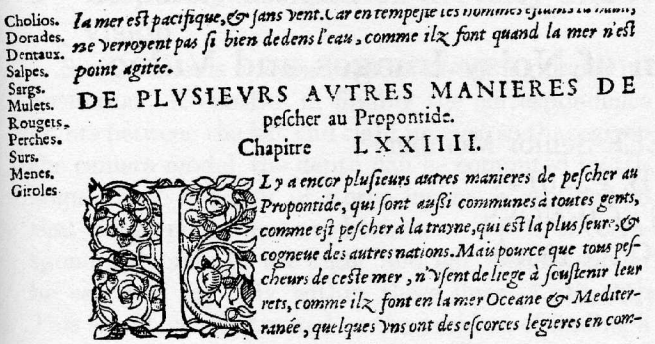


Figure 12: Résultat d'une binarisation suivant le placement de la grille

B. La correction de l'inclinaison

Généralement, les scanners professionnels possèdent des règles et des points d'appuis permettant de maintenir le document horizontalement face à la caméra. Mais les feuilles de documents peuvent laisser apparaître une légère inclinaison à cause du support et de la qualité de la reliure. Il existe de nombreuses solutions pour calculer l'angle d'inclinaison d'une image de document à partir de l'inclinaison des lignes et des traits qui devraient apparaître horizontaux [8]. Le calcul de l'angle d'inclinaison est très fiable et les nombreuses méthodes existantes sont très robustes. La correction s'effectue par rotation de toute l'image en fonction de l'angle d'inclinaison.


Cholios. Dorades. Dentaux. Salpes. Sargs. Mulets. Rougets. Perches. Surs. Menes. Giroles.

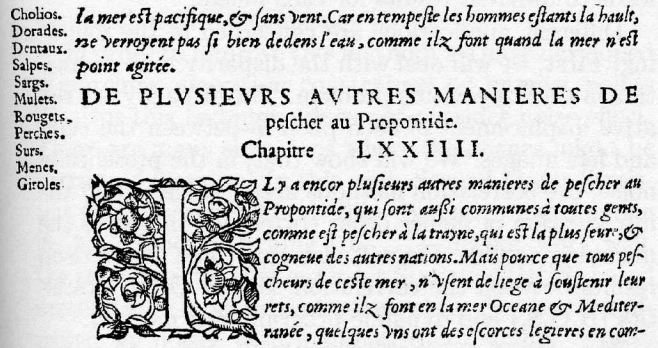
La mer eſt pacifique, & ſans vent. Car en tempeſte les hommes eſtants la hault, ne verroyent pas ſi bien dedens l'eau, comme ilz font quand la mer n'eſt point agitée.

DE PLVSIEVRS AVTRES MANIERES DE peſcher au Propontide.

Chapitre LXXIIII.

Ly a encor pluſieurs autres manieres de peſcher au Propontide, qui ſont auſſi communes à toutes gents, comme eſt peſcher à la trayne, qui eſt la plus ſeure & cogneue des autres nations. Mais pource que tous peſcheurs de ceſte mer, n'vſent de liege à ſouſtenir leur rets, comme ilz font en la mer Oceane & Mediterranée, quelques vns ont des eſcorces legieres en com-

Figure 13: Image d'origine inclinée


Cholios. Dorades. Dentaux. Salpes. Sargs. Mulets. Rougets. Perches. Surs. Menes. Giroles.

La mer eſt pacifique, & ſans vent. Car en tempeſte les hommes eſtants la hault, ne verroyent pas ſi bien dedens l'eau, comme ilz font quand la mer n'eſt point agitée.

DE PLVSIEVRS AVTRES MANIERES DE peſcher au Propontide.

Chapitre LXXIIII.

Ly a encor pluſieurs autres manieres de peſcher au Propontide, qui ſont auſſi communes à toutes gents, comme eſt peſcher à la trayne, qui eſt la plus ſeure & cogneue des autres nations. Mais pource que tous peſcheurs de ceſte mer, n'vſent de liege à ſouſtenir leur rets, comme ilz font en la mer Oceane & Mediterranée, quelques vns ont des eſcorces legieres en com-

Figure 14: Image redressée par rotation

Afin d'améliorer la qualité optique du résultat, on utilise lors de la rotation une interpolation des pixels en fonction de ses voisins [12]. Dans l'exemple de la Figure 15, nous avons fait pivoter l'image initiale (a) d'un angle de 5° (b) en appliquant une interpolation bilinéaire. Comme chaque point est calculé selon une moyenne pondérée de ses voisins, cette interpolation effectue un filtrage passe-bas de l'image et élimine, de fait, toutes les hautes fréquences surtout présentes sur les contours des caractères. Ce phénomène provoque une détérioration des performances des algorithmes d'analyse, car il réduit très nettement la différence entre la forme et le fond.

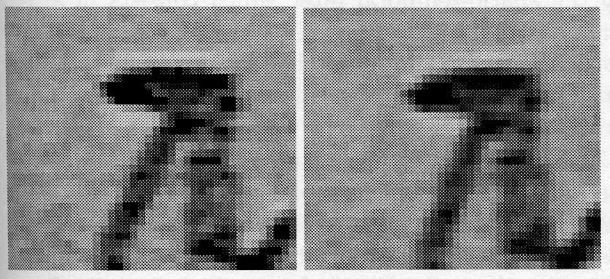


Figure 15: Image initiale (a) Image pivotée de 5° avec une interpolation bilinéaire (b)

De même, une interpolation au plus proche voisin conservera mieux les hautes fréquences de l'image ; elle altérera, en revanche, très nettement les contours.

V. CONCLUSION

L'expérience acquise lors du projet DEBORA nous a guidé pour présenter différents problèmes liés à la numérisation des documents textuels. Nous avons résumé les recommandations sur la qualité minimale requise des images pour les traitements informatiques présents et futurs. Une étude sur les effets de la compression avec perte JPEG et JPEG2000 a été menée car ce mode est encore trop souvent utilisé pour comprimer les images de documents. Nous avons montré que ces compressions affectent de façon significative la qualité visuelle des images ainsi que la reconnaissance de caractères. La compression JPEG et, à moindre mesure, JPEG2000 ne sont donc pas adaptées aux images de documents qui présentent des caractéristiques particulières. Nous avons mis en évidence les problèmes de la binarisation ainsi que la mesure de sa qualité du point de vue de la reconnaissance. Mais cette mesure de qualité ne tient pas compte des déformations dues au déplacement aléatoire de la grille d'échantillonnage. Nous avons enfin décrit les effets d'une correction géométrique de base nécessaire à la segmentation des images. Ce type de correction affecte aussi les formes des caractères car elles emploient toutes une interpolation. Compte tenu du nombre élevé de bases documentaires déjà numérisées, se pose maintenant le problème de la restauration des images et des informations qui ont été perdues lors de la numérisation.

VI. RÉFÉRENCES

[1] *Information technology – Digital compression and coding of continuous-tone still images – requirements and guidelines* Recommendation T.81, ITU/CCITT, 09/92

[2] *JPEG2000, PartI Final Comittee Draft Version 1.0,* ISO/IEC JTC1/SC29 WG1, 16 mars 2000, 205 pages

[3] H.S. Baird, *Document image defect models*, In H.S. Baird, H. Bunke, and K. Yamamoto, editors, Structured Document Image Analysis, Springer Verlag, 1992, pages 546-556

[4] R. Bouché, H. Emptoz et al, *DEBORA*, projet européen n° LB5608A, http://debora.enssib.fr, juin 2000

[5] F. LeBourgeois, E. Trinh, *Compression des images de documents par redondance de formes et compensation,* Rapport Interne RFV n° 2-01, INSA Lyon, février 2001

[6] Omid E. Kia, *Document Image Compression and Analysis*, Thèse sci. : University of Maryland, 1997, 191 p

[7] Niblack W, *An introduction to image processing*" Englewood Cliffs, N.J., Prentice Hall, p 115-116

[8] James R. Parker, *Algorithms for Image Processing and Computer Vision*, Wiley, 432 pages, 1996

[9] P. Sarkar, G. Nagy, J. Zhou and D. Lopresti, *Spatial Sampling of Printed Patterns*, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol 29, No 3, mars 1998

[10] Oivind Due Trier, Anil K. Jain, *Goal Directed Evaluation of Binarization Methods*, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol 17, No 12, Decembre 1995

[11] G.K. Wallace, *The JPEG Still Picture Compression Standard*,. Communication of the ACM, vol 34, n°4, avril 1991, pages 31-45.

[12] George Wolberg, *Digital Image Warping*, IEEE Computer Society Press Monograph, 1990, 318p