الجمهوريـة الجزائـريـة الديمقراطيـة الشعبيـة
République Algérienne Démocratique et Populaire
وزارة التعليـم العالـي والبحـث العلمـي
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Nº Réf :……

Centre Universitaire
Abd elhafid Boussouf Mila

Institut des sciences et de la technologie **Département de Mathématiques et Informatiques**

Mémoire préparé en vue de l'obtention du diplôme de Master

En : Informatique

Spécialité : Sciences et Technologies de l'Information et de la Communication (STIC)

Analyse et reconnaissance automatique des documents numérisés

Préparé par : - **Khenfi Moussaab**

Soutenue devant le jury :

Encadré par : M. Boulmerka Aissa…….....M.A.A
Président : Mme. Yassaadi Sabrina……...M.A.A
Examinateur : M. Dib Abderrahim..……..……M.A.A

Année universitaire : 2015/2016

Dédicaces

Au nom du dieu le clément et le miséricordieux
*louange à **ALLAH** le tout puissant.*
Je dédie ce modeste travail :
À mes chers parents, pour leur soutien, leur patience et leur Amour
Mes frères et sœurs qui n'ont cessé d'être pour moi des exemples de courage et de générosité
À tous mes professeurs de CUM
À tous mes collègues
À tous mes amis
En particulier à la mémoire de mon ami proche Daikh Aissa

Moussaab

Remerciements

En tout premier lieu, je remercie le bon Dieu, tout puissant, de m'avoir donné la force pour survivre, ainsi que l'audace pour dépasser toutes les difficultés.

Je remercie vivement l'encadreur ***Monsieur Aissa Boulmerka*** *qui n'a épargné aucun effort pour que ce travail prenne forme. Je le remercie pour l'attention particulière qu'il a portée à ce travail et de la confiance qu'il m'a accordée tout au long de ce parcours.*

Mes remerciements s'adressent également à tous les professeurs et docteurs de CUM qui nous ont enseigné. Leurs enseignements ont révolutionné et élargi notre connaissance scientifique, nous pensons particulièrement aux professeurs : ***Ben Abderrahmane Fatiha, Yassaadi Sabrina, Nerdjess Bouchemal, Dib Abderrahim, Bilal douas, Ben Cheikh Madjid, Bourideh Adel.***

Mes reconnaissances s'adressent également au directeur des Transmissions National de la Wilaya de Mila ***Monsieur Mohammed Bougeffa*** *pour son soutien et son aide.*

Je serais traité d'ingrat si je laisse passer inaperçu le soutien moral de nos collègues du Service Maintenance : ***Riade Siari***, ***Abdelhak Tir, Jamal Boulekroune***, ***Farid Meziani***, ***Sebti Maiche***, ***Foul Daoud***, ***Facih Abdelhak.***

Sans oublier tout le personnel de la Direction des Transmissions National de la Wilaya de Mila pour leur aide et leur soutien.

J'adresse mes plus sincères remerciements à tous ma famille, mes proches et amis, qui m'ont toujours soutenue et encouragée au cours de la réalisation de ce mémoire.

Je tiens à remercier finalement toute personne qui a, de près ou de loin, contribué d'une manière ou d'une autre au succès de ce travail, et spécialement ceux dont les noms ne sont pas mentionnés, mais qui sont présents dans mon esprit et dans mon cœur.

Table des matières

Liste des figures **iv**

Liste des tableaux **vii**

Liste des abréviations, sigles et acronymes **viii**

Résumé **ix**

1 Introduction Générale **1**

2 Prétraitement **4**

2.1 Introduction 4

2.1.1 Image numérique 4

2.2 Le processus de traitement des documents 5

2.3 Définition du prétraitement 5

2.3.1 Les origines du problème de qualité d'image 6

2.3.2 Méthodes du prétraitement 7

2.3.3 Lissage 8

2.3.4 Debruitage 8

2.3.5 Effet miroir et inversion 9

2.3.6 Réalignement automatique de l'image 9

2.4 Type des prétraitement 9

2.4.1 Amélioration du contraste 9

2.4.2 Filtrage 11

2.4.3 Amélioration des couleurs 12

2.5 Binarisation 14

2.5.1 Seuillage global 15
2.5.2 Seuillage local 18
2.5.3 Seuillage hybride 20
2.5.4 Seuillage par combinaison 20
2.6 Morphologie 21
2.7 Conclusion 24

3 Analyse de la structure des documents numériques 25
3.1 Introduction 25
3.2 Définition d'un document 25
3.2.1 Structure physique 26
3.2.2 Structure logique 26
3.3 Segmentation 27
3.3.1 Définition 27
3.3.2 Les étapes de la segmentation 27
3.3.3 Les classes de mise en page 28
3.4 Les méthodes de la segmentation et de l'analyse du document . . . 29
3.4.1 Les méthodes descendantes 29
3.4.2 Les méthodes ascendantes 33
3.4.3 Les méthodes mixtes 36
3.5 Conclusion 37

4 Outils pour l'analyse et la reconnaissance des documents 38
4.1 Introduction 38
4.2 Tesseract 38
4.3 Les Outils du labo Prima 39
4.3.1 TesseractToPAGE 1.3 39
4.3.2 Aletheia 43
4.3.3 PAGE Viewer 50
4.4 Leptonica 50
4.5 Conclusion 52

5 Implémentation et réalisation 53
5.1 Introduction 53

5.2 Les outils de développement 53
5.3 Les fonctionnalités de notre application 56
5.3.1 Les fonctions fichier (File) 57
5.3.2 Les fonctions de prétraitement 58
5.3.3 Les fonctions d'analyse des documents 63
5.3.4 Exemple d'un déroulement des fonctions sur un document . 69
5.4 Conclusion 71

6 Conclusion et Perspectives **72**

Bibliographie **74**

Liste des figures

1.1 Document avant et après le prétraitement. 2
1.2 Exemple de l'analyse et de la structure d'un document. 3

2.1 Image pixellisée 5
2.2 Le prétraitement d'un document. 6
2.3 La restauration d'un document. 7
2.4 Contraste amélioré d'un document. 9
2.5 Egalisation d'un histogramme. 11
2.6 Application du filtre median. 12
2.7 Modes de codage des couleurs. 12
2.8 Conversion des espaces colorimétriques. 13
2.9 Texte binarisé. 14
2.10 Binarisation d'un document par un seuillage global. 16
2.11 Document binarisé par la méthode d'Otsu 17
2.12 Les niveaux de gris d'un document. 18
2.13 Document binarisée par la méthode de Niblack. 19
2.14 Document binarisé par la méthode de Sauvola. 20
2.15 Exemple sur le déroulement de l'érosion. 22
2.16 Exemple sur le déroulement de la dilatation. 22
2.17 Document dilaté 23
2.18 La squelettisation d'un texte. 24

3.1 Structure physique d'un document. 26
3.2 Structure logique d'un document. 27
3.3 Segmentation d'un document. 28

3.4 Les classes de mise en page. 28
3.5 Approches descendante et ascendante. 29
3.6 Méthode d'arbre X-Y 30
3.7 Profiles de projection. 30
3.8 Application de la méthode RLSA. 32
3.9 Application de la méthode RLSO. 33
3.10 Application de la méthode DOCSTRUM. 35
3.11 Diagramme Area Voronoi. 36

4.1 Image représentatif de TesseractToPAGE. 39
4.2 L'effet du Layout analysis. 41
4.3 L'effet du Layout analysis et OCR on niveau region 42
4.4 L'effet de l'analyse de Layout et OCR au niveau region, text, ligne, et mot. 42
4.5 Capture d'écran de l'outil Aletheia. 43
4.6 Les opérations d'images dans Aletheia. 44
4.7 Border et Espace d'impression dans Aletheia. 45
4.8 Régions et mise en page dans Aletheia. 45
4.9 Création des régions dans Aletheia. 46
4.10 Modification des régions dans Aletheia. 47
4.11 Superposition de texte dans Aletheia. 47
4.12 Lignes de texte dans Aletheia. 48
4.13 Les mots et les Glyphes dans Aletheia. 49
4.14 Format XML dans Aletheia. 49
4.15 Capture écran de l'outil PAGE Viewer. 50
4.16 Génération des texts blocks. 51
4.17 Génération des texts lines. 52

5.1 Capture d'écran de l'environnement MATLAB. 54
5.2 La structure des menus de notre application. 56
5.3 Capture d'écran de la fonction ouvrir. 57
5.4 Capture d'écran de la fonction sauvegarder l'image transformé. . . . 58
5.5 Capture d'écran de la fonction transformation en niveau de gris. . . 59

5.6 Capture d'écran de la fonction réduction de bruit avec l'image originale. 59
5.7 Capture d'écran de la fonction réduction de bruit avec l'image résultante. 60
5.8 Capture d'écran de la fonction égalisation de l'histogramme avec l'image originale. 60
5.9 Capture d'écran de la fonction égalisation de l'histogramme avec l'image résultante. 61
5.10 Capture d'écran de la fonction binarisation d'Otsu. 61
5.11 Capture d'écran de la fonction binarisation de Sauvola. 62
5.12 Capture d'écran de la fonction binarisation de Niblack. 63
5.13 Capture d'écran de la fonction analyser les régions. 63
5.14 Capture d'écran de la fonction analyser les lignes. 64
5.15 Capture d'écran de la fonction analyser les mots. 65
5.16 Capture d'écran de la fonction analyser les glyphes. 65
5.17 Capture d'écran de la fonction Reconnaissance des caractères (regions). 66
5.18 Capture d'écran de la fonction Reconnaissance des caractères (mots). 66
5.19 Capture d'écran de la fonction Save XML Files (correction des erreurs). 67
5.20 Capture d'écran de la fonction sauvegarder le fichier XML (sauvegarder). 67
5.21 Capture d'écran de la fonction voir le fichier XML (ouvrir). 68
5.22 Capture d'écran de la fonction voir le fichier XML (appliquer les opérations). 68
5.23 Capture d'écran de la fonction lecture et affichage d'un document . 69
5.24 Capture d'écran de la fonction suppression du bruit du document utilisé. 69
5.25 Capture d'écran du document résultant. 70
5.26 Capture d'écran des fonctions analyse et reconnaissance. 70

Liste des tableaux

2.1 Les techniques de Seuillage 15

Liste des abréviations, sigles et acronymes

OCR Optical Character Recognition

RVB Rouge Vert Bleu

CIE Commission Internationale de l'Éclairage

CMJN Cyan, Magenta, Jaune, Noir

CC Color Centroid

RLSA Run Length Smoothing Algorithm

RLSO Run-Length Smoothing with OR

OSD On-Screen Display

XML Extensible Markup Language

ALTO Analyzed Layout and Text Object

Résumé

L'analyse et la reconnaissance automatique des documents numérisés est un domaine très vaste. Se caractérisent par divers problèmes pertinents tels que la présence des bruits, la variation des documents, la langue, etc.
Actuellement, on utilise les documents numérisés pour accomplir plusieurs tâches d'information, cependant des grands problèmes entourent les documents mal scannés et la mauvaise écriture car cela engendre une perte d'information et une mauvaise expérience de lecture. La question qui se pose c'est comment régler ces problèmes et comment extraire une bonne information ?
A l'aide des outils tel que Prima et Tesseract, nous avons suggéré dans ce mémoire une solution aux problèmes déjà cités. Cette méthode se base sur le prétraitement et la préparation des documents numériques, puis sur l'analyse et la segmentation de la structure des documents numériques, et enfin une reconnaissance des caractères. Nous avons fait une étude bibliographique sur les différents phénomènes rencontrés dans le processus d'analyse et de reconnaissance de documents numérisés.
On a utilisé une méthode caractérise par la préservation maximale du texte, une analyse et une segmentation des documents et enfin une reconnaissance des documents avec moins de faute. Le résultat final obtenu par cette méthode est un document utile, lisible et capable d'extraire de bonnes informations. Les résultats obtenus sont très satisfaisantes.

Mots Clés :Prétraitement, Analyse et segmentation de document, Reconnaissance de caractère (OCR).

Abstract

The analysis and automatic recognition of scanned documents is a very broad area. It is characterized by various relevant issues such as the presence of noise, the change of documents, language, ect.
Currently, we use scanned documents to perform several tasks, however the big problems is the poorly scanned documents and bad writing, causing a loss of information and a bad reading experience. The question is how to address these issues and how to extract a good information ?
Using tools like Prima and Tesseract, we suggested in this manuscript a solution to the problems already mentioned. This method is based on pre-processing and preparation of digital documents and the analysis and segmentation of the structure of digital documents, and finally the character recognition. We made a bibliographic study on the different phenomena encountered in the process of analysis and recognition of scanned documents.
We used a method characterized by the maximum preservation of the text, analysis and segmentation of documents and finally a recognition of documents with less fault. The final result obtained by this method is a useful document, readable and capable of extracting good information. The results are very satisfactory.
Keywords : Preprocessing, Document analysis and segmentation, Optical Character Recognition (OCR).

Chapitre 1

Introduction Générale

Le domaine du traitement automatique des documents a connu une avance technologique considérable dans tous les aspects causant un rôle important dans plusieurs applications informatiques, en raison de la propagation progressive des documents numériques dans tous les domaines.
Cela a commencé avec une tentative de lecture automatique des textes imprimés. Avant la naissance des ordinateurs numériques, avec les capacités émergentes de simulation et de calcul scientifique, un développement produit et ça continue sur divers sujets tels que l'amélioration et la structuration des documents, l'analyse et la mise en page, ainsi que la reconnaissance des caractères manuscrits, la récupération et l'extraction des données.

Des applications réelles telles que Tesseract et Aletheia ont ajouté une grande valeur pour le domaine du traitement automatique des documents. En fait, ce type d'applications a permis aux techniques de traitement automatique des documents d'être connues par un grand communauté mondiale .

Intitulé "Analyse et reconnaissance automatiques des documents numérisés", ce mémoire tend ainsi à démontrer qu'une application peut reconstituer et analyser le contenu d'un document à partir de son image. Pour cela, nous avons fait une étude bibliographique sur les différents problèmes rencontrés dans le processus d'analyse et reconnaissance de documents numérisés.

Ces problèmes doivent être traités d'abord par un prétraitement et une préparation des documents, ensuite une analyse de la structure des documents est appliquée, et enfin la reconnaissance des caractères permet d'extraire les caractères du texte. La première partie de ce mémoire est consacrée au prétraitement. C'est là où on doit améliorer la qualité des documents et supprimer tous les bruits et les tâches (voir Figure 1.1).

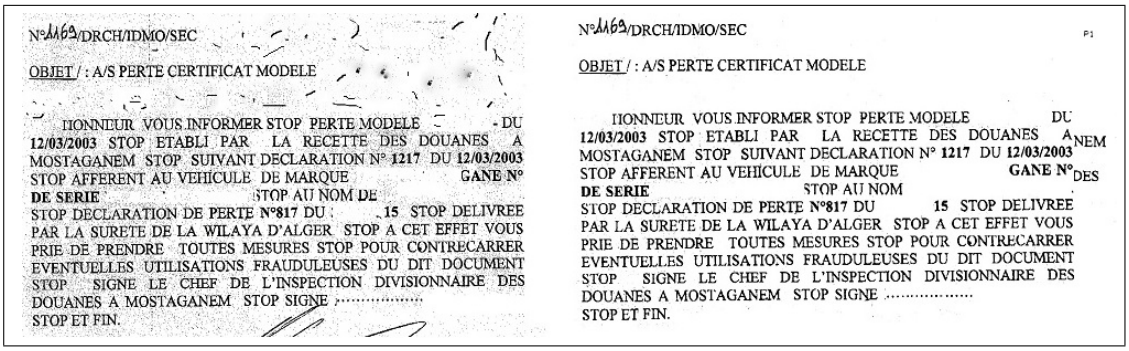
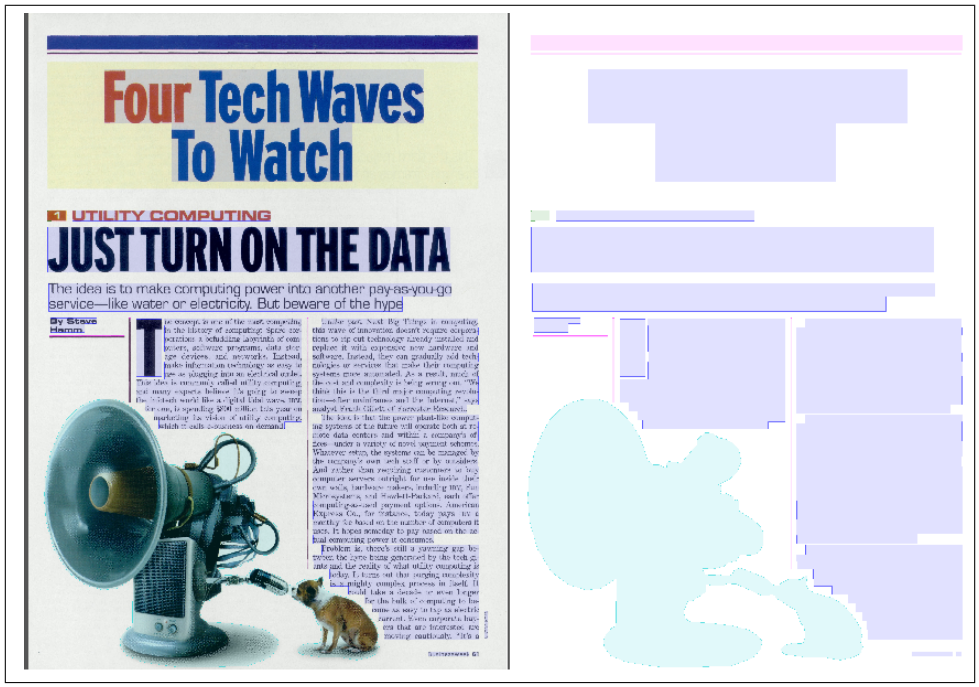


FIGURE 1.1 – Document avant et après le prétraitement.

La deuxième partie est dédiée à l'analyse de la structure des documents numériques, elle permet, en effet, de décomposer un document donné en régions constitutives et de saisir leurs rôles fonctionnels (voir Figure 1.2).


Four Tech Waves To Watch

UTILITY COMPUTING

JUST TURN ON THE DATA

The idea is to make computing power into another pay-as-you-go service—like water or electricity. But beware of the hype

By Steve Hamm

FIGURE 1.2 – Exemple de l'analyse et structure d'un document.

Puis, on décrira, dans la troisième partie, les outils pour l'analyse et la reconnaissance des documents. Ces outils nous permettent de faciliter la tâche d'implémentation et de la réalisation de projet.

En quatrième chapitre, on présentera l'application implémentée et réalisée durant la période de ce travail.

Et on terminera ce travail avec une conclusion et des perspectives.

Chapitre 2

Prétraitement

2.1 Introduction

L'acquisition des documents est le processus d'obtention d'une image électronique qui est sur papier. Dans la plupart des cas, un scanner est utilisé, mais il ne conserve pas toujours le document en état sain. Les opérations de prétraitement dans l'analyse d'images de documents transforment l'image originale en une image améliorée plus appropriée pour une analyse ultérieure. Dans ce chapitre, nous allons détailler quelques opérations de prétraitement. Commençons tout d'abord par connaître quelques notions :

2.1.1 Image numérique

Définition 1 : c'est une matrice de $X * Y$ pixels [1] correspondant à l'échantillonnage et la quantification d'un signal acquis avec une caméra.

Chaque pixel est associé à un niveau de gris n ou des niveaux de composantes couleurs codés sur N bits et qui représentent respectivement le niveau de luminosité ou de couleur de la zone [35].

Chaque pixel est localisé par ses coordonnées x et y dans l'image.

1. l'unité de base permettant de mesurer la définition d'une image numérique matricielle.

Définition 2 : une image est une matrice (tableau) de point et chaque point est un "pixel" (voir Figure 2.1), chaque pixel contient une couleur.

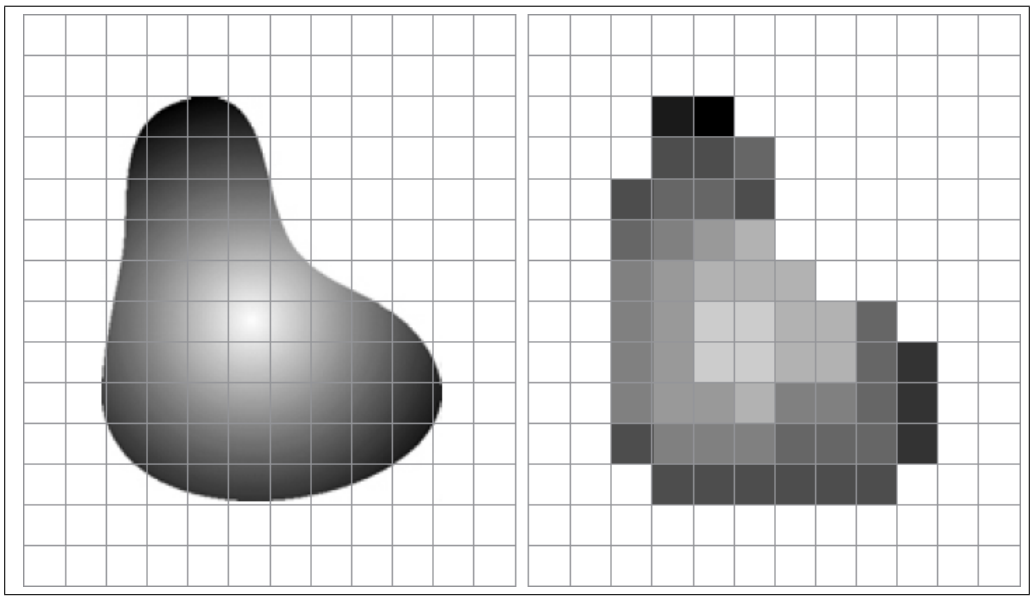


FIGURE 2.1 – Image pixellisée [12].

Définition 3 : l'image numérique est un ensemble d'informations, ces informations sont structurées afin d'avoir une signification pour l'oeil humain.

2.2 Le processus de traitement des documents

Il existe trois types de traitement des documents scannés qu'on peut regrouper dans trois catégories principales : (i) le prétraitement, (ii) la segmentation des documents, et (iii) la reconnaissance des caractères.

2.3 Définition du prétraitement

C'est un ensemble d'opérations consistant à rendre des données brutes en données pouvant subir une analyse spécifique [4].

Donc, le prétraitement consiste à améliorer l'image par la clarté, le contour, et éliminer les informations indésirables.
Voilà un exemple sur le prétraitement d'image (voir Figure 2.2).

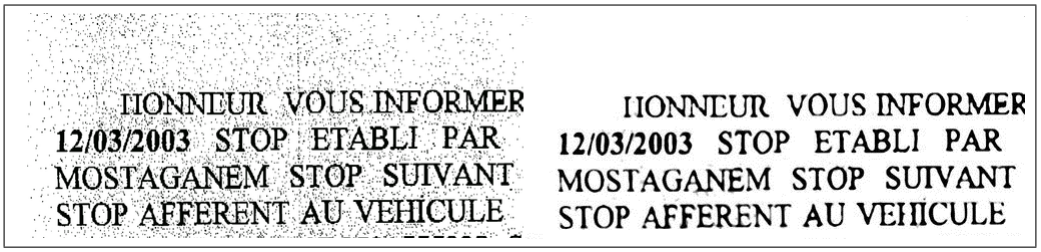


FIGURE 2.2 – Le prétraitement d'un document.

2.3.1 Les origines du problème de qualité d'image

Il existe plusieurs problèmes qui affectent la qualité de l'image. On peut citer, parmi ces problèmes :

2.3.1.1 La nature de la scène

On cite les nuages, la poussière dans les scènes industrielles, ou le brouillard pour les scènes routières, etc...

2.3.1.2 Le contexte d'acquisition

Les sur/sous illumination, la perturbation des capteurs, le bougé, ou encore les objets en mouvement.

2.3.1.3 La qualité du capteurs

Le capteur mal réglé, ou de mauvaise qualité (distorsion de la gamme des niveaux de gris ou en flou)

2.3.2 Méthodes du prétraitement

2.3.2.1 Restauration d'images

La restauration d'image est le résultat d'une ancienne image remise à neuf par un infographiste à l'aide d'un ordinateur et d'un logiciel d'image [34].

Exemple : une image déchirée, grafignée, décolorée, peut avec ce procédé, être rendue à son état d'origine.

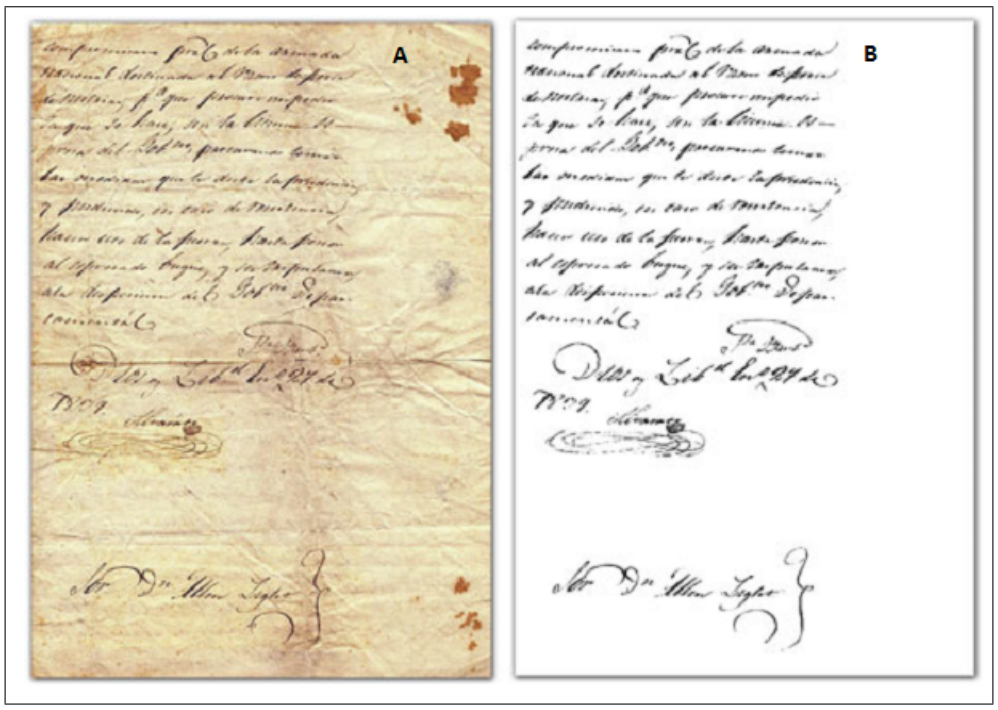


FIGURE 2.3 – Restauration d'un document. A :document originale, B :document restauré.

2.3.2.2 Amélioration d'images

L'amélioration a pour but de satisfaire l'oeil de l'observateur humain. L'oeil humain est essentiellement sensible aux forts contrastes [34]. C'est pourquoi les techniques d'amélioration tentent d'augmenter ceux-ci dans le but d'accroître la séparabilité des régions composant un document.

2.3.2.3 Seuillage d'image

Le seuillage d'image est utilisé pour créer une image comportant uniquement deux valeurs, noir ou blanc.
Le seuillage est divisé en trois types :

Global : un seuil pour toute l'image.

Local : un seuil est appliqué sur chaque partie de l'image.

Adaptatif : un seuil s'ajustant selon les parties de l'image.

Seuillage de base (2 classes) :
Si valeur (pixel) $>$ seuil alors valeur (pixel) $= 1$
Si valeur (pixel) $<$ seuil alors valeur (pixel) $= 0$
Donc ce résultat du seuillage est une image binaire qui contient des pixels avec valeurs 0 ou 1.

2.3.3 Lissage

On appelle "lissage" (parfois débruitage ou filtre anti-bruit) l'opération de filtrage visant à éliminer le bruit dans une image.
L'opération de lissage spécifique consistant à atténuer l'effet d'escalier produit par les pixels en bordure d'une forme géométrique est appelée anti-crénelage (en anglais anti-aliasing).

2.3.4 Debruitage

Lors de la numérisation des documents de qualité médiocre, il se peut que vous obteniez des images comportant beaucoup de (bruit), c'est-à-dire beaucoup de points et de traces. Ces traces, lorsqu'elles apparaissent près de lettres ou de nombres, peuvent affecter la qualité de l'OCR [2]

Cette fonctionnalité de nettoyage supprime ce bruit. La taille des traces à supprimer peut-être définie par l'utilisateur.

2. Optical Character Recognition.

2.3.5 Effet miroir et inversion

Une option disponible permet de réfléchir l'image préparée sur son axe vertical. Il est également possible d'inverser les couleurs de l'image préparée.

2.3.6 Réalignement automatique de l'image

Il s'agit d'une fonction d'imagerie des documents indispensable qui est appliquée aux documents numérisés nécessitant un réajustement au niveau de l'alignement des images. Il existe plusieurs méthodes de réalignement d'images : par paires de carrés noirs, par lignes ou par lignes de texte.

2.4 Type des prétraitement

Pour réaliser un prétraitement fiable et efficace on a utilisé une approche basée sur la qualité des documents et qui doit assurer l'information. Voici quelques types d'approches de prétraitement que nous avons utilisé : amélioration du contraste, le filtrage, amélioration des couleurs.

2.4.1 Amélioration du contraste

Lorsque la quantité de lumière incidente est faible ou bien les surfaces observées présentent des teintes voisines (en couleur comme en noir et blanc), le contraste peut être insuffisant. L'étirement d'histogramme et son égalisation peuvent améliorer grandement la lisibilité de l'image [6].

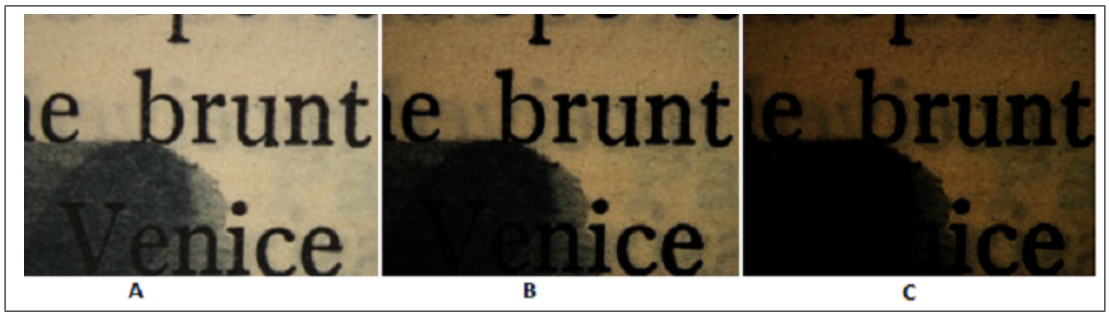


FIGURE 2.4 – Contraste amélioré d'un document avec plusieurs échantillons.

2.4.1.1 Techniques

2.4.1.2 Transformation linéaire

On étire la dynamique du document en restructuration les niveaux de gris entre 0 et 255.

$$\boxed{I'(i,j) = \frac{255}{max-min}(I(i,j) - min)}$$

2.4.1.3 Transformation linéaire avec saturation

Définie par 2 seuils : $\boxed{\text{imin} < \text{Smin} < \text{Smax} < \text{imax.}}$

Effet : rehaussement du contraste des niveaux vérifiant Smin $<$ i $<$ Smax et saturation

- à 0 des niveaux vérifiant imin $<$ i $<$ Smin.
- à 255 des niveaux vérifiant Smax $<$ i $<$ imax.

2.4.1.4 Transformation non-linéaire(gamma)

Il corrige les défaut selon γ :

$$\boxed{I' = 255(\frac{i}{255})^{1/\gamma}}$$

$1/\gamma < 1$ éclaircit principalement les parties foncées.
$1/\gamma > 1$ assombrit principalement les parties foncées.

2.4.1.5 Egalisation de l'histogramme

C'est une méthode d'ajustement du contraste qui vise à égaliser l'histogramme d'une image afin que les niveaux de gris soient également répartis .

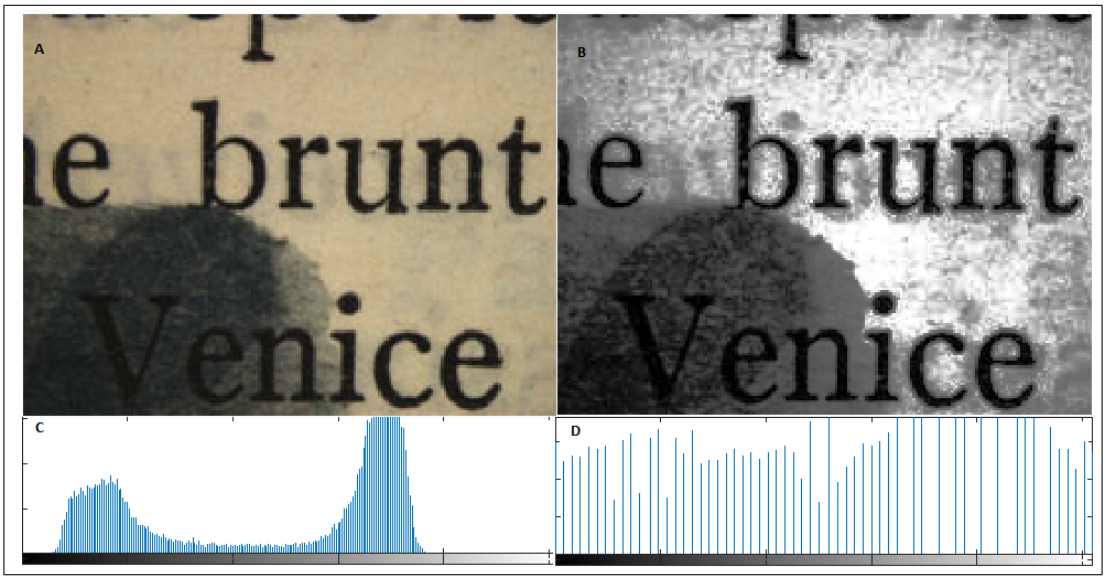


FIGURE 2.5 – L'opération d'égalisation d'un histogramme. A :document originale, B :document après égalisation, C :histogramme du document originale, D :histogramme du document égalisé.

2.4.2 Filtrage

2.4.2.1 Définition de filtrage

Le filtrage consiste à remplacer un pixel par une valeur qui est à proximité du pixel afin d'enlever les composantes indésirables de l'image.
Les buts du filtrage sont :

- Lissage d'image.
- Atténuation du bruit [3].
- Accentuation des discontinuités (contours).

2.4.2.2 Le Filtre Médian

Il permet d'atténuer et réduire le bruit dans une image. Dans ce cas de figure, le filtre médian donne des résultats satisfaisants, mais il faut l'utiliser avec précaution

3. un signal (parasite) qui s'ajoutent de façon aléatoire.

puisqu'il peut causer une dégradation de la qualité des documents dans certains cas.

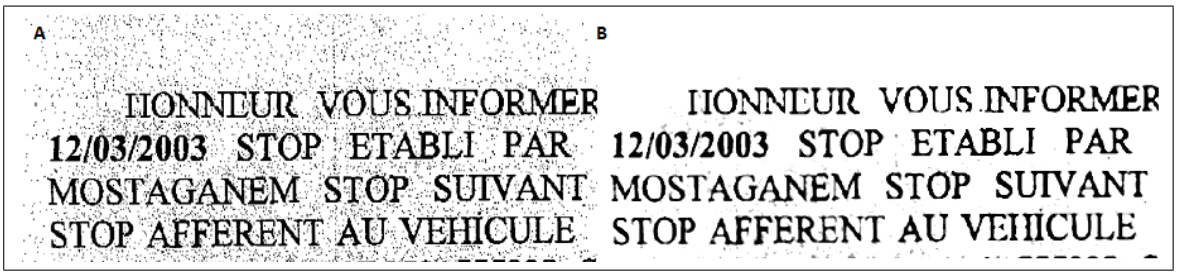


FIGURE 2.6 – Application du filtre median. A :document original, B :document après utilisation du filtre médian.

2.4.3 Amélioration des couleurs

2.4.3.1 Modes de codage des couleurs

L'espace colorimétrique rouge, vert, bleu (RVB). Cet espace est un mélange des trois composantes R, V, B [27].
Il existe aussi le CIE, le CIE LAB et CMJN (voir Figure 2.7).

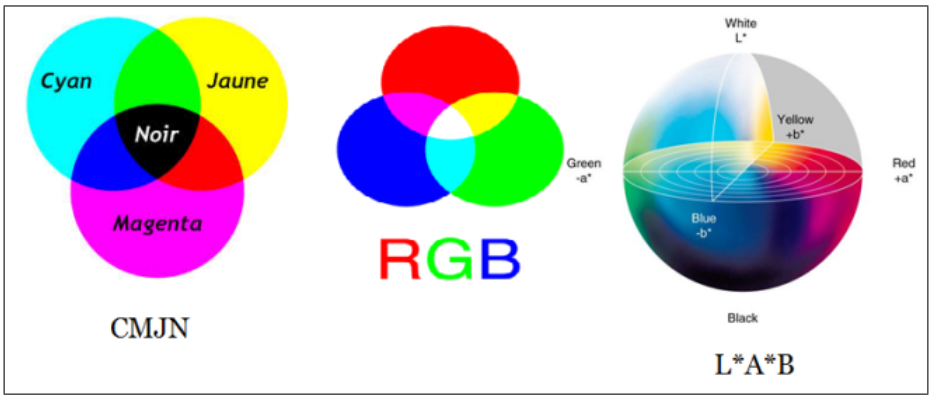


FIGURE 2.7 – Modes de codage des couleurs.

La conversion entre les espaces colorimétriques se passe comme suit :

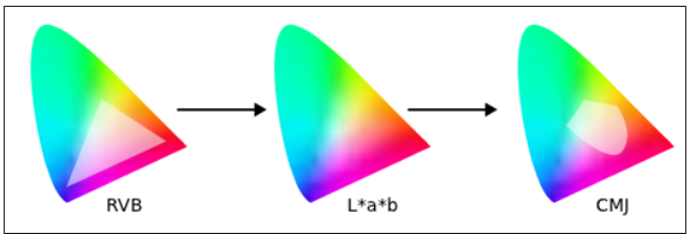


FIGURE 2.8 – Conversion des espaces colorimétriques.

2.4.3.2 Conversion Image Couleur en niveau de gris

Il existe une méthode simple pour convertir une image couleur en niveau de gris, et cela en calculant la moyenne des trois composantes RGB et d'utiliser cette valeur pour chacune des composantes [27].

$$\boxed{Gris = \frac{(Rouge+Vert+Bleu)}{3}}$$

Il existe une autre méthode selon La C.I.E (Commission Internationale de l'Éclairage) propose, de caractériser l'information de luminance (la valeur de gris) d'un pixel par deux formules :

Dans sa recommandation 709, qui concerne les couleurs (Vrais) ou naturelles :

$$\boxed{Gris = 0.2125.Rouge + 0.7154.Vert + 0.0721.Bleu}$$

Dans sa recommandation 601 pour les couleurs non-linéaires, c'est-à-dire avec correction du gamma (image vue à partir d'un écran vidéo) :

$$\boxed{Gris = 0.299.Rouge + 0.587.Vert + 0.114.Bleu}$$

Ces formules rendent compte de la manière dont l'oeil humain perçoit les trois composantes RVB, de la lumière. Pour chacune d'elles, la somme des 3 coefficients vaut 1 [27].

2.4.3.3 Saturer ou désaturer une couleur

Saturer les couleurs d'une image signifie lui donner plus de couleur, plus d'éclat. Au contraire, désaturer signifie enlever de la couleur ou de l'éclat. Une couleur totalement désaturée est noire et blanc.

2.5 Binarisation

La binarisation signifie que si un pixel de l'image a une intensité supérieure à une certaine valeur de seuil. Il lui sera attribué la couleur blanche sinon il sera noir. Cette tâche est appliquée sur chaque pixel de l'image. Nous obtenons donc une image comportant seulement deux niveaux (valeur 0 ou 1).
Dans les documents, la binarisation est utilisé afin de séparer le texte des régions d'arrière-plan en utilisant une technique de sélection des seuils triés, elle permet de classer tous les pixels sous forme de texte ou de non-texte. C'est une étape importante et critique car elle permet de minimiser l'espace de stockage de l'image, augmenter la lisibilité des zones de texte et permet l'efficacité et la rapidité de traitement ultérieur pour la segmentation et la reconnaissance de page [5].

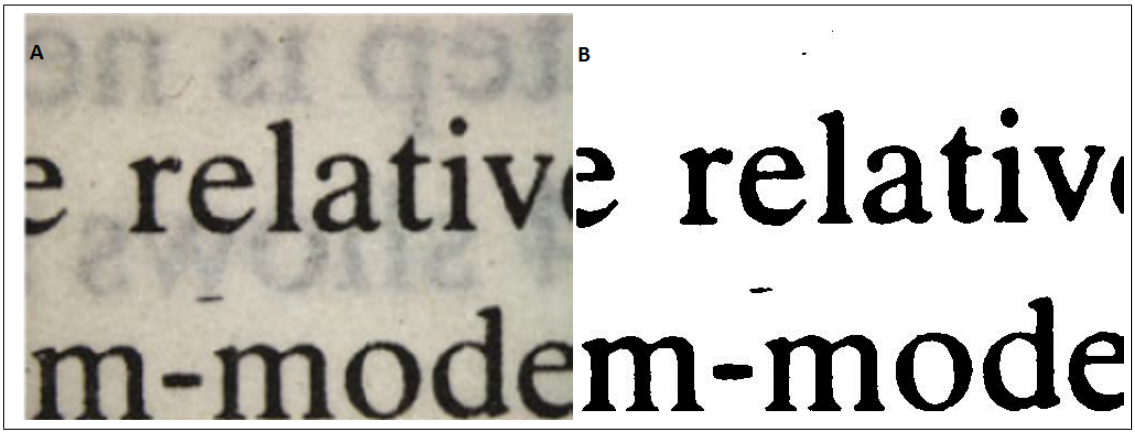


FIGURE 2.9 – Texte binarisé. A :texte originale, B :texte binarisée avec un seuil=50.

Il existe quatre techniques de seuillage : (i) seuillage global, (ii) seuillage local, (iii) seuillage hybride, (iv) seuillage par combinaison. Voici un tableau qui regroupe toutes les techniques utilisées en seuillage :

Référence	Catégorie	Brève description	Remarque
Otsu [23]	Seuillage global	le seuil optimal est calculé en séparant les deux classes de pixels de sorte que la variance entre les classes est maximisée	succès que pour les documents ayant un histogramme de distribution bimodale
Solihin et al [30]	Seuillage global	un seuillage qui exigent que chaque pixel doivent être affecté à une de trois classes (au premier plan, fond, zone floue)	conçu pour binariser l'écriture en niveaux de gris
Niblack [21]	Seuillage local	un algorithme de binarisation locale qui calcule le seuil d'un pixel "sage" en changeant une forme de fenêtre rectangulaire sur l'image	le bruit qui est présenté dans le fond peut rester dominante dans l'image binaire finale
Sauvola et Pietikainen [29]	Seuillage local	une modification qui ajoute une hypothèse sur les valeurs de gris de texte et au fond	le bruit est éliminé, mais les régions texte peuvent manquer
Tseng and Lee [33]	Seuillage hybride	basé sur l'analyse et la disposition du document. Les pixels détectés hors les blocs sont binarisées en utilisant une technique Otsu	Testé sur un large ensemble des images comportant articles journaux, magazines et des cartes de visite
Gatos et al [10]	Combinaison des techniques	Combine les résultats des méthodes de binarisation globale et adaptatif en utilisant une majorité des stratégies de vote et des informations intègres	Pour les documents historique et dégradé

TABLE 2.1 – Les techniques de Seuillage

2.5.1 Seuillage global

Cette tâche consiste à calculer le seuil optimal T afin de diviser l'image en deux classes. La sélection d'un seuil T affecte directement la qualité des images. Une valeur de T plus petit peut conduire à des cassés ou des caractères faibles tandis

qu'une plus grande valeur de T peut générer des caractères bruyants ou fusionnés dans l'image binaire résultante.
Le seuillage global a une bonne performance quand il y a une bonne séparation entre le premier plan et les zones d'arrière-plan. S'il y a un chevauchement entre ces deux classes, alors les techniques globales de seuillage peuvent échouer [5].

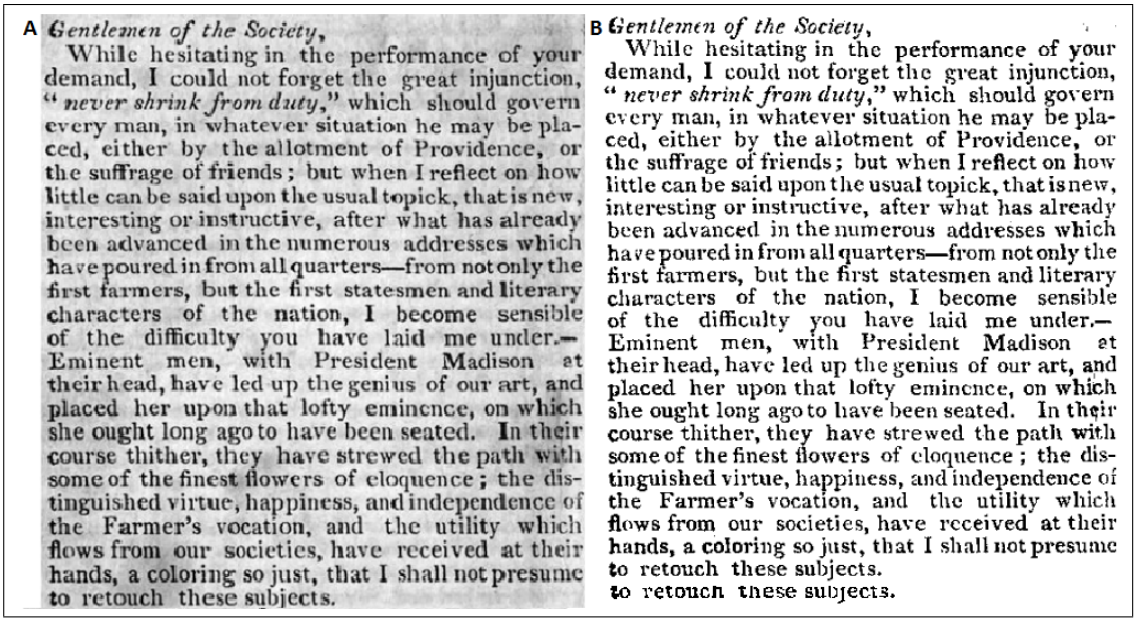


FIGURE 2.10 – Binarisation d'un document., A :document originale, B :document binarisée avec un seuillage global Seuil =100

2.5.1.1 Seuillage par Otsu

L'une des techniques de seuillage global efficace et fréquemment utilisé a été proposé par Otsu, elle est basée sur l'analyse de l'histogramme. L'opération de seuil considère que le partitionnement des pixels d'une image en deux classes $\mathcal{C}_0$ et $\mathcal{C}_1$, où $\mathcal{C}_0$ est le premier plan (texte) et $\mathcal{C}_1$ c'est l'arrière-plan.
Lorsque on utilise le seuil global t pour les mesures, les critères discriminants maximisent η sont utilisés où η est une mesure de séparabilité définie comme :

$$\eta = \frac{\delta_B^2}{\delta_T^2} \tag{2.1}$$

où δ_B^2 and δ_T^2 sont la variance de la classe et la variance totale, respectivement, et

$$\delta_T^2 = \sum_{i=0}^{255}(i-\mu_T)^2 p_i, \mu_T = \sum_{i=0}^{255} i p_i \tag{2.2}$$

$$\delta_B^2 = \omega_0\omega_1(\mu_1-\mu_0)^2 \tag{2.3}$$

p_i est la probabilité d'occurrence de niveau de gris i et définie comme

$$p_i = \frac{H_i}{N} \tag{2.4}$$

Où H_i c'est l'histogramme de niveaux de gris, et N est le nombre total de pixels. ω_0,ω_1,μ_0 et μ_1 sont définis comme suit :

$$\omega_0 = \sum_{i=0}^{t} p_i, \omega_1 = 1-\omega_0 \tag{2.5}$$

$$\mu_0 = \frac{\sum_{i=0}^{t} tp_i}{\omega_0}, \mu_1 = \frac{\sum_{i=t+1}^{255} tp(i)}{\omega_1} \tag{2.6}$$

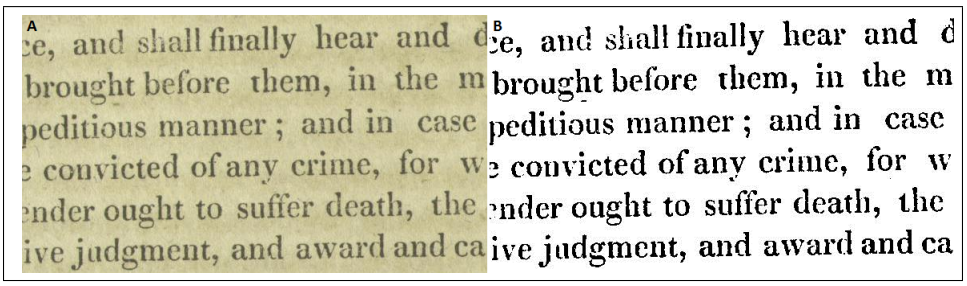


FIGURE 2.11 – Document binarisé par la méthode d'Otsu. A :document original, B :document binarisé.

2.5.2 Seuillage local

Selon cette technique, les informations locales guident la valeur du seuil pour chaque pixel de l'image. Ces techniques ont été largement utilisés dans l'analyse des documents images, car ils ont une meilleure performance dans l'extraction des traces de caractères à partir d'une image qui contient des niveaux de gris spatialement inégaux en raison des dégradations (voir Figure 2.12) [5].

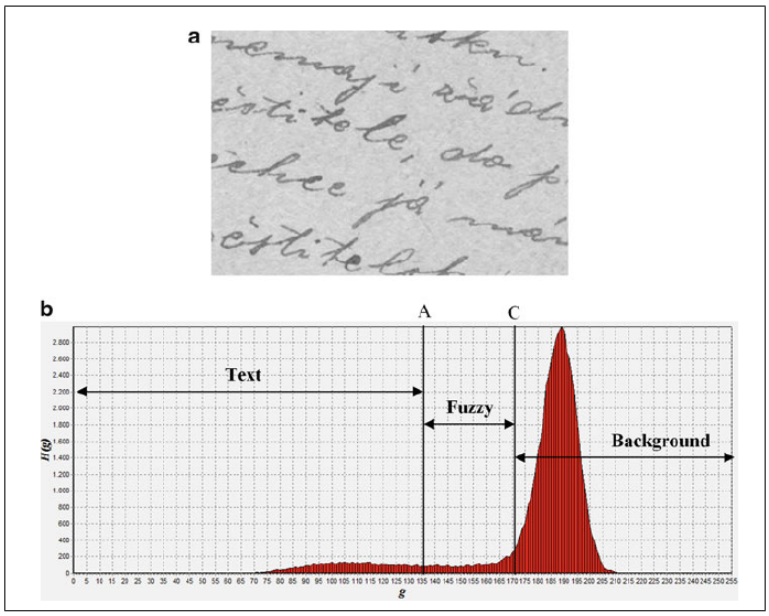


FIGURE 2.12 – Les niveaux de gris d'un document.

2.5.2.1 Seuillage par Niblack

Niblack introduit un algorithme de binarisation locale qui calcule le seuil en déplaçant une fenêtre rectangulaire sur l'image. Le seuil T pour le pixel central de la fenêtre est calculé en utilisant la moyenne m et la variance s des valeurs de gris dans la fenêtre [5].

$$\boxed{T = m + ks} \tag{2.7}$$

Où k est un ensemble constant égale à -0.2, la valeur de k est utilisée pour déterminer combien est de la limite totale de l'objet d'impression est pris comme une

partie de l'objet donné. Cette méthode peut distinguer l'objet de l'arrière-plan de manière efficace dans les zones proches des objets. Les résultats ne sont pas très sensibles à la taille de la fenêtre aussi longtemps que la fenêtre recouvre au moins une et de deux caractères. Cependant, le bruit qui est présent dans l'arrière-plan reste dominant dans l'image binaire finale. Par conséquent, si les objets sont rares dans une image, beaucoup de bruit de fond sera laissé [5].

FIGURE 2.13 – Document binarisée par la méthode de Niblack. A :document originale, B :document binarisée

2.5.2.2 Seuillage par Sauvola

Sauvola propose une méthode qui résout ce problème de seuillage de Niblack par l'ajout d'une hypothèse sur les valeurs de gris de texte et d'arrière-plan (les pixels qui ont des valeurs de gris près de 0 et des pixels de fond qui ont des valeurs de gris près de 255) [5]. Exemple sur le résultat de seuillage par Sauvola est présentés sur la Figure 2.21

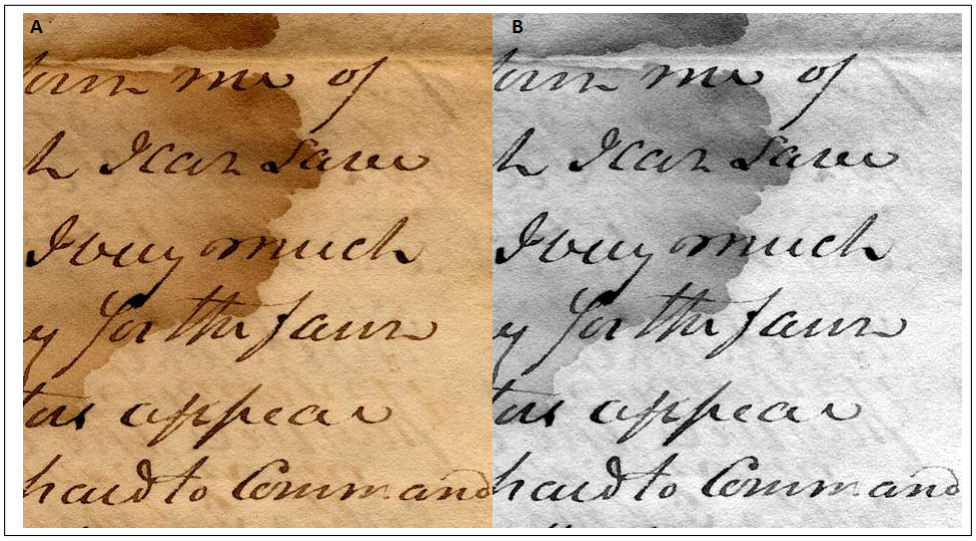


FIGURE 2.14 – Document binarisée par la méthode de Sauvola. A :document original, B :document binarisé.

2.5.3 Seuillage hybride

Proposé pour la binarisation des documents d'image qui utilisent à la fois l'information globale et locale afin de décider si un pixel appartient au texte ou à une catégorie d'arrière-plan.
Les techniques de seuillage hybrides permettent de distinguer la différence entre l'arrière-plan et le premier plan, puis à binariser l'image améliorée par un algorithme global simple [5].

2.5.4 Seuillage par combinaison

Récemment, plusieurs techniques de binarisation combinent les résultats d'un ensemble de techniques de binarisation afin d'utiliser la complémentarité dans la réussite de chaque technique. Pour plus de détails voire le tableau 2.1

2.5.4.1 Méthode Su et al

Su et al. [32] dit que si en combinant les différentes techniques de binarisation, le rendement obtenu peut être amélioré. Avec une analyse approfondie, les pixels

qui sont étiquetés par des méthodes différentes sont habituellement classés correctement, les pixels qui sont classés en tant que texte par certaines méthodes et les pixels qui sont classés comme fond par d'autres méthodes ont plus de grande possibilité d'être mal classés.
Basé sur cette observation, nous divisons tous les pixels de l'image en trois ensembles :

- Jeu de premier plan : où les pixels sont classés en premier plan par tous les méthodes de binarisation ;
- Jeu de fond : où les pixels sont classés en arrière-plan par tous les méthodes de binarisation ;
- Jeu incertain : où le reste des pixels appartiennent à une zone incertaine.

2.6 Morphologie

2.6.0.2 Définition

Le mot morphologie désigne une branche de la biologie qui traite de la forme et la structure des animaux et plantes. On utilise le même mot ici, en traitement d'image, c'est comme un outil pour extraire les composantes de l'image qui sont utiles dans la représentation et la description de la région de forme, telles que les frontières, squelettes... etc [19].

2.6.0.3 L'utilisation de la Morphologie

- La segmentation d'images,
- La régularisation des formes,
- L'affinage des contours,
- La détection de défauts,

2.6.0.4 Les Opérateurs morphologiques

Erosion : si un des éléments du voisinage du masque (élément structurant) correspond à un pixel du fond (valeur 0) alors le pixel central devient fond (pixel de l'objet) [19].

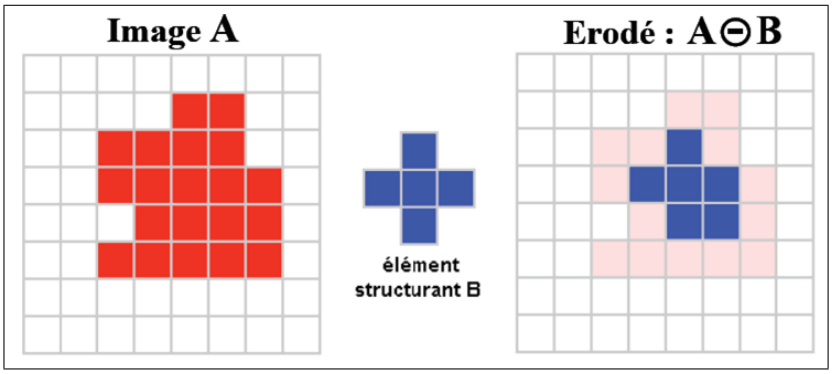


FIGURE 2.15 – Exemple sur le déroulement de l'érosion.

Dilatation : si un des éléments du voisinage du masque est à 1 alors les éléments du masque appartiennent à l'image dilatée [19].

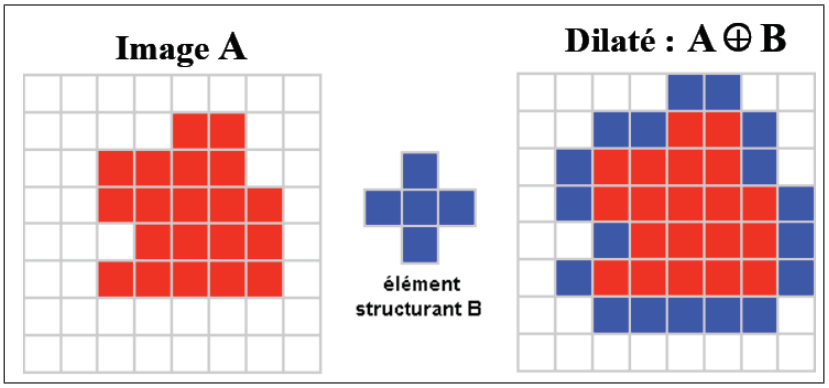


FIGURE 2.16 – Exemple sur le déroulement de la dilatation.

Voici l'effet de la dilatation sur un document :

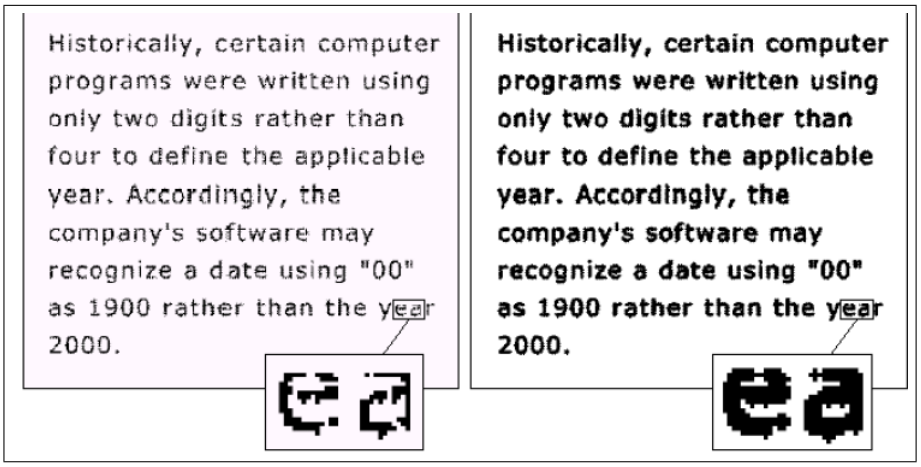


FIGURE 2.17 – Document dilaté [12].

Ouverture : Erosion suivie d'une dilatation par transposé de B afin d'isoler les surfaces présentes dans l'image et lisser les contours.

Fermeture : Dilatation suivie d'une érosion par transposé de B [12].

2.6.0.5 But de l'ouverture et la fermeture

- Le but de l'ouverture est d'isoler les surfaces présentes dans l'image et lisser les contours.
- Le but de la fermeture est de recoller des morceaux de surfaces proches de manière à fermer des contours disjoints et lisser les contours [19].

2.6.0.6 Squelettisation

La squelettisation donne une représentation compacte des objets. En dimension 2, les squelettes représentent des lignes inter connectées au centre d'un objet. En dimension 3, les squelettes peuvent représenter des lignes centrales ou bien des surfaces centrales [9].

La squelettisation est très utilisée en analyse d'image et reconnaissance de forme, car elles permettent de d'écrire synthétiquement non seulement la forme, mais aussi certaines propriétés mathématiques des objets, comme par exemple la longueur ou la surface [9].

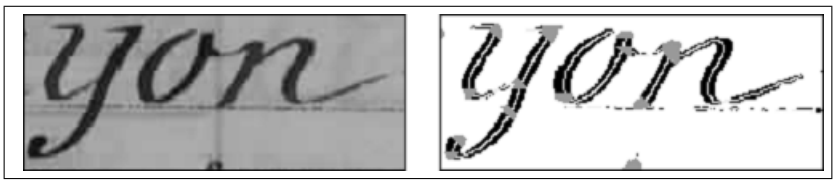


FIGURE 2.18 – La squelettisation d'un texte [9].

2.7 Conclusion

Pour éliminer les bruits parasites et les artéfacts nuisibles, il est recommandé de commencer le traitement d'un document scanné par une étape de prétraitement. On peut dire que le prétraitement est une étape très importante dans le traitement et l'analyse de document car si on entame l'analyse de document sans prétraitement, les résultats vont être dégradés. Dans le chapitre qui suit, nous allons présenter quelques approches pour l'analyse de la structure des documents numérisés.

Chapitre 3

Analyse de la structure des documents numériques

3.1 Introduction

Un document est composé d'une variété d'entités ou des régions tels que des blocs de texte, des lignes, des mots, des chiffres, des tableaux et des arrière-plans. Aussi des variétés des étiquettes fonctionnelles ou logiques comme les titres, les légendes, les noms des auteurs et les adresses.
Le processus d'analyse de la structure du document, tente de décomposer un document donné en des régions fondamentales afin de comprendre chaque rôle de ces régions.
Ce chapitre vise à identifier ce processus et distinguer quelles sont les méthodes adéquates à utilisées ?

3.2 Définition d'un document

Le mot document vient du mot latin " Documentum ", comme le verbe " doceo " qui signifie "enseigner" d'après L'ISO un document est défini comme un :
"ensemble formé par un support et une information, généralement enregistrée de façon permanente, et tel qu'il puisse être lu par l'homme et la machine "

Un document papier peut être un questionnaire, avec des formes graphiques, des images et des champs de texte, éventuellement numériques. Comme il peut ne contenir que du texte, page d'un roman par exemple [13].

Un document dispose de deux types de structures :

3.2.1 Structure physique

La structure physique du document est une modélisation qui représente l'organisation des éléments graphiques et des symboles utilisés pour le codage de l'information. Il s'agit de la description du support physique et de son apparence visuelle et matérielle. Dans la forme papier, les entités physiques sont de plusieurs niveaux : le document lui-même dans sa globalité, l'ensemble de pages, la page, le cadre, le bloc, la ligne, jusqu' à l'entité graphique élémentaire. La structure physique est alors souvent représentée comme une hiérarchie de ces entités physiques [13].

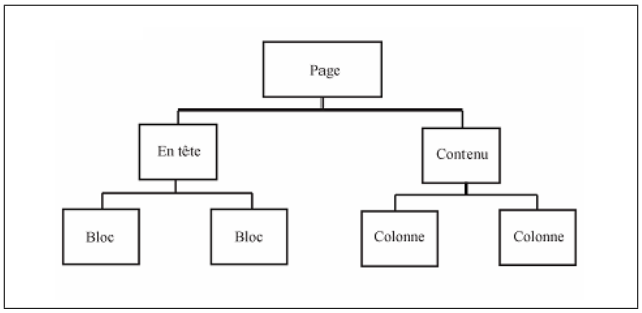


FIGURE 3.1 – Structure physique d'un document.

3.2.2 Structure logique

La structure logique est une représentation abstraite du document ne tenant compte ni du support ni de sa présentation. Elle décrit l'organisation du discours de l'auteur, c'est-à-dire la façon dont il a articulé, structuré sa pensée pour communiquer l'information. Elle décrit un agencement d'entités, chacune associée à une fonction telle que section, chapitre, paragraphe, titre, etc... [13].

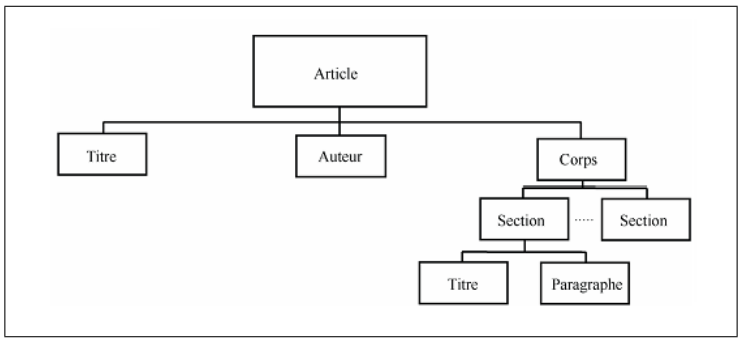


FIGURE 3.2 – Structure logique d'un document.

3.3 Segmentation

3.3.1 Définition

C'est une technique fondamentale dans l'analyse des documents, qui reçoit pour entree un document précis et renvoie des partitions contenant des pièces ou des informations significatives.

3.3.2 Les étapes de la segmentation

Prétraitement : éliminer les informations qui sont superflues ou pertinentes afin de faire une tâche spécifique (par exemple : utiliser le filtre médian).

Extraction des caractéristiques : rassembler des données nécessaires pour un partitionnement réels.

Decision : partitionner effectivement l'image du document [7].

Le résultat d'un tel processus donne un ensemble de blocs constant et importants (assez significatif). Pour une manipulation facile, les blocs et les cadres sont généralement représentés par leur boîte englobante, puis la page segmentée doit être traitée de manière plus efficace que celui du document original [7].

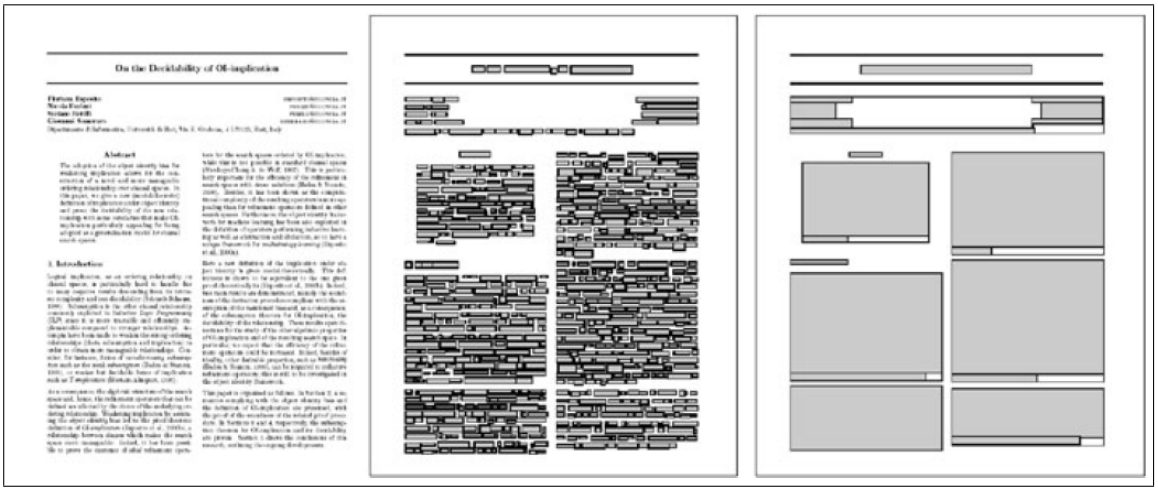


FIGURE 3.3 – Segmentation d'un document [7].

3.3.3 Les classes de mise en page

Il existe quatre types de classes de mise en page, qui sont :

- plan Rectangulaire ;
- plan Manhattan ;
- plan Non-Manhattan ;
- plan Chevauchement .

Remarque : le type Manhattan est celui dont les régions de la page sont toutes des rectangles de même orientation. En réalité ce nom fait référence à la disposition des bâtiments de quartier de Manhattan à New York.

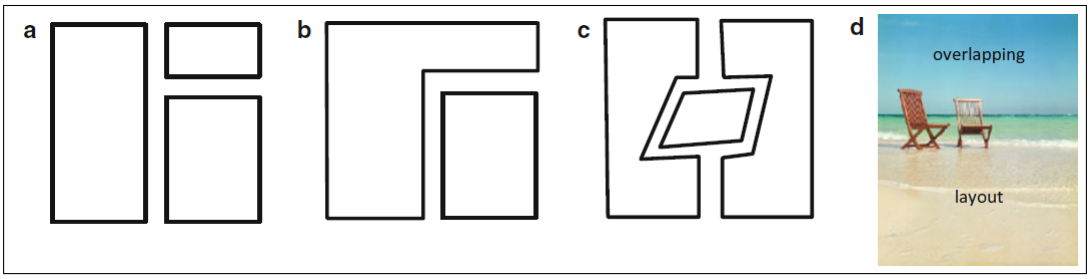


FIGURE 3.4 – Les classes de mise en page. (a) Rectangulaire, (b) Manhattan, (c) non-Manhattan, (d) Chevauchement plan [5].

3.4 Les méthodes de la segmentation et de l'analyse du document

Les méthodes de la reconnaissance de structures physiques, peuvent être classées en deux grandes classes : les méthodes descendantes et les méthodes ascendantes. Actuellement, une troisième classe figure parmi les méthodes descendantes et ascendantes c'est les méthodes mixtes. Celles-ci combinent les deux approches citées avant [31].

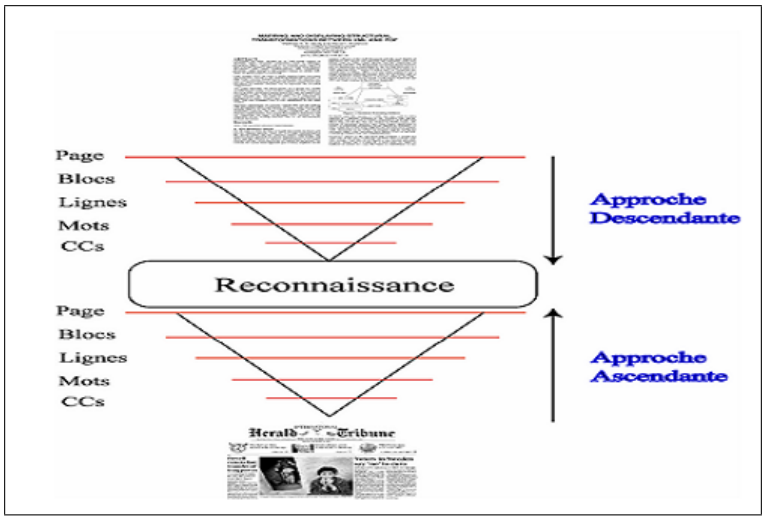


FIGURE 3.5 – Approches descendante et ascendante.

3.4.1 Les méthodes descendantes

Les méthodes descendantes commencent par le niveau le plus élevé à savoir la page et descendent jusqu'à arriver au niveau des composantes connexes ou au niveau pixel. Les algorithmes utilisant la stratégie descendante sont plus appropriés aux structures de type Manhattan.

3.4.1.1 Méthode de découpage X-Y (X-Y Trees)

L'algorithme de découpage X-Y utilisant les profils de projection a été introduit par Nagy [20]. L'hypothèse de base repose sur le fait que les éléments structurés de la page sont généralement présentés dans des blocs rectangulaires. Mais aussi

sur le fait que les blocs peuvent être divisés en groupes de telle sorte que les blocs qui sont adjacents l'un à l'autre, dans un groupe, ont une dimension en commun. Le document est successivement divisé en de petits blocs rectangulaires en faisant une alternance de découpages horizontal et vertical le long des espaces blancs. Ces espaces blancs sont trouvés en utilisant un seuil de profil de projection. Le résultat d'une telle segmentation peut être représenté dans un arbre X-Y, dans lequel la racine correspond à la page toute entière et les feuilles représentent les blocs de la page et chaque niveau de l'arbre représente alternativement les résultats de la segmentation horizontale ou verticale [20].

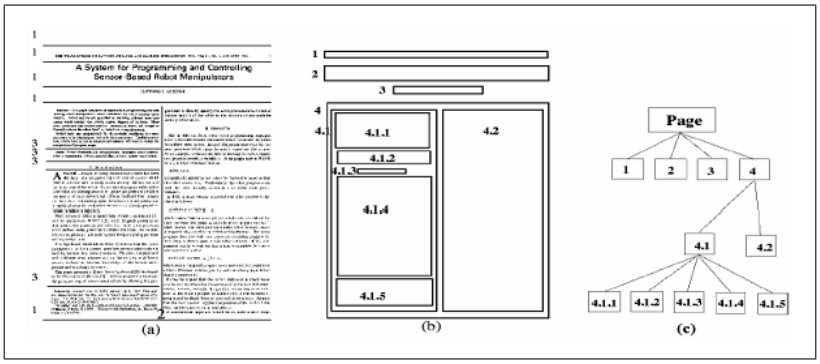


FIGURE 3.6 – (a) document original, (b) découpage X-Y, (c) arbre de découpage.

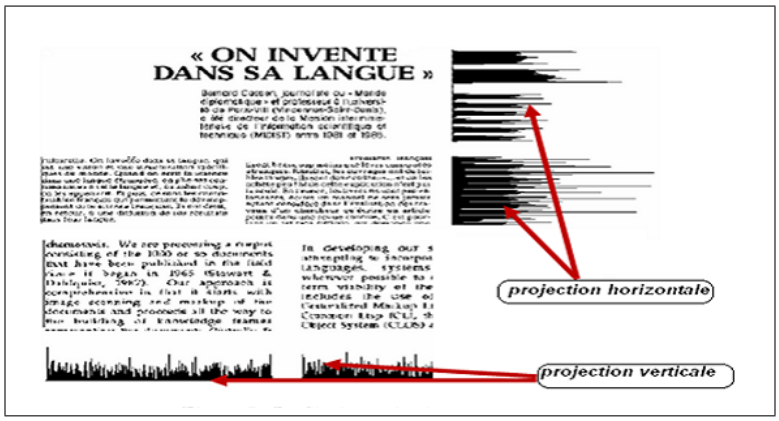


FIGURE 3.7 – Profiles de projection.

3.4.1.2 Méthode RLSA (Run Length Smoothing Algorithm)

La méthode est développée pour les systèmes d'analyse des documents, Le principe de cette méthode est de noircir les espaces blancs sur chaque ligne, de

longueur inférieure à un seuil S.
Par exemple, on applique le lissage sur la ligne suivante :
0 0 1 1 0 0 0 1 1 1 1 1 1 0 0 1 1 0 0 0 0 0 0 1 1 1 1 0 0 0 0 1 1 0 Un lissage horizontal avec S=4 donne le résultat suivant :
1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 0 0 0 0 0 1 1 1 1 1 1 1 1 1 1 1
En particulier, le RLSA fonctionne en quatre étapes, dont chacune applique un opérateur :

1. Lissage horizontal mené par des rangées sur l'image avec un seuil t_h ;
2. Lissage vertical, réalisé par des colonnes sur l'image de seuil t_v ;
3. ET logique des images obtenues dans les étapes 1 et 2 (le résultat a un pixel noir dans des positions où les deux images d'entrée ont un, ou un pixel blanc autrement) ;
4. Nouveau lissage horizontal avec un seuil t_a sur l'image obtenue à l'étape 3, pour remplir les pistes blanches à l'intérieur des blocs découverts [7].

Le résultat des trois premières étapes de RLSA sur un document est représenté sur la Figure ci-dessous.

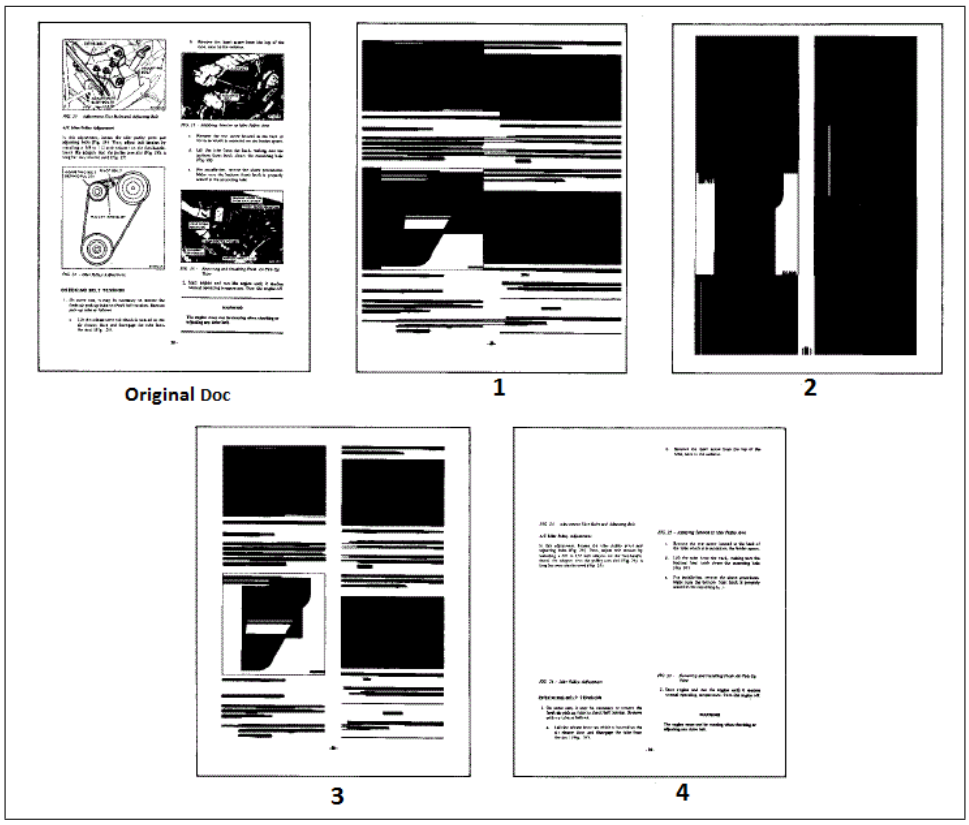


FIGURE 3.8 – Application de la méthode RLSA. 1 :Lissage horizontale, 2 :Lissage vertical, 3 :le ET logique des étapes 1, 2, 4 :Nouveau lissage horizontale.

3.4.1.3 Méthode RLSO (Run-Length Smoothing with OR)

RLSO est une variante de la RLSA qui effectue :

1. lissage horizontale de l'image, réalisée par rangées avec seuil t_h ;
2. Défroissage vertical de l'image, réalisée par des colonnes avec un seuil t_c ;
3. OU logique des images obtenues dans les étapes 1 et 2 (le résultat a un noir pixel dans des positions où au moins l'une des images d'entrée a un, ou un pixel blanc autrement).

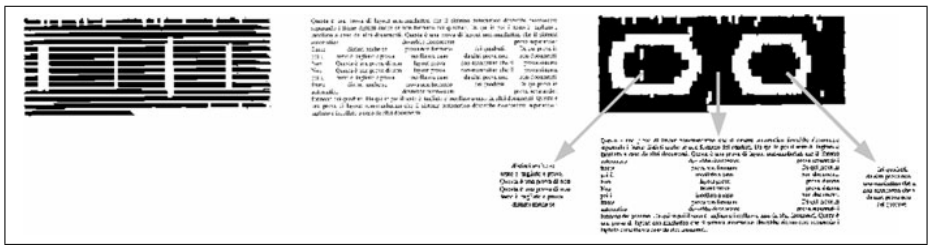


FIGURE 3.9 – Application des trois étapes RLSO a un document [7].

3.4.1.4 Comparaison des Méthodes RLSA et RLSO

Comparée à RLSA, RLSO a une étape de moins (pas de lissage horizontal final est effectué, depuis l'opération OU, contrairement à l'ET, conserve tout des deux images lissées), et exige des seuils plus courts pour remplir moins les pistes [7].

3.4.2 Les méthodes ascendantes

Les méthodes ascendantes commencent par le niveau le plus bas et remontent d'un niveau à un autre jusqu'à compléter la page.

Le principe des méthodes ascendantes est le suivant : elles commencent par fusionner du plus bas niveau, en formant les mots à partir des composantes connexes, et puis remontent à un niveau supérieur en fusionnant les mots en lignes, les lignes en blocs, etc..., jusqu'à ce que la page soit complètement reconstituée.

3.4.2.1 Méthodes utilisant les composantes connexes :

Fisher [8]. A combiné l'algorithme de lissage avec l'extraction des composantes connexes. Les composantes connexes et leurs rectangles englobant constituent les blocs de la structure physique du document. Un ensemble de caractéristiques des composantes connexes est utilisé.
Cette approche permet d'identifier les zones textes et non textes mais reste cependant sensible à la rotation de l'image du document [8].

Saitoh [28]. Aussi procède par échantillon de 8x4 pixels sur toute l'image, puis extrait les composantes connexes. Ces dernières sont classées en texte, bruit, table,

diagramme, image, ou filet, en utilisant les attributs des blocs tels que la hauteur du bloc, le ratio hauteur/largeur et la présence ou non de filets. Les blocs sont divisés selon les critères suivants : la distance verticale entre les lignes et la hauteur des lignes dans les blocs [28].

3.4.2.2 Méthodes utilisant le filtrage à base de fenêtres :

Les méthodes ascendantes, utilisant le filtrage à base de fenêtres, reposent sur un balayage d'une fenêtre d'une certaine taille sur toute l'image du document.
Lebourgeois [17]. Utilise un filtre de 8x3 pixels. L'image échantillonnée est dilatée horizontalement pour rassembler les caractères adjacents. Il est à noter que chaque composante connexe est caractérisée par son rectangle englobant et par la moyenne des longueurs de plages de valeurs de pixels noirs.
Ensuite, si la composante connexe est à l'intérieur de l'intervalle, celle-ci sera classée en une zone de texte, sinon elle sera classée en zone non texte. Les composantes connexes classées en zone texte sont fusionnées verticalement en blocs selon des règles prenant en considération l'alignement [17].

3.4.2.3 Méthodes utilisant la technique docstrum

O'Gorman [22]. Introduit la technique "docstrum" qui est une technique d'analyse de structures physiques de page, basée sur la combinaison de l'analyse ascendante qui fait intervenir le calcul des k plus proches voisins pour chaque composante connexe de la page.
Chaque paire de voisins les plus proches possède un angle et une distance associée. En regroupant les composants à travers les caractéristiques citées précédemment, les régions géométriques de structures physiques de la page peuvent être déterminées. La méthode proposée est indépendante du changement de l'orientation de la page mais aussi de l'espacement intertexte. Cependant, la valeur du k est dépendante de la structure de la page [22].

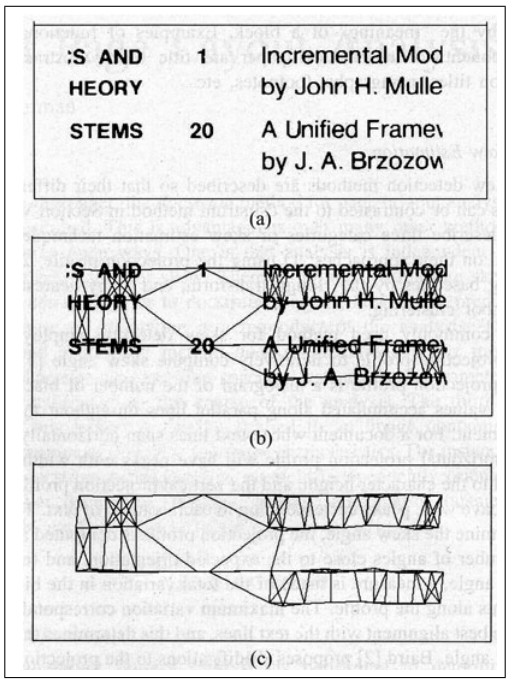


FIGURE 3.10 – Application de la méthode DOCSTRUM. (a) le fragment de document original ; (b) le fragment avec des segments superposés, (c) résultat du méthode [7].

3.4.2.4 Méthode utilisant les diagrammes de Voronoï

Kise [15]. Présente une méthode de segmentation de pages basée sur la surface approximée des diagrammes de Voronoï. La méthode repose sur les étapes suivantes : au début, un point du diagramme de Voronoï est construit à partir de l'ensemble des pixels noirs sur les contours des composantes connexes. Ensuite, une surface est obtenue en éliminant du point du diagramme Voronoï toutes les arêtes générées à partir d'une paire de points sur la même composante connexe. Une caractéristique de cette méthode est qu'elle s'applique sur des images de documents possédant une structure de type Manhattan et ayant subi une rotation. Il est à noter que cette méthode est efficace pour l'extraction des zones de texte [15].

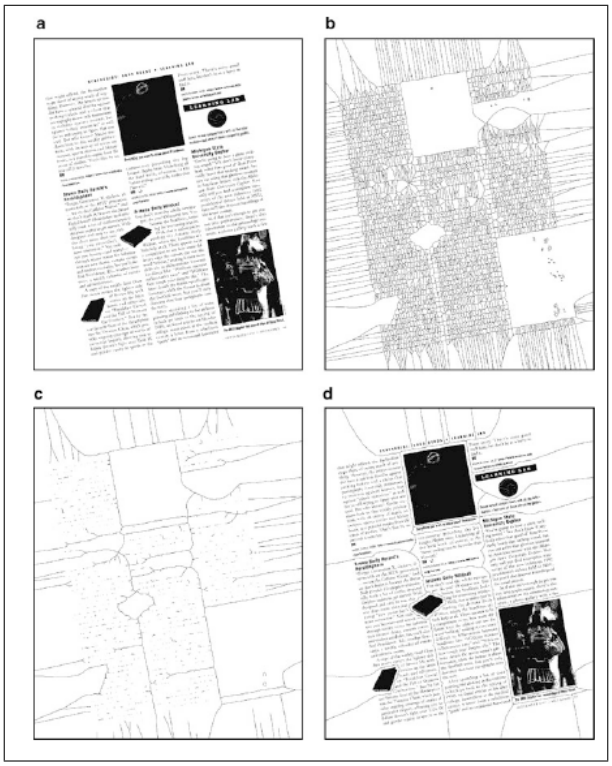


FIGURE 3.11 – Diagramme Area Voronoi et son utilisation pour la segmentation du page [5].

3.4.3 Les méthodes mixtes

Les méthodes descendantes et ascendantes ont leurs limites. En effet, ces méthodes purement descendantes ou ascendantes donnent de bons résultats en présence de classes spécifiques de documents et peinent en présence d'autres types de classes.

Devant ces insuffisances mutuelles de deux types de méthodes, il y a eu naissance d'un nouveau type de méthode ; les méthodes mixtes. Les méthodes mixtes résultent de la combinaison des méthodes descendantes et ascendantes ou de l'utilisation conjointe d'une de ces dernières avec une autre méthode [20].

comme par exemple l'analyse syntaxique. L'analyse syntaxique tient ses origines de la compilation des programmes informatiques écrits dans un langage de

programmation à côté de l'analyse lexicale. Cette technique d'analyse a été approchée pour la reconnaissance de structures physiques de documents.
En effet, Krishnamoorthy [16]. Propose une méthode permettant la combinaison de la segmentation d'une image avec l'étiquetage. Cette méthode est différente des autres méthodes d'analyse syntaxique des documents en deux points.

- Le premier point c'est que les grammaires utilisées forment une hiérarchie.
- Le deuxième point c'est la combinaison de formules syntaxiques avec la méthode de recherche du "branch and bound" [16].

3.5 Conclusion

Beaucoup de méthodes ont été utilisées dans la segmentation et l'analyse des documents. On s'est intéressé a choisir des méthodes plus générales qui sont basées sur des contraintes spécifiques a afin de déterminer un meilleur résultat et une performance des méthodes. Des pages difficiles à segmenter par une méthode sont également difficiles par d'autres méthodes. Ces méthodes présentent un avantage en ce qui concerne le temps de calcul et de mémoire. Le chapitre suivant résume quelques outils et plateformes développés pour traiter les problèmes d'analyse des documents scannés.

Chapitre 4

Outils pour l'analyse et la reconnaissance des documents

4.1 Introduction

Depuis de nombreuses années, la numérisation des documents imprimés a été l'objet d'un grand développement et reste encore loin d'être fini. Poussé par les besoins toujours croissant de l'accessibilité à l'information. Des efforts considérables sont déployés pour transformer le patrimoine de livres, de journaux et d'autres documents anciens en des médias numériques modernes. Les méthodes et applications proposées doivent être très efficaces et fiables pour être compétitifs vers le contexte de la numérisation de documents. Dans ce chapitre, nous présentons quelques outils que nous avons utilisés afin de nous faciliter le travail.

4.2 Tesseract

Tesseract est un logiciel de reconnaissance optique de caractères sous licence Apache. Conçu par les ingénieurs de Hewlett Packard de 1985 à 1995, son développement est abandonné pendant les dix années suivantes. En 2005, les sources du logiciel sont libérées sous licence Apache et le logiciel est actuellement développé par Google. Initialement limité aux caractères ASCII. Actuellement, il supporte

parfaitement les caractères UTF-8 et reconnaît maintenant plus de 40 langues [24].

4.3 Les Outils du labo Prima

Prima est un laboratoire construit sur plusieurs années. Il est spécialisé dans une variété de domaines d'expertise : bibliothèques numériques, numérisation et OCR, développement de logiciels.

4.3.1 TesseractToPAGE 1.3

TesseractToPAGE est un outil de ligne de commande pour analyser un document en utilisant le moteur OCR. Tesseract est un logiciel open source qui permet d'exporter les résultats de la page au format XML [24].

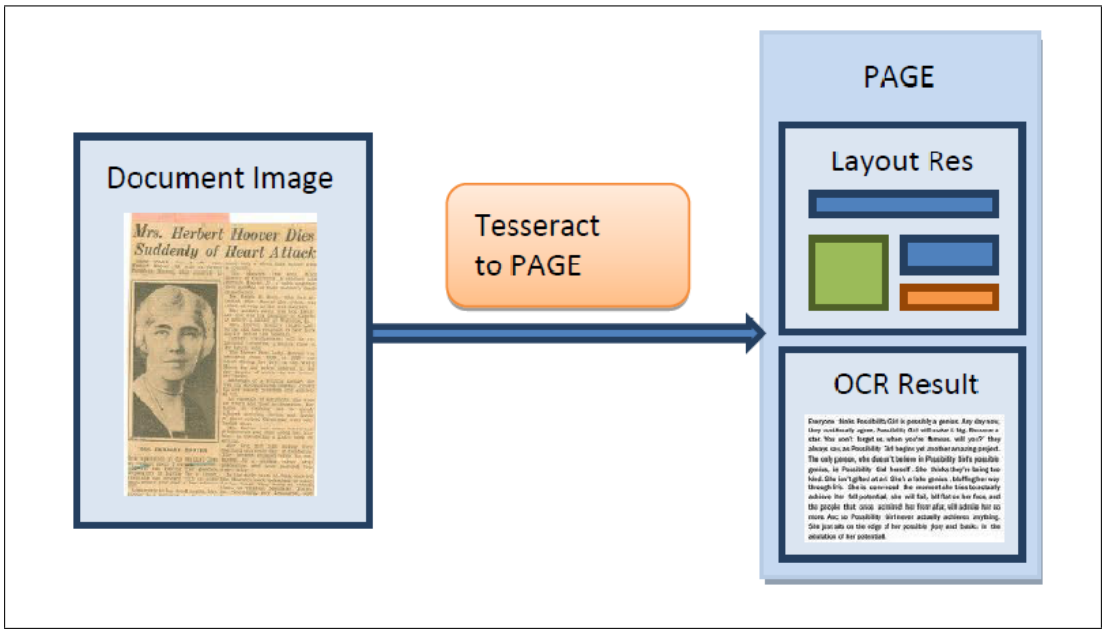


FIGURE 4.1 – Image représentatif de TesseractToPAGE [24].

4.3.1.1 L'usage de TesseractToPAGE 1.3

La syntaxe de commande de TesseractToPAGE peut être définie comme suit :

... $-arg1value1[-arg2value2...][-option1-option2...]$ Avec :

arg : nom de l'argument.

value : valeur de l'argument.

option : l'option de ligne de commande [24].

4.3.1.2 Les Arguments

-inp-img <file path> : fichier d'image d'entrée.
-out-xml <file path> : le fichier de sortie (PAGE XML).
-out-img <file path> : chemin du fichier de l'image de sortie binaire (optionnel).

4.3.1.3 Les Modes

-rec-mode <mode> : mode de reconnaissance (niveau de données pour remplir le fichier PAGE).

l'un de :

Ocr-Regions : Layout et text (régions).

Ocr-Lines : Layout et text (régions, lignes).

Ocr-Words : Layout et text (régions, lignes, mots).

Ocr-Glyphs : Layout et text (régions, lignes, mots, glyphes) [24].

Les Modes pris en charge :

AUTO : Segmentation de page entièrement automatique, mais pas de l'OSD.

AUTO OSD (par défaut) : La segmentation automatique avec détection de l'orientation et l'écriture (OSD).

AUTO ONLY : Segmentation de page automatique, mais pas OSD, ou OCR.

OSD ONLY∗ : Orientation et de détection de script uniquement.

SINGLE COLUMN : Supposons une seule colonne de texte (taille variable).

SINGLE BLOCK VERT TEXT∗ : Supposons un seul bloc uniforme de texte aligné verticalement.

SINGLE BLOCK : Supposons un seul bloc uniforme de texte.

SINGLE LINE∗ : Traiter l'image comme une seule ligne de texte.

SINGLE WORD∗ : Traiter l'image comme un seul mot.

CIRCLE WORD∗ : Traiter l'image comme un seul mot dans un cercle.

SINGLE CHAR∗ : Traiter l'image comme un seul caractère [24].

Remarque : Les modes marqués d'un * ne sont pris en charge pour 'ocr' rec-modes.

4.3.1.4 Les options

orig-outlines : pour exporter les contours tel que produit par Tesseract.

reading-order : pour essayer de créer l'ordre de lecture.

debug : pour des informations de sortie de débogage.

-lang <language> : langue de reconnaissance.

-tessdata <folder> : dossier racine des données.

Par défaut : le dossier parent du dossier avec exécutable.

4.3.1.5 Exemples

Mode Layout analysis only (les segments dans les régions) :
$...exe - inp - img"001.tif" - out - xml"001.xml" - rec - modelayout$

FIGURE 4.2 – L'effet du Layout analysis [24].

Mode Layout analysis and OCR on region level :

...exe – inp – img"001.tif" – out – xml"001.xml" – rec – modeocr – regions

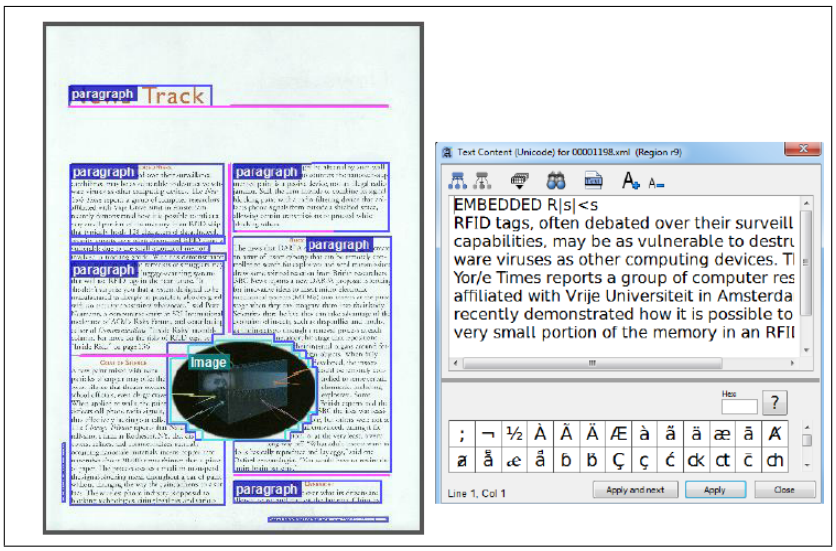


FIGURE 4.3 – L'effet du Layout analysis et OCR on on niveau region [24].

Mode Layout analysis and OCR on region, text line and word level :

...exe – inp – img"001.tif" – out – xml"001.xml" – rec – modeocr – words

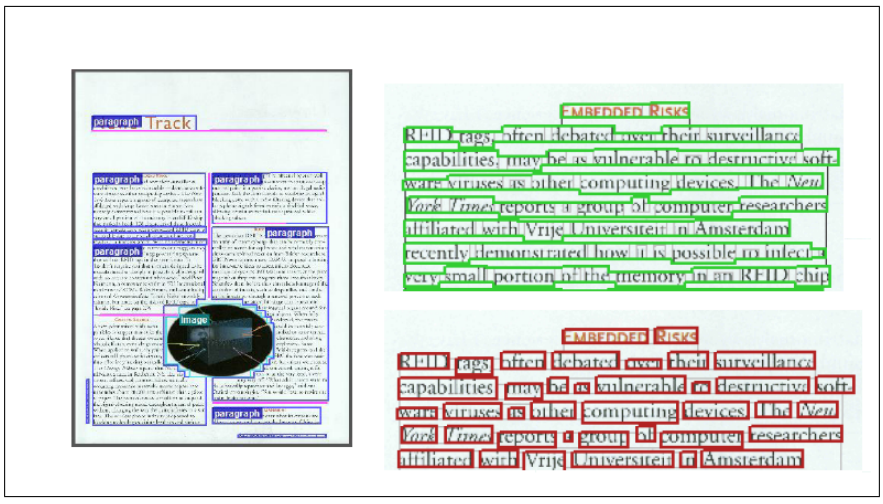


FIGURE 4.4 – L'effet de l'analyse de Layout et OCR au niveau region, text, ligne, et mot [24].

4.3.2 Aletheia

Aletheia vient du mot grec (vérité), Aletheia a été partiellement développé dans le contexte de l'amélioration de l'accès au texte, c'est un projet visant à améliorer les technologies de numérisation des différents documents, l'objectif principal de ce système est l'efficacité, la précision, la flexibilité et la facilité de l'utilisation, et tout ce qui concerne l'évaluation à grande échelle.
Aletheia a une performance très élevée dans l'analyse de documents, les résultats de la segmentation sont représentés selon un schéma XML sophistiqué qui contient les composantes de la page.
En plus, les régions de texte peuvent être structurés en lignes, mots et glyphes ainsi que des résultats OCR [3].

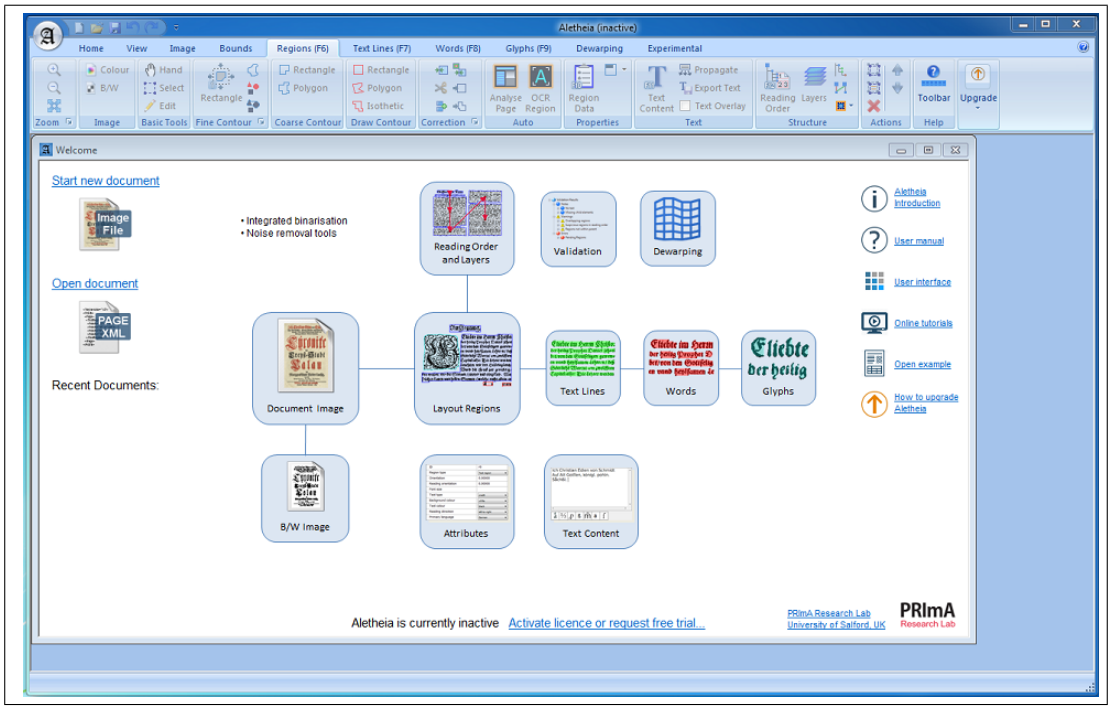


FIGURE 4.5 – Capture d'écran de l'outil Aletheia.

4.3.2.1 les Options de Aletheia

Il existe plusieurs options dans l'application Aletheia, parmi ces options on peut citer :

4.3.2.2 Les Opérations d'image

- Binarisation : seuillage par Otsu ou Sauvola.
- Suppression du bruit : bruit Poivre et sel pour les composants sélectionnés.

FIGURE 4.6 – Les opérations d'image dans Aletheia [25].

4.3.2.3 Bordure et Espace d'impression

- Border : marquer le bord d'un document numérisé sous forme d'un polygone unique.
- Imprimer espace : c'est un polygone qui marquent le corps principal du texte d'un document sans numéros de page, marginalia, etc [25].

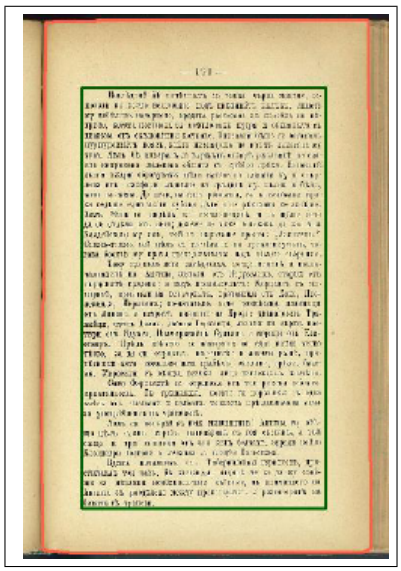


FIGURE 4.7 – Border et Espace d'impression dans Aletheia [25].

4.3.2.4 Régions et mise en page

FIGURE 4.8 – Régions et mise en page dans Aletheia [25].

- Forme polygonale.
- 11 types : texte, image, tableau, séparateur, ...
- Sous-types : texte, paragraphe, titre, numéro de page, ...

– D'autres attributs (par exemple le contenu du texte, la langue, l'orientation)

4.3.2.5 Création des régions

– Outils manuels (rectangles, polygone arbitraire et esthétique)
– Outils semi-automatiques (réducteurs) : basé sur les boites englobantes des composantes connectées [25].

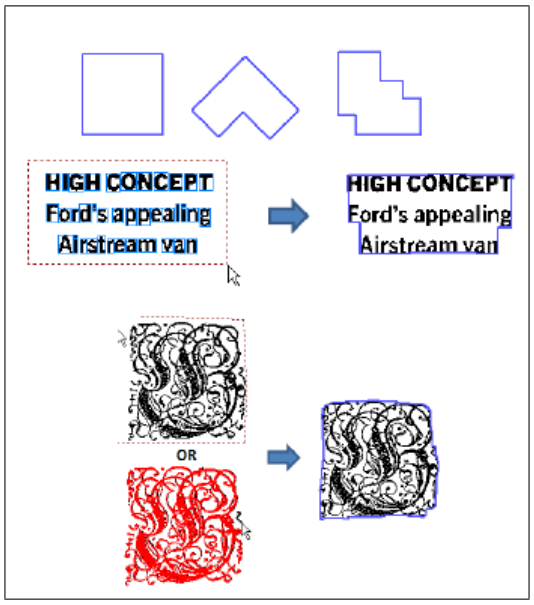


FIGURE 4.9 – Création des régions dans Aletheia [25].

4.3.2.6 Modification des régions

Correction des données de pré-production par :
– Ajouter, déplacer et supprimer des points de polygone.

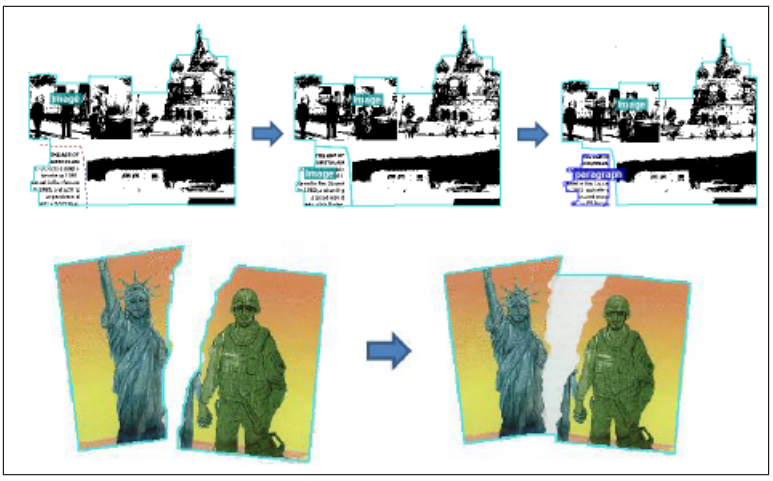


FIGURE 4.10 – Modification des régions dans Aletheia [25].

4.3.2.7 Contenu du Text

- Unicode
- Police spéciale (support des caractères qui ne sont pas (encore) partie de la norme Unicode)
- Recherche de texte

4.3.2.8 Superposition de texte

Superpositionner le textes sur l'image du document afin d'assurer la qualité.

FIGURE 4.11 – Superposition de texte dans Aletheia [25].

4.3.2.9 Lignes de texte

Marquage des lignes de texte [25] :

- division en régions avec un seul click split.
- Par rétrécissement des composants connectés.
- Par la fusion des fragments de ligne :
- Par les mots combinés.
- Par dessin manuellement.

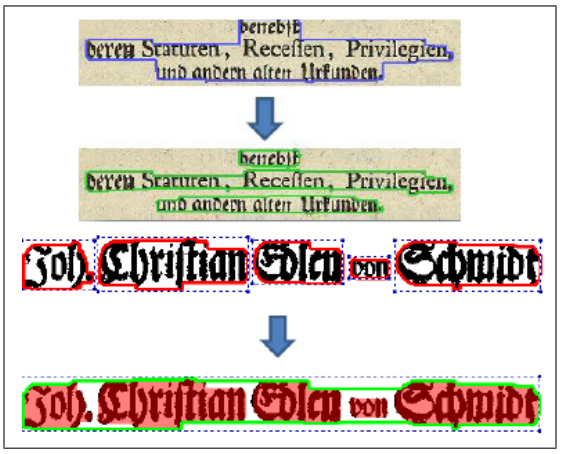


FIGURE 4.12 – Lignes de texte dans Aletheia [25].

4.3.2.10 Les mots et les Glyphes

Afin de marquer les lignes de texte, on utilise :

- Outils split.
- Outil merge.
- Shrinkers.
- Dessin manuel.

Le texte peut être propagé à tous les niveaux (par exemple, des régions à glyphes) [25].

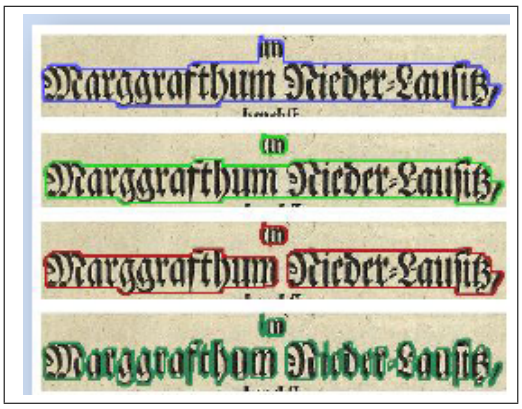


FIGURE 4.13 – Les mots et les Glyphes dans Aletheia [25].

4.3.2.11 Format XML

- Générer un schéma XML qui fait partie de la page.
- Plus de compatibilité ascendante (tous les outils peuvent gérer les anciennes versions du schéma XML).

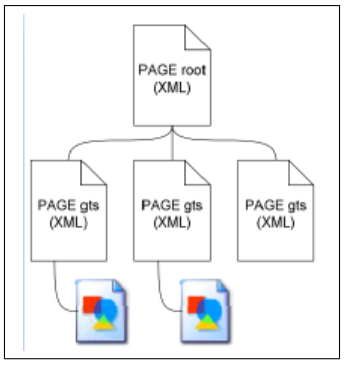


FIGURE 4.14 – Format XML dans Aletheia [25].

4.3.3 PAGE Viewer

FIGURE 4.15 – Capture écran de l'outil PAGE Viewer.

Page Viewer est une application pour la visualisation de la mise en page, le contenu du texte de la segmentation et les résultats de la page de reconnaissance et systèmes OCR. Le format de fichier naïf pris en charge est la PAGE XML. Cependant, les formats ALTO XML, FineReader XML et hOCR peuvent être ouvert aussi.
Page Viewer montre la mise en page en transparence sur l'image du document. Le contenu et les attributs des objets texte sont affichés sous forme d'infobulles.

4.4 Leptonica

Leptonica est le mot italien pour "leptonic", qui est un adjectif qui se rapporte à trois familles très similaires de particules fondamentales, appelées leptons, dont l'électron et son neutrino comprennent le plus familier. Leptonica est un logiciel open source orienté pédagogique qui est largement utile pour les applications de traitement d'image et d'analyse d'image.
Leptonica est utilisé aussi par d'autres outils tels que Tesseract et Aletheia afin

d'offrir plusieurs fonctionnalités dans leurs applications.

Leptonica avait plusieurs fonctions dans le traitement des images comme :

- La morphologie.
- La quantification des couleurs.
- L'amélioration des images.
- La rotation des images.
- La mise à l'échelle des images [11].

Dans le domaine de l'analyse des documents, Leptonica contient certaines fonctions pour aider à l'analyse des documents, par exemple :

- La détection d'inclinaison ;
- L'orientation du texte ;
- La segmentation des pages textes ;
- La classification non supervisée des composantes connectées, des caractères et des mots ;
- La reconnaissance de chiffres et des glyphes et des mots [11].

Prenons par exemple la segmentation des pages textes voici les résultats obtenus en utilisant Leptonica :

FIGURE 4.16 – Génération des texts blocks [11].

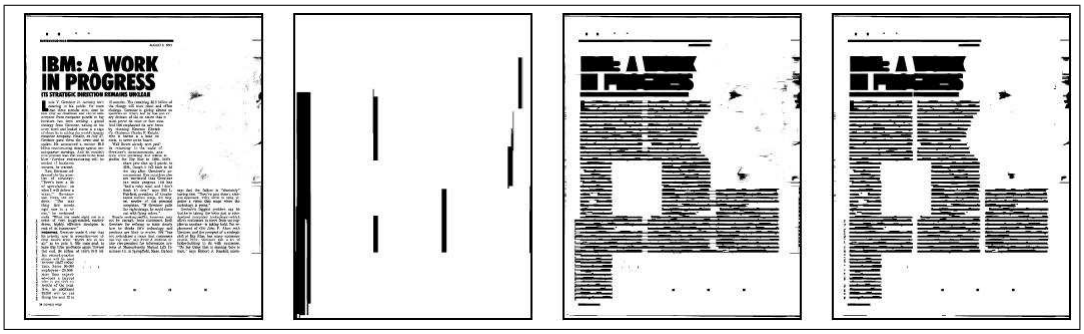


FIGURE 4.17 – Génération des texts lines [11].

4.5 Conclusion

Tous les outils que nous avons cités, visent à répondre à toutes les objectifs pour un système conçu pour analyser un document scanné. Ces outils seront utilisés pour nous aider á rendre l'utilisation facile et flexible afin de fournir un bon résultat d'analyse des document et reconnaissance de texte.

Chapitre 5

Implémentation et réalisation

5.1 Introduction

Dans cette partie de notre travail, nous allons présenter les résultats de quelques tests dans les différents cas, en illustrant avec des images les différents documents. On montrera les outils que nous avons utilisés afin de valider notre application, ainsi que les résultats des fonctions utilisés.

5.2 Les outils de développement

Pour réaliser notre application, nous avons utilisé le langage de programmation Matlab, et l'environnement de développement utilisé c'est le Matlab 2015a.

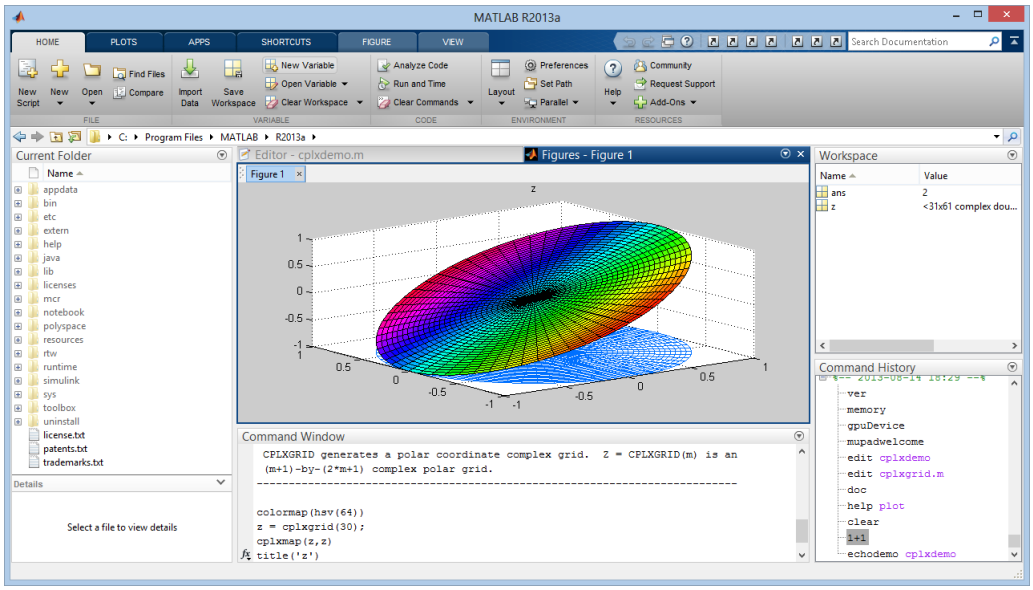


FIGURE 5.1 – Capture d'écran de l'environnement MATLAB.

5.2.0.1 Matlab et ses fonctionnalités

Le nom MATLAB vient de l'anglais "MATrix LABoratory", une traduction littérale nous amène à voir MATLAB comme un laboratoire pour manipuler des matrices [14].

Il est important de définir ce qu'est MATLAB, pour certains, c'est un logiciel, un outil, pour d'autres un langage, en fait :

- Lorsque l'on parle du logiciel MATLAB, on fait référence à l'outil que l'on utilise, l'interface utilisateur.
- Lorsque l'on parle du langage MATLAB, on désigne la syntaxe spécifique que l'on met en oeuvre dans cet outil.

MATLAB comprend de nombreuses fonctions parmi celles-ci, nous citons :

- Les calculs de traitement des données ;
- Tracés les courbes ;
- Résolution des systèmes et d'algorithmes de calculs numériques.

5.2.0.2 La syntax du Matlab

Le logiciel MATLAB est construit autour du langage MATLAB, une interface en ligne de commande, elle permet d'exécuter des commandes simples. Des séquences de commandes peuvent être sauvegardées dans un fichier texte, typiquement avec l'éditeur MATLAB, sous la forme d'un "script" ou encapsulées dans une fonction [36].

5.2.0.3 Les domaines d'application du Matlab

MATLAB comprend de nombreuses fonctions prédéfinies pour le calcul matriciel, mais pas seulement ça, en effet les domaines d'application sont extrêmement variés, on peut citer par exemple :

- Le traitement des images.
- Le calcul numérique dans le corps des réels ou des complexes.
- Le calcul de probabilités ou les statistiques.
- Le calcul intégral ou la dérivation.
- Le traitement du signal.
- L'optimisation.
- L'automatisme [14].

5.2.0.4 Les avantages du langage Matlab

Il existe plusieurs avantages du langage Matlab parmi lesquelles on retrouve :

- La capacité à traiter les images et les vidéos.
- La possibilité d'appeler d'autre bibliothèques externes, comme OpenCV.
- Une grande communauté d'utilisateurs avec beaucoup du codes libres et partages des connaissances.
- La capacité de lire une grande variété des formats des images, communs et spécifiques à un domaine.
- L'environnement MATLAB Desktop, qui vous permet de travailler de façon interactive avec vos données, et vous aide à garder les traces des fichiers et des variables, et simplifie les tâches de programmation ou débogage.
- MATLAB vous permet de tester immédiatement vos algorithmes sans recompilation, vous pouvez taper quelque chose à la ligne de commande où

d'exécuter une section dans l'éditeur et voir immédiatement les résultats, ce qui facilite grandement le développement de l'algorithme [18].

5.2.0.5 La librairie Image Processing

"Image Processing Toolbox" fournit un ensemble d'algorithmes de référence standard, des fonctions et des applications de traitement d'image, analyse et visualisation. Elle fournit aussi plusieurs opérations sur le domaine d'imagerie tel que l'analyse, la segmentation, l'amélioration de l'image, la réduction du bruit, des transformations géométriques, et enregistrement de l'image.
"Image Processing Toolbox" prend en charge un ensemble de divers types d'images, y compris aussi des fonctions de visualisation, d'examiner des régions de pixel, d'ajuster la couleur et le contraste, de créer des contours ou des histogrammes [18].

5.3 Les fonctionnalités de notre application

Dans notre application, l'utilisateur peut ouvrir n'importe quelle image document qui se trouve dans son ordinateur. Une fois l'image choisie, il a la possibilité de faire des opérations de prétraitement et d'analyse des documents puis sauvegarder l'image transformée.
La reconnaissance s'effectue en cliquant sur les parties du document analysé soit une région, ligne ou caractère.
La correction des erreurs dans le document se fera manuellement et l'utilisateur peut sauvegarder la version corrigée dans un fichier XML pour l'ouvrir ultérieurement en cas de besoin.

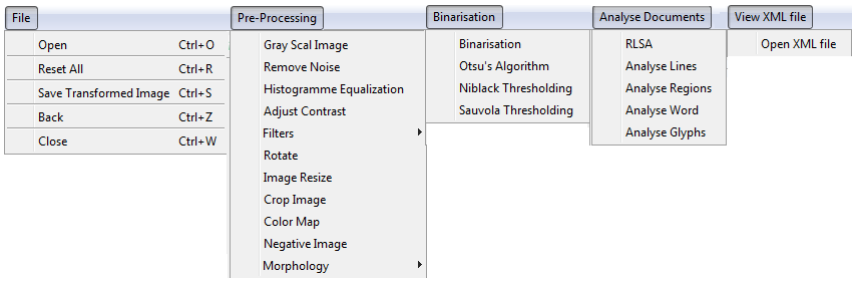


FIGURE 5.2 – La structure des menus de notre application.

Dans ce qui suit, on décrira les fonctionnalités principales de notre application :

5.3.1 Les fonctions fichier (File)

5.3.1.1 Fonction ouvrir (Open)

Elle permet d'ouvrir n'importe quelle image se trouvant dans un répertoire en utilisant un sélectionneur de fichiers de Windows et permet aussi l'affichage de l'image sélectionnée.

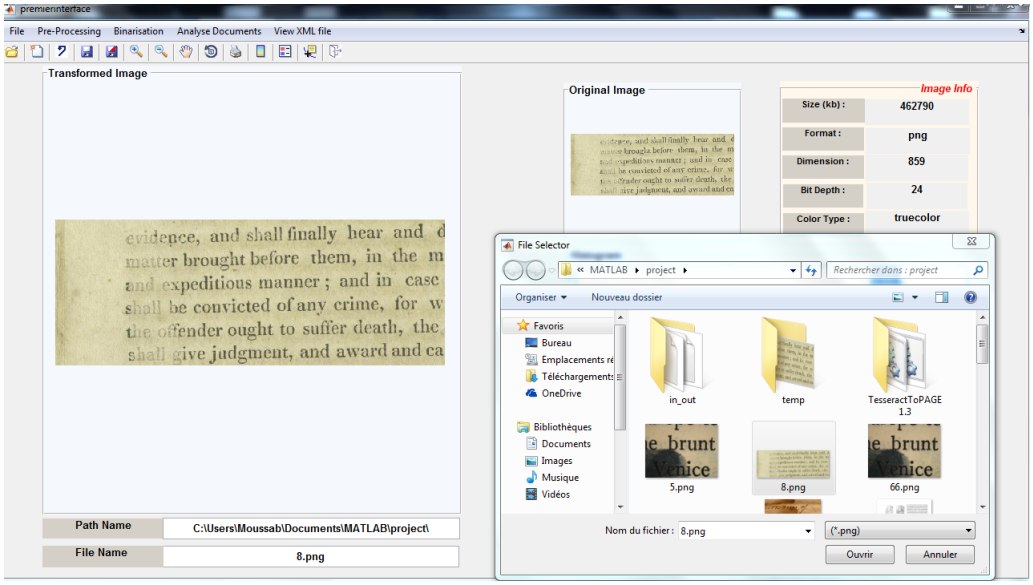


FIGURE 5.3 – Capture d'écran de la fonction ouvrir.

5.3.1.2 Fonction sauvegarder l'image transformé (Save Transformed Image)

Elle permet de sauvegarder l'image transformée dans un emplacement choisi.

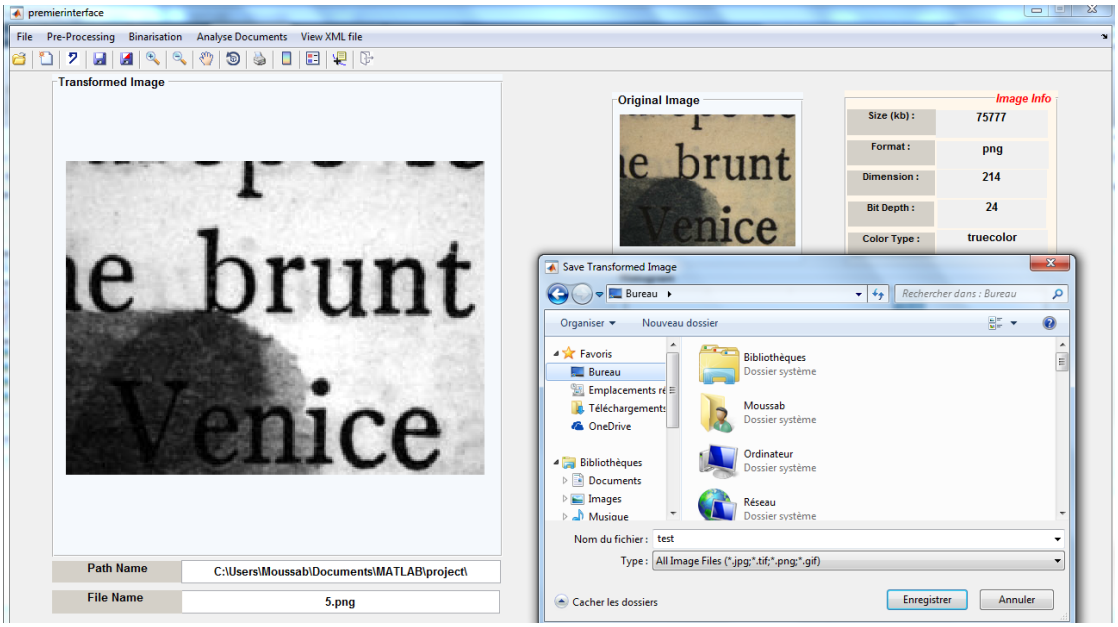


FIGURE 5.4 – Capture d'écran de la fonction sauvegarder l'image transformé.

5.3.1.3 Fonction retourner (Back)

Elle permet d'annuler l'opération courante et retourner à l'opération précédente afin d'exécuter d'autres opérations.

5.3.1.4 Fonction initialiser (Reset all)

Elle permet l'initialisation de l'application afin de lancer de nouveaux traitements.

5.3.2 Les fonctions de prétraitement

5.3.2.1 Fonction transformation en niveau de gris (Grayscale)

Elle permet la transformation d'une image couleur en niveau de gris.

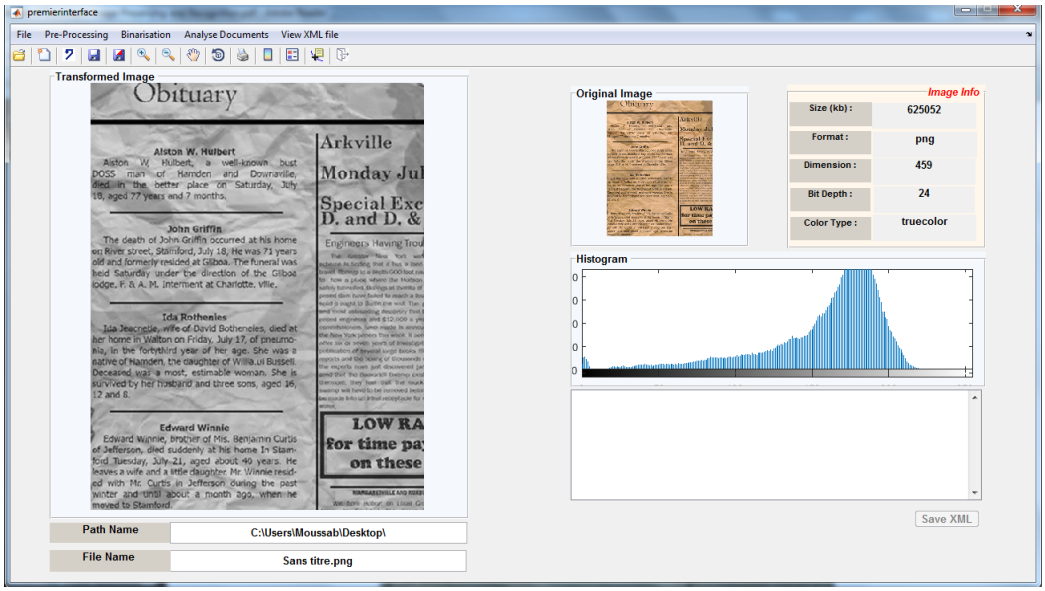


FIGURE 5.5 – Capture d'écran de la fonction transformation en niveau de gris.

5.3.2.2 Fonction suppression de bruit (Remove noise)

Elle permet la réduction du bruit dans une image.

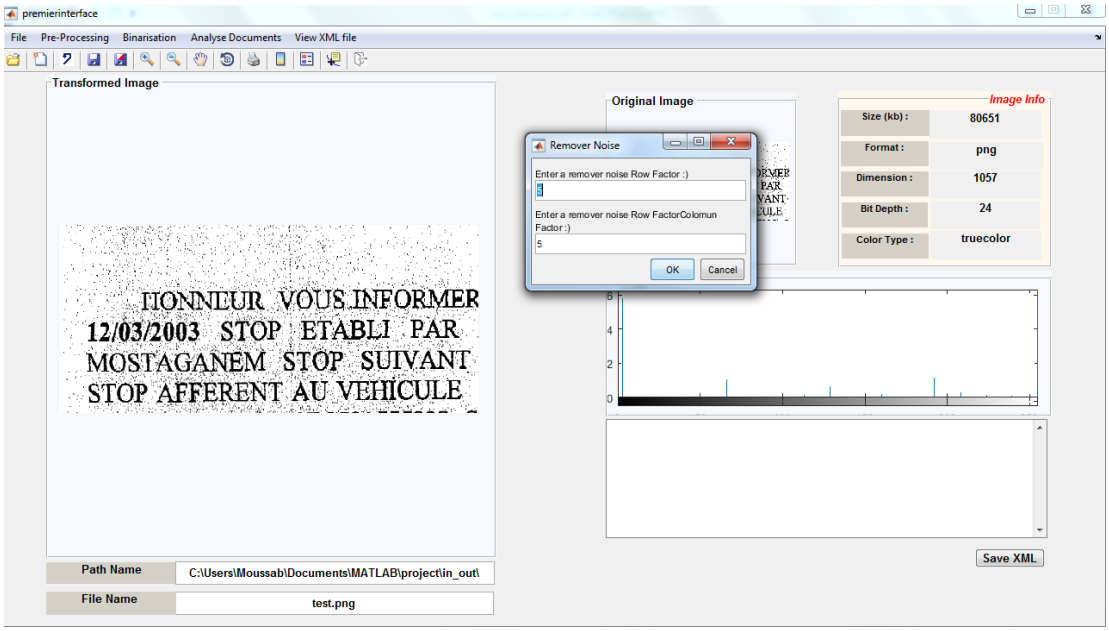


FIGURE 5.6 – Capture d'écran de la fonction réduction de bruit avec l'image originale.

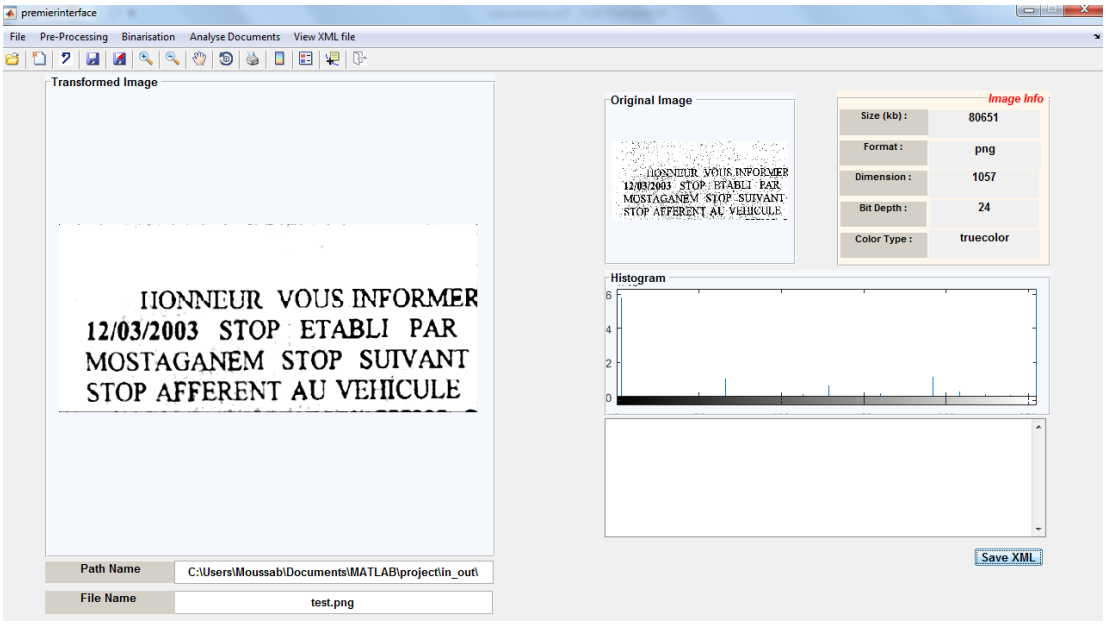


FIGURE 5.7 – Capture d'écran de la fonction réduction de bruit avec l'image résultante.

5.3.2.3 Fonction égalisation de l'histogramme (equalisation histogram)

Elle permet d'égaliser l'histogramme de l'image correspondante.

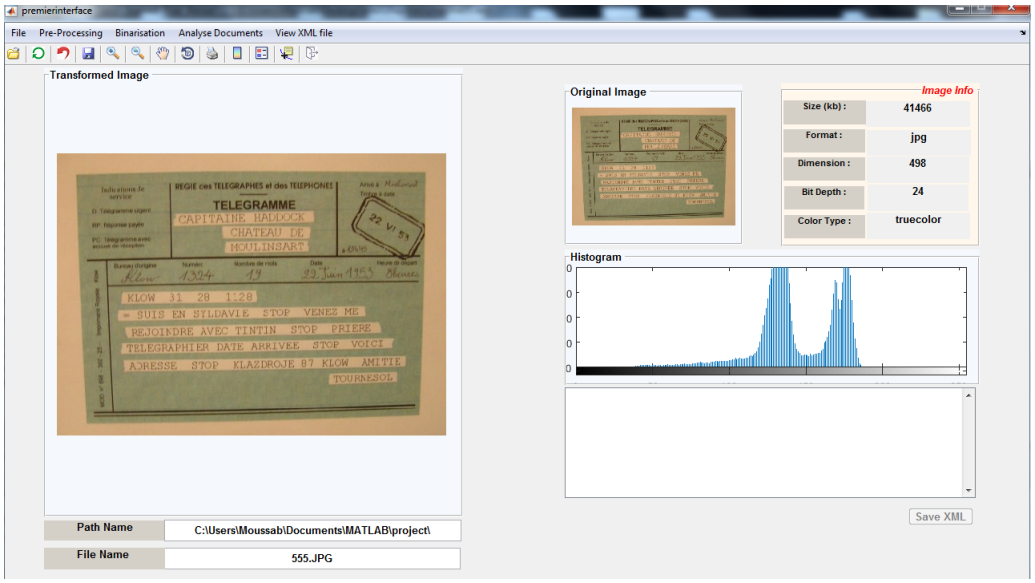


FIGURE 5.8 – Capture d'écran de la fonction égalisation de l'histogramme avec l'image originale.

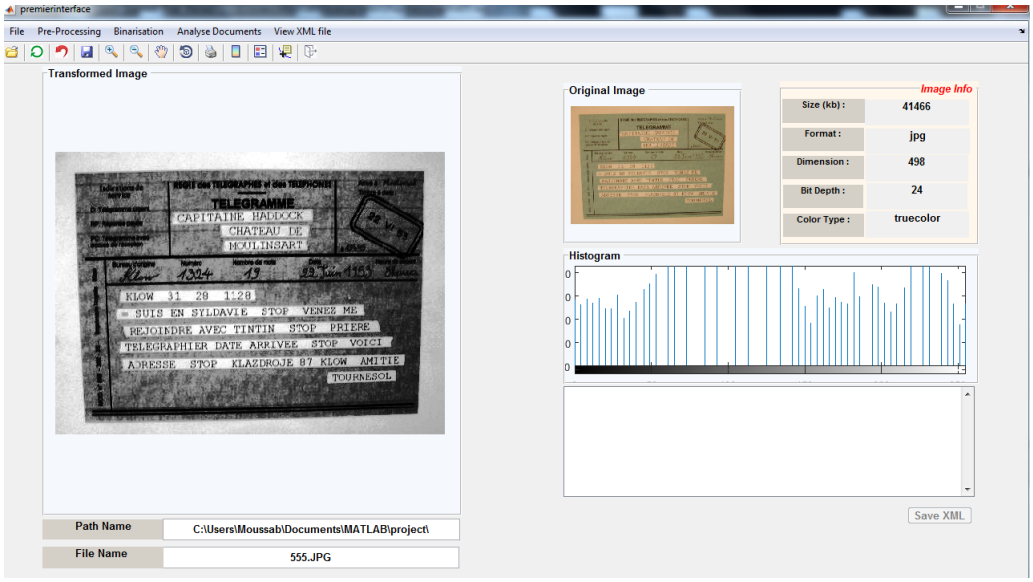


FIGURE 5.9 – Capture d'écran de la fonction égalisation de l'histogramme avec l'image résultante.

5.3.2.4 Fonction binarisation d'Otsu (Otsu Binarisation)

Elle permet la transformation de l'image au niveau de gris en image binarisée (noire et blanc) avec la méthode d'Otsu.

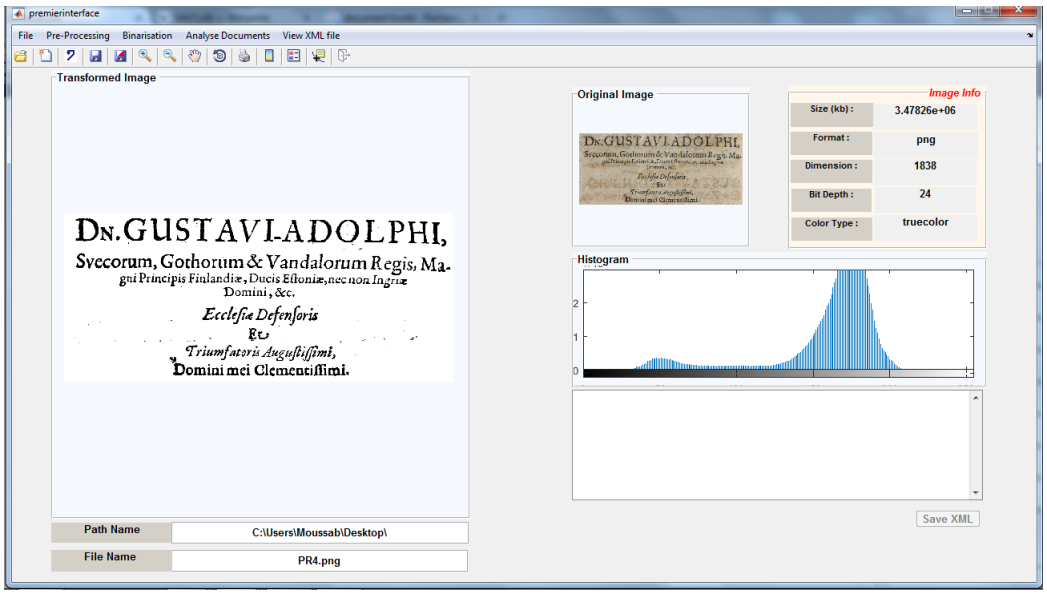


FIGURE 5.10 – Capture d'écran de la fonction binarisation d'Otsu.

5.3.2.5 Fonction binarisation de Sauvola (Sauvola Binarisation)

Elle permet la transformation de l'image niveau de gris en image binarisée (noire et blanc) avec la méthode Sauvola.

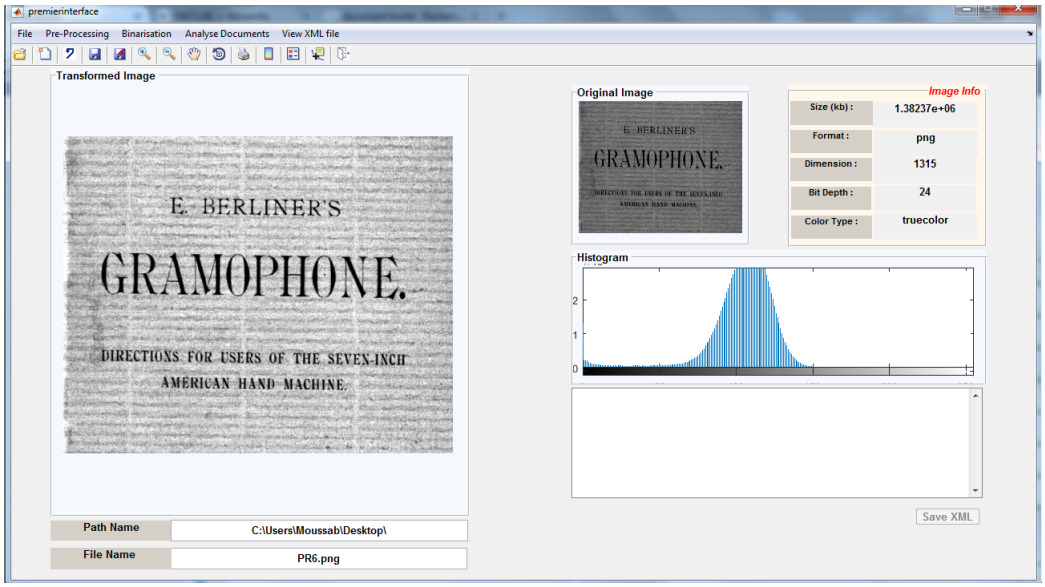


FIGURE 5.11 – Capture d'écran de la fonction binarisation de Sauvola.

5.3.2.6 Fonction binarisation de Niblack (Niblack Binarisation)

Elle permet la transformation de l'image niveau de gris en image binarisée (noire et blanc) avec la méthode Niblack.

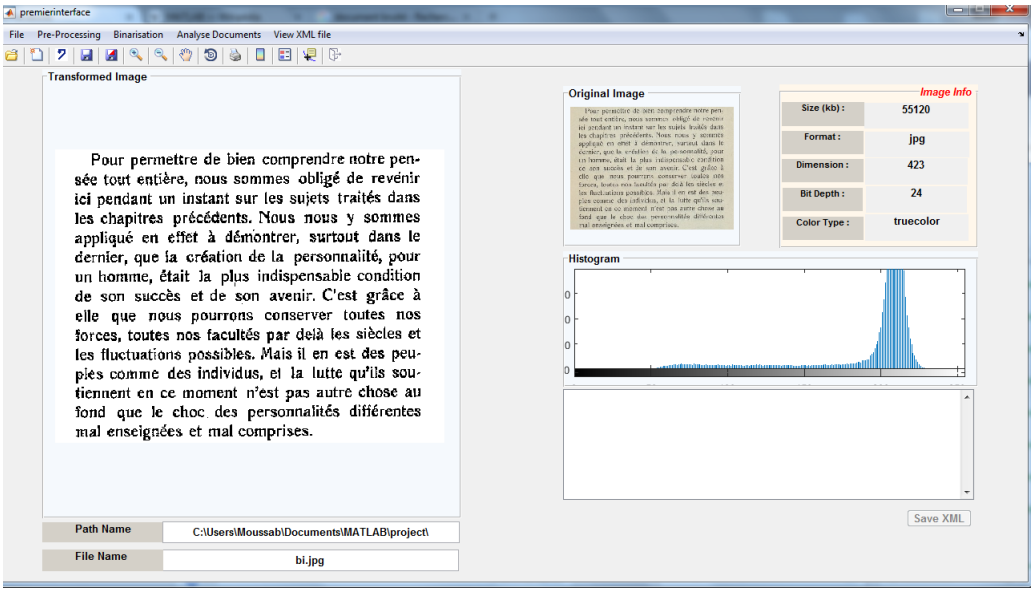


FIGURE 5.12 – Capture d'écran de la fonction binarisation de Niblack.

5.3.3 Les fonctions d'analyse des documents

5.3.3.1 Fonction d'analyser les régions (Analyse Regions)

Elle permet de faire l'analyse d'un document et d'extraire les régions.

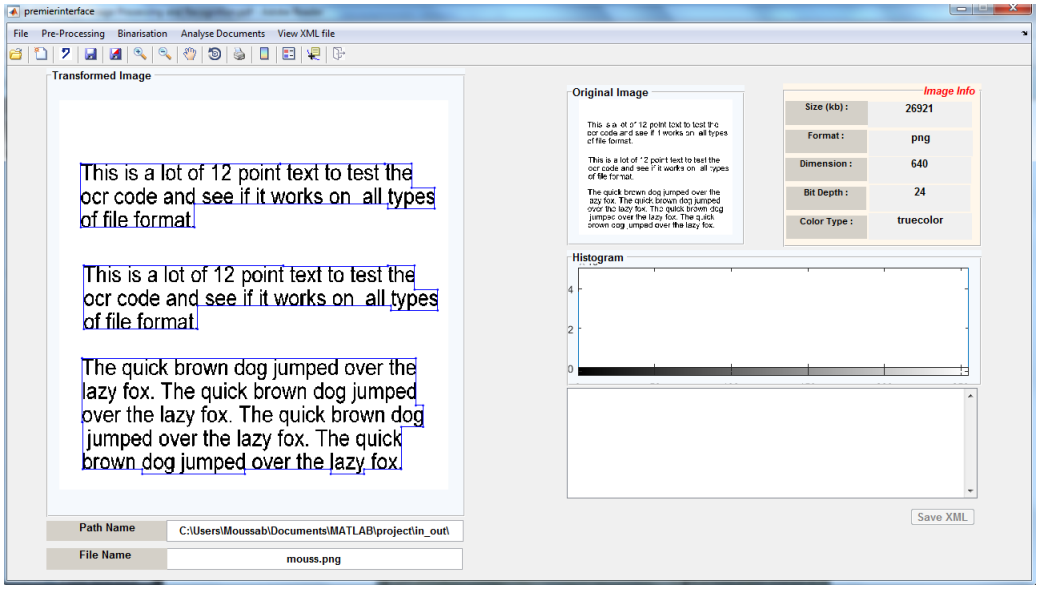


FIGURE 5.13 – Capture d'écran de la fonction analyser les régions.

5.3.3.2 Fonction d'analyser les lignes (Analyse lines)

Elle permet de faire l'analyse d'un document et d'extraire les lignes dans le document.

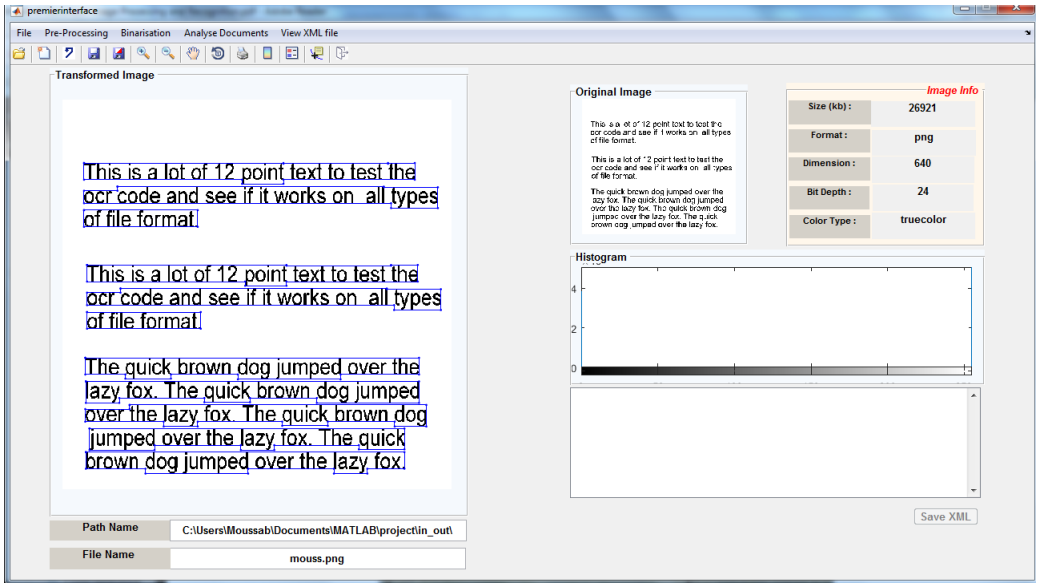


FIGURE 5.14 – Capture d'écran de la fonction analyser les lignes.

5.3.3.3 Fonction d'analyser les mots (Analyse Word)

Elle permet de faire l'analyse d'un document et d'extraire les mots dans le document

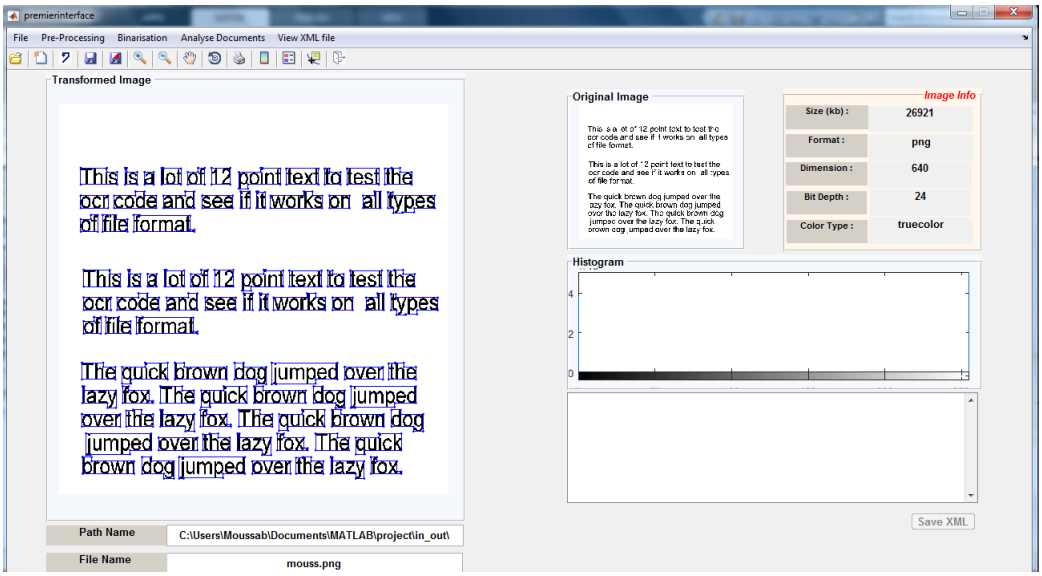


FIGURE 5.15 – Capture d'écran de la fonction analyser les mots.

5.3.3.4 Fonction d'analyser les glyphes (Analyse Glyph)

Elle permet de faire l'analyse d'un document et d'extraire les caractères dans le document

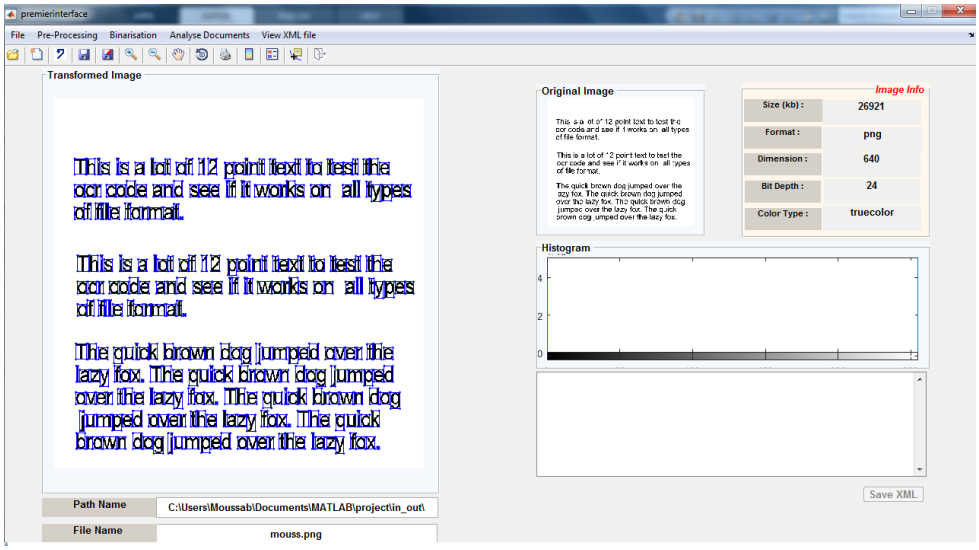


FIGURE 5.16 – Capture d'écran de la fonction analyser les glyphes.

5.3.3.5 Fonction reconnaissance des caractères (OCR)

Elle permet de faire la reconnaissance de tous les types d'analyse des documents que nous avons traités. En cliquant par exemple sur le caractère ou la région, le résultat apparaît dans une boite de texte dans le côté droit de l'application.

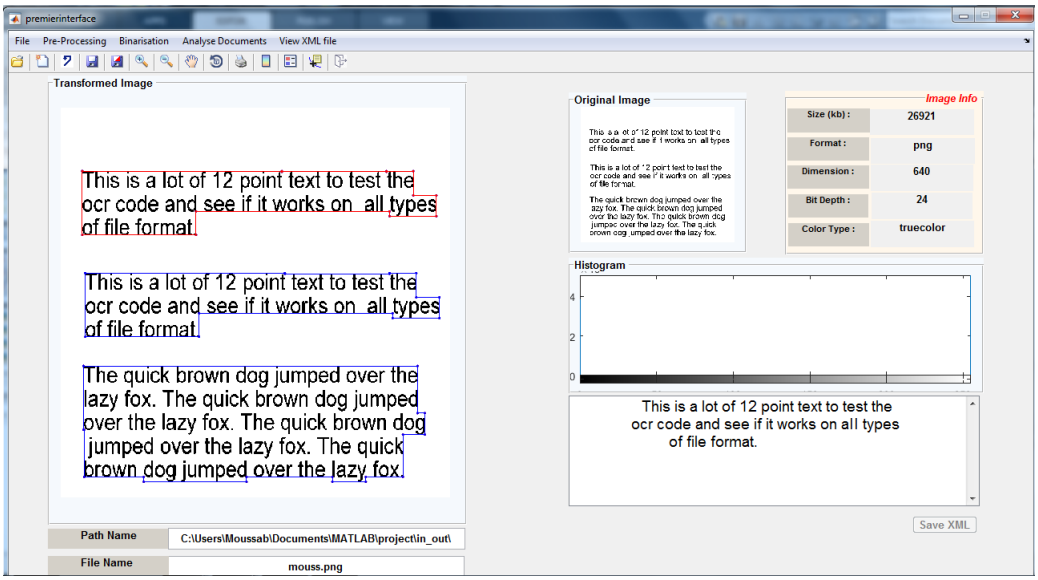


FIGURE 5.17 – Capture d'écran de la fonction Reconnaissance des caractères (regions).

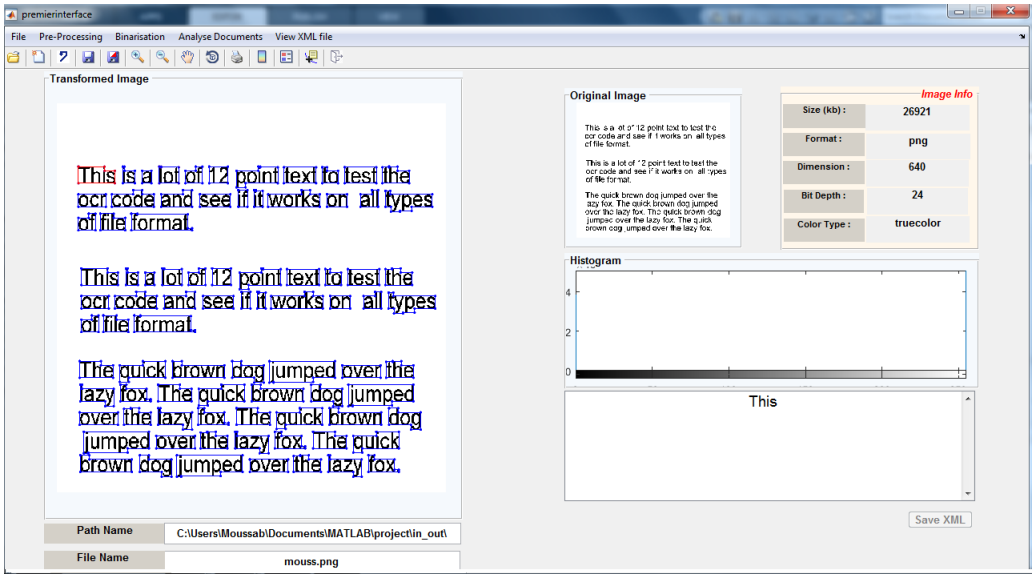


FIGURE 5.18 – Capture d'écran de la fonction Reconnaissance des caractères (mots).

5.3.3.6 Fonction sauvegarder le fichier XML (Save XML Files)

Elle permet la correction des erreurs dans le document puis sauvegarder le texte corrigé dans un fichier XML corrigé dans un emplacement choisi par l'utilisateur.

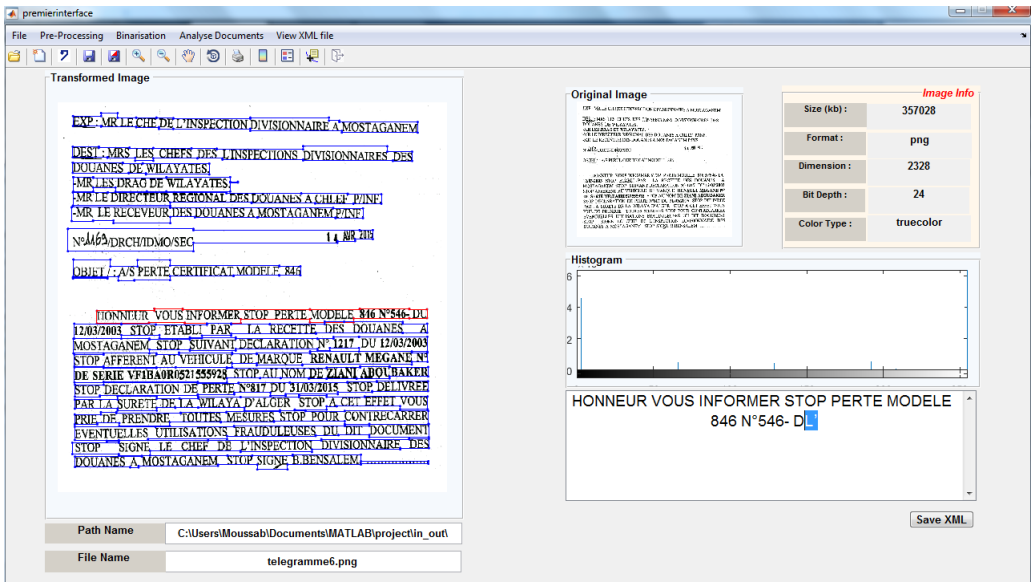


FIGURE 5.19 – Capture d'écran de la fonction Save XML Files (correction des erreurs).

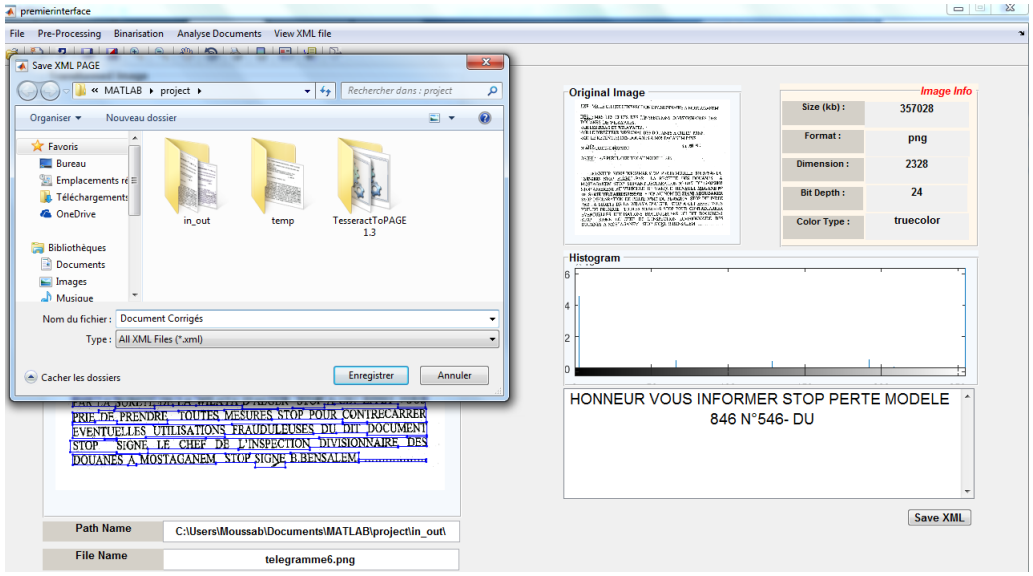


FIGURE 5.20 – Capture d'écran de la fonction sauvegarder le fichier XML (Sauvegarder).

5.3.3.7 Fonction voire les fichiers XML (View XML Files)

Elle permet de choisir le fichier XML d'un document, puis visualiser le document et choisir d'appliquer des opérations d'analyse.

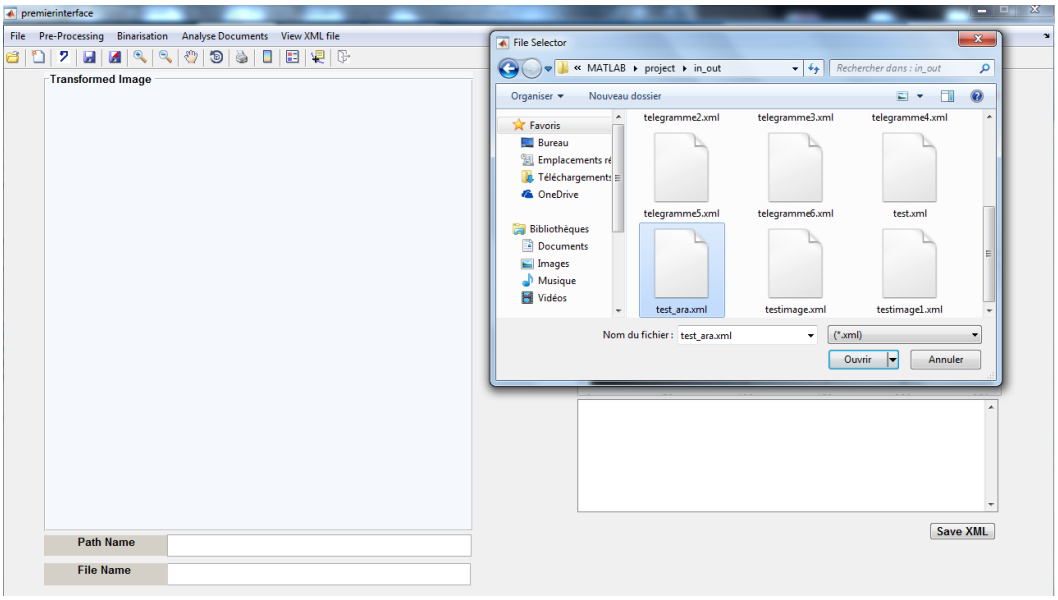


FIGURE 5.21 – Capture d'écran de la fonction voir les fichiers XML (ouvrir).

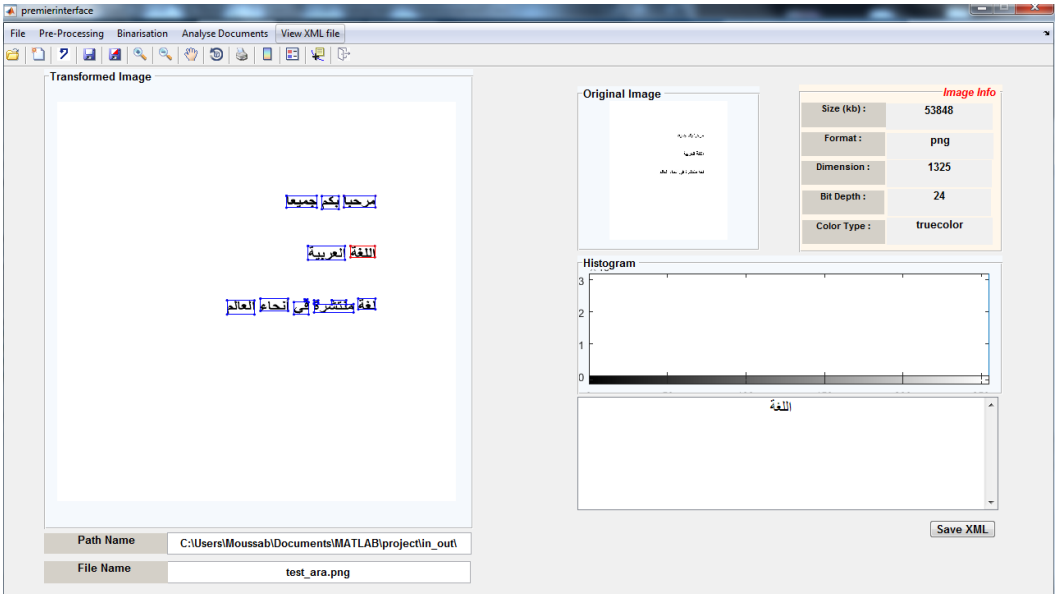


FIGURE 5.22 – Capture d'écran de la fonction voir les fichiers XML (appliquer les opérations).

5.3.4 Exemple d'un déroulement des fonctions sur un document

Voici un exemple sur le déroulement du processus de prétraitement et analyse et reconnaissance d'un document :

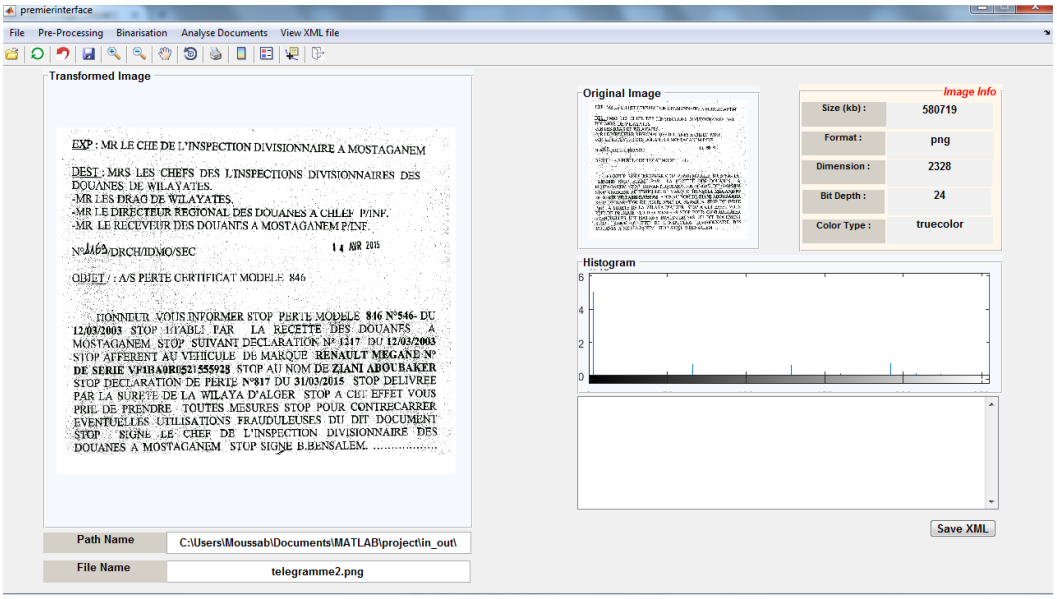


FIGURE 5.23 – Capture d'écran du fonction lecture et l'affichage d'un document.

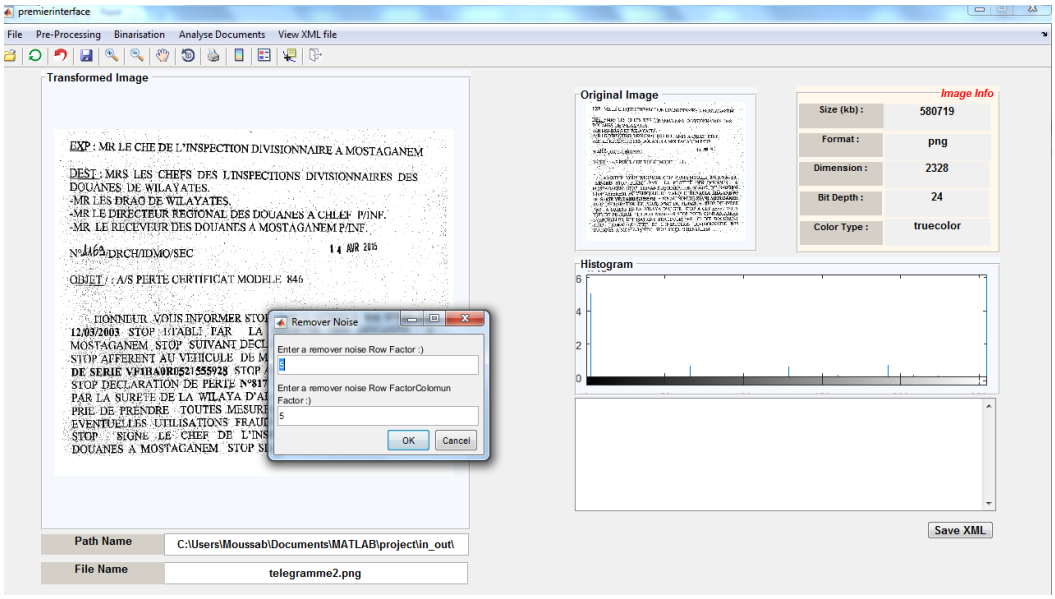


FIGURE 5.24 – Capture d'écran du fonction suppression du bruit.

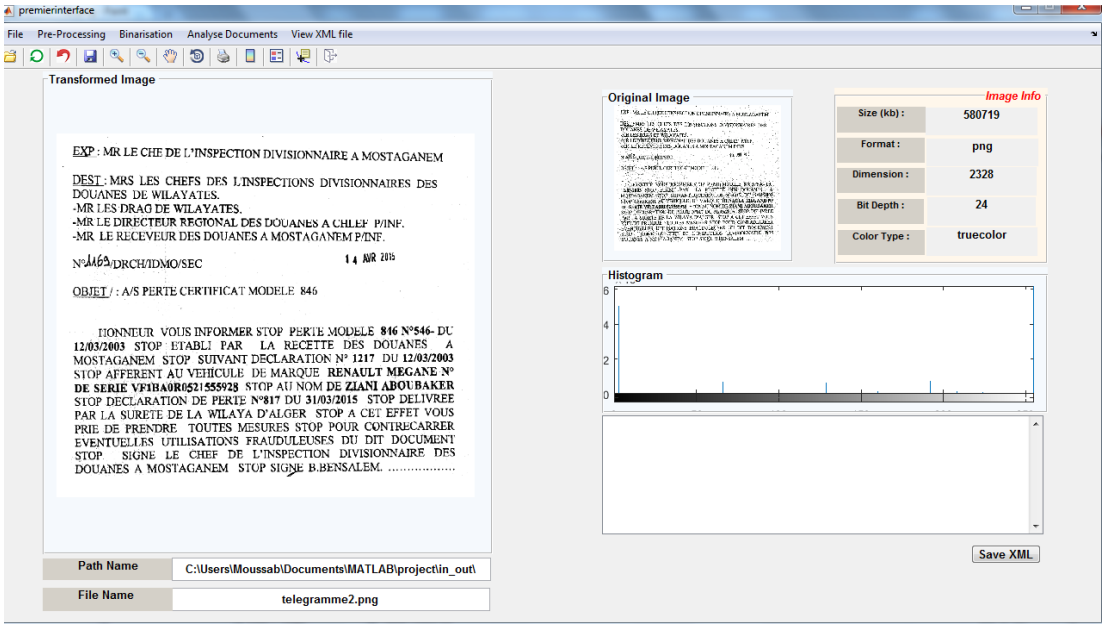


FIGURE 5.25 – Capture d'écran du document résultant.

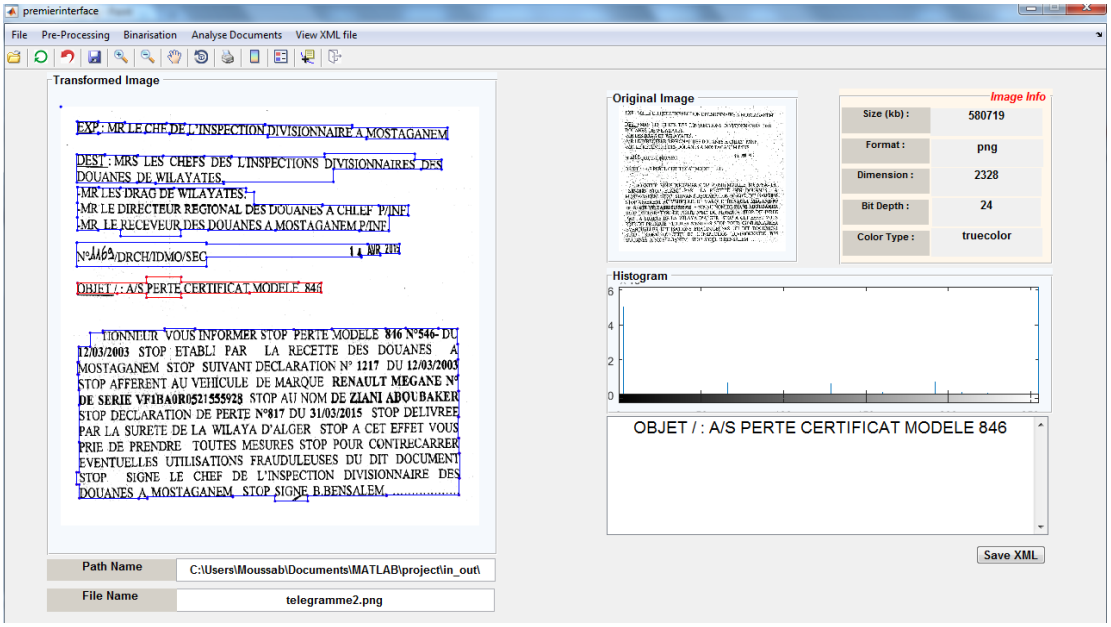


FIGURE 5.26 – Capture d'écran des fonctions analyse et reconnaissance.

5.4 Conclusion

Dans ce chapitre, nous avons montré les résultats de l'application implémentée sur différents types de documents. Nous avons introduit et analysé les fonctions menées au cours de développement de notre application. L'analyse des résultats a été effectuée par une inspection visuelle des documents transformés afin de voir tous les effets produits.

Chapitre 6

Conclusion et Perspectives

Ce travail nous a montré que le domaine du traitement automatique de documents est un vaste chantier à explorer.

Notre travail nous a donc permis de comprendre comment on peut extraire une bonne information à travers un document texte bruité et mal scanné, à l'aide d'un processus d'analyse et de reconnaissance des caractères, afin d'utiliser l'information extraite pour d'autres expériences.

L'implémentation et la réalisation de ce projet n'était pas facile. A la fin nous avons réussi à produire une application opérationnelle, prête à être utilisée. Notre seul regret, c'est que nous n'avons pas pu réaliser toutes les fonctionnalités par contrainte de temps.

En fin, les futures perspectives envisagées peuvent être résumés dans les points suivants :

- La classification des documents.
- Etudier le processus de reconnaissance avec d'autre langues, notamment la reconnaissance de la langue arabe.
- La prise en charge d'autres formats des documents (PDF).
- Étudier les nouvelles approches dans le domaine d'analyse et la segmentation des documents.

– La Prise en charge des autres types des documents (historiques, scientifiques ,...)

Le présent travail, étant réalisé par un humain, il n'est pas exempté d'imperfection. Ainsi, les remarques et les suggestions pour son amélioration sont donc les bienvenues.

Bibliographie

[1] G. Borgefors, I. Nyström, G. S. Di. Baja. Computing skeletons in three dimensions. *J.Pattern recognition*, 32(7) : 1225-1236, 2008.

[2] B. B. Chaudhuri. *Digital document processing : major directions and recent advances*. Springer Science and Business Media, 2007.

[3] C. Clausner, S. Pletschacher, A. Antonacopoulos. Aletheia-an advanced document layout and text ground-truthing system for production environments. *International Conference on Document Analysis and Recognition (ICDAR)*, 48-52, 2011.

[4] Dictionnaire sensagent : http ://dictionnaire.sensagent.com/prC3A9traitement +d+image/fr-fr/

[5] D. Doermann, K. Tombre. *Handbook of Document Image Processing and Recognition*. Springer, 2014.

[6] M. V. Droogenbroeck. Traitement d'images numériques au moyen d'algorithmes utilisant la morphologie mathématique et la notion d'objet : application au codage. *Thèse Doctorat*, 1994.

[7] S. Ferilli. *Automatic Digital Document Management*. Springer Science and Business Media, 2011.

[8] J. L. Fisher, S. C. Hinds, D. P. Amato. A rule-based system for document image segmentation. *10th International Conference on Pattern Recognition*, 567-572, 1990.

[9] C. Fouard, G. Malandain, S. Prohaska, M. Westerhoff, F. Cassot. Squelettisation par blocs pour des grands volumes de données 3D. *J. Microcirculation*, 13(1) :1-18, 2006.

[10] B. Gatos, I. Pratikakis, S. J. Perantonis. Efficient binarization of historical and degraded document images. *The Eighth International Workshop on Document Analysis Systems (DAS)*, 447-454, 2008.

[11] J. Goutsias, L. Vincent, D. S. Bloomberg. *Mathematical morphology and its applications to image and signal processing*. Springer Science and Business Media, 2006.

[12] R. C. Gonzalez, R. E. Woods. *Digital Image Processing*. Prentice hall Upper Saddle River, 2002.

[13] P. Héroux. *Contribution au problème de la rétro-conversion des documents structurés*. Thèse de doctorat, Rouen, 2001.

[14] Initiation MATLAB :http ://nte.mines-albi.fr/MATLAB/co/Generalites.html

[15] K. Kise, A. Sato, M. Iwata. Segmentation of page images using the area Voronoi diagram. *J.Computer Vision and Image Understanding*, 70(3) :370-382, 1998.

[16] M. Krishnamoorthy, G. Nagy, S. Seth, M. Viswanathan. Syntactic segmentation and labeling of digitized pages from technical journals. *J.Pattern Analysis and Machine Intelligence*, 15(7) :737-747, 1993.

[17] F. Lebourgeois, H. Emptoz. A fast and efficient method for extracting text paragraphs and graphics from unconstrained documents. *11th International Conference on Pattern Recognition*, 272-276, 1992.

[18] MathWorks : http ://www.mathworks.com

[19] T. B. Moeslund. *Introduction to video and image processing : Building real systems and applications*. Springer Science and Business Media, 2012.

[20] G. Nagy, S. Seth. Hierarchical representation of optically scanned documents. *International Conference on Pattern Recognition*, 347-349, 1984.

[21] W. Niblack. *An introduction to digital image processing*, Strandberg Publishing Company, 1985.

[22] L. O'Gorman. The document spectrum for page layout analysis. *J.Pattern Analysis and Machine Intelligence*, 15(11) :1162-1173, 1993.

[23] N. Otsu. A threshold selection method from Gray-level histograms. *J.Trans Syst Man Cybern*, 9(1) :377-393, 1979.

[24] PRImA Research Lab. *Tesseract to PAGE User Guide*. PRImA, 2014.

[25] PRImA Research Lab. *Aletheia Introduction*. PRImA, 2015.

[26] P. Rey. *Le Traitement d'Images Avec Silverlight 5*. Gambh, 2012.

[27] J. C. Russ. *The image processing handbook*. CRC press, 2011.

[28] T. Saitoh, T. Pavlidis. Page segmentation without rectangle assumption. *11th International Conference on Pattern Recognition*, 277-280, 1992.

[29] J. Sauvola, M. Pietikainen. Adaptive document image binarization. *J.Pattern Recognit*, 33(2) :225-236, 2000.

[30] Y. Solihin, C. G. Leedham. Integral ratio : a new class of global thresholding techniques for handwriting images. *J.Pattern Analysis and Machine Intelligence*, 21(8) :761-768, 1999.

[31] S. N. Srihari, G. W. Zack. Document image analysis. *International Conference on on Pattern Recognition*, 434-436, 1986.

[32] B. Su, S. Lu, U. Pal, C. L. Tan. An effective staff detection and removal technique for musical documents. *10th IAPR International Workshop on Document Analysis Systems (DAS)*, 160-164, 2012.

[33] Y. H. Tseng, H. J. Lee .Document image dewarping using robust estimation of curled text lines. *J.Pattern Analysis and Applications*, 11(1) :33-44, 2008.

[34] N. Vandenbroucke. Cours traitement d'images sous Matlab. 2009.

[35] N. Vandenbroucke. Cours traitement d'images. 2009.

[36] Wikipedia MATLAB : https ://fr.wikipedia.org/wiki/MATLAB