République **A**lgérienne **D**émocratique et **P**opulaire
Ministère de l'**E**nseignement **S**upérieur et de la **R**echerche **S**cientifique

UNIVERSITE **M**OULOUD **M**AMMERI DE **T**IZI-**O**UZOU

FACULTE DU **G**ENIE **E**LECTRIQUE ET D'INFORMATIQUE
DEPARTEMENT D'INFORMATIQUE

Mémoire de Fin d'Etudes de MASTER ACADEMIQUE

Domaine : **Mathématiques et Informatique**
Filière : **Informatique**
Spécialité : **Réseau, Mobilité et Systèmes Embarqués**

Présenté par
Hacène BELKACEMI

Thème

Extraction automatique de métadonnées à partir d'images de manuscrits anciens numérisés

Mémoire soutenu publiquement le 29/09/ 2016. devant le jury composé de :

Président : M. HAMMACHE Arezki

Encadreur : M. SOUALAH Mohammed Ou Rabah

Examinateur : M. HABET Mohammed-Saïd

Examinateur : M. DEMRI Mohamed

Dédicace :

Je dédie ce modeste travail à toute personne qui le lira, et à toute personne qui le poursuivra.

Hacène

Résumé :

Dans notre travail nous traitons un problème d'actualité relatif au traitement d'image : l'extraction automatique de métadonnées à partir d'images de manuscrits anciens numérisés.

Par conséquent, nous avons mis en œuvre un système d'extraction du nombre de lignes en utilisant un algorithme original (segmentation RLSA) auquel nous avons apportés des modifications qui se sont avérées pertinentes.

Un second système concernant l'extraction de figures sur des images de manuscrits anciens numérisés à l'aide de traitements usuels d'images a été implémenté Des résultats probants ont montrés la possibilité d'utiliser les outils de traitement d'images pour l'extraction automatique de métadonnées.

Soammaire

Introduction générale

La numérisation des documents manuscrits anciens est un besoin crucial au niveau de toutes les bibliothèques. Les projets de numérisation n'a à aucun moment cessé d'augmenter. Bien au contraire, il semble que le mot d'ordre des bibliothèques virtuelles est la préservation des manuscrits anciens via leur numérisation.

Les manuscrits anciens sont des trésors inestimables pour la mémoire de l'humanité..

La conservation des manuscrits anciens dans des lieux sécurisés et adéquats n'empêche nullement leur dégradation: humidité, risques d'incendies, et le risque d'endommagement du papier lors de leurs consultations, présence de parasites et de bactéries qui rongent le papier provoquant parfois, des dommages irréversibles.

La numérisation réduit le risque de perte des documents en créant des versions numériques duplicables. Elle permet également de répondre au besoin de l'accessibilité visé par la création de bibliothèques numériques. La représentation numérique des documents nécessitent des moyens et approches spécifiques pour parvenir à traiter, à indexer et à accéder au contenu des images de ces documents.

[Jou 09] rapporte que seulement un taux de 30 % des pixels de ces images contiennent les informations essentielles (textuelles, graphiques).Pour extraire ces pixels significatifs, nous avons recours à des approches d'accès au contenu.

L'accès au contenu de l'image d'un manuscrit numérisé se donne pour objectif l'extraction de métadonnées (cf. page 9). Cet objectif est réalisable par l'accessibilité aux informations et aux données sources.

L'extraction automatique de métadonnées à partir de ressources numériques est un puissant outil qui permet d'ajouter une couche sémantique au manuscrit après interprétation de la métadonnée extraite.

En effet, le catalogue est un puissant outil qui permet d'accéder à un manuscrit, mais l'acte de catalogage s'avère très fastidieux et dans la majorité des cas, pénible et répétitif pour certaines tâches. Nous citons entre autre le comptage du nombre de ligne, qui semble être une tâche très ennuyeuse pour le catalogueur. A ce niveau d'analyse vient la problématique de notre travail :

Problématique : Comment utiliser l'outil de traitement d'images pour venir en aide au catalogueur ? En d'autre terme, nous pouvons se poser une autre question en disant, est-il possible d'utiliser le traitement d'image pour éviter au catalogueur d'exécuter les tâches répétitives ? D'une manière implicite, est-il possible de compter automatiquement le nombre de lignes ? Est-il possible de détecter la complétude d'un manuscrit ? Est-il possible de statuer automatiquement sur l'état du manuscrit ?

Notre travail porte sur l'étude des images de manuscrit anciens numérisés. Nous nous intéressons à dénombrer automatiquement le nombre de lignes d'une page d'un manuscrit et à extraire des figures présente éventuellement, dans le manuscrit. Ces différents aspects nous ont emmené à étudier les différentes techniques de traitements d'images appliquées sur les images de documents, nous citons entre autre les prétraitements et la segmentation :

- Les prétraitements : ce sont les premiers traitements appliqués sur l'image. Ils consistent à améliorer la qualité des images en éliminant certains défauts comme les taches, et à les simplifier en suppriment l'information inutile pour l'analyse ultérieure de l'image. Les prétraitements qu'on applique sur les images de documents sont les modifications d'histogramme, les filtrages et les opérateurs morpho mathématiques.
- La segmentation : Elle se donne pour objectif de fournir une description des objets contenus dans l'image à travers l'extraction de diverses indications visuelles telles que les contours des objets ou les régions homogènes. Les techniques de segmentation sont divisées en quatre classes : la binarisation, la segmentation ascendante, la segmentation descendante et la segmentation par texture.

Dans ce mémoire, nous avons mis en œuvre un système d'extraction du nombre de lignes dans une image de manuscrit ancien numérisé. Nous avons également, mis en œuvre un système d'extraction de figures.

Notre mémoire est composé de cinq chapitres :

- Dans le premier chapitre, nous avons défini des généralités sur les manuscrits anciens, leurs modèles de dégradations qui peuvent survenir tout au long de la durée de vie du manuscrit ainsi que les difficultés structurelles que peuvent présenter les manuscrits anciens. Nous avons défini l'extraction de métadonnées et la différence entre l'extraction manuelle et automatique ;
- Dans le deuxième chapitre, nous avons présenté les différentes techniques de traitement d'images. Nous avons d'abord élaboré les principaux prétraitements appliqués sur les images de documents. Nous avons par la suite défini la segmentation avec ses différentes techniques ;
- Dans le troisième chapitre, nous avons décrits quelques méthodes de segmentation de texte, notamment la binarisation d'Otsu, le RLSA et la technique de projection horizontale ;
- Dans le quatrième chapitre, nous avons conçu un système d'extraction du nombre de lignes et un système d'extraction de figures à partir d'images de manuscrits anciens numérisés, en exploitant les différentes techniques de prétraitement et de segmentation traitées dans les chapitres précédents et en modifiant la méthode du RLSA ;
- Dans le cinquième et dernier chapitre, nous avons mis en œuvre les deux systèmes dans le but de les évaluer.

Partie I : Etat de l’Art

Chapitre I : Généralités

1. Introduction :

Les manuscrits anciens sont considérés comme un trésor et un héritage culturel au vue de leur importance sur les divers aspects scientifiques, historiques et archéologique. Cette importance justifie la nécessité d'accès et d'analyse du contenu par les utilisateurs. Cet accès peut altérer la qualité du manuscrit. Par conséquent, on a recours à la définition d'autres méthodes d'accès. Avec la numérisation des documents manuscrits, L'accès est sûr et distant. Pour accéder au contenu des images du manuscrit numérisé nous devons définir un ensemble de méthodes nécessaires en utilisant le traitement d'images. D'où l'usage de l'extraction automatique de métadonnées par l'analyse du contenu des images en vue d'une éventuelle indexation des manuscrits.

Ce chapitre est consacré à introduire la notion de métadonnée, l'extraction de métadonnées et à faire une analyse des manuscrits anciens avec leurs dégradations.

2. Les manuscrits anciens :

2.1.Définition du manuscrit:

Un manuscrit est un texte écrit à la main, sur un support souple (papier, papyrus, parchemin) par son auteur ou par un copiste[1].

2.2.Catégories de documents manuscrits :

On peut classer les documents manuscrits dans quatre catégories en fonction de leur aspect physique, indépendamment de la langue [Ouw 10] :

a) *Documents mono-orientés* : les lignes sont orientées dans une seule direction.

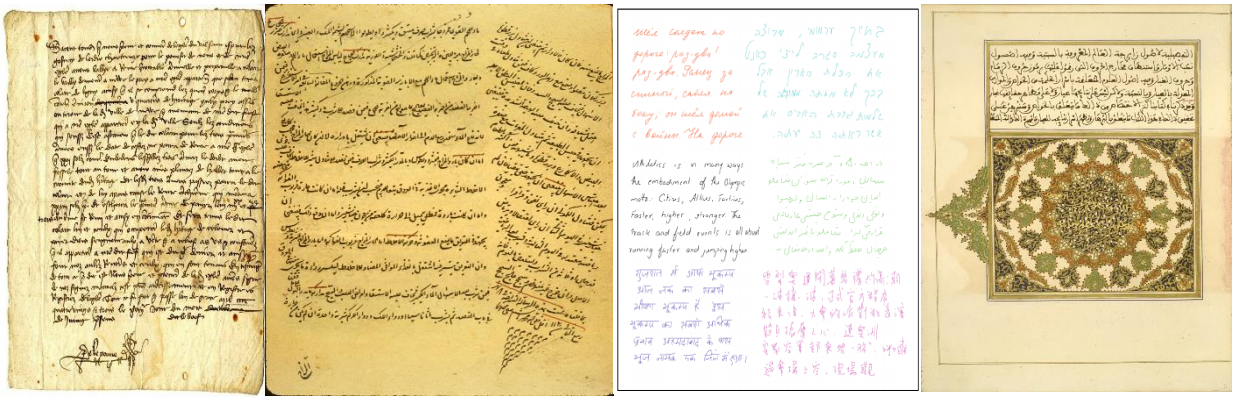
b) *Documents multi-orientés* : les lignes sont rangées par blocs ou elles peuvent occuper toute la largeur du document. Elles ont des orientations différentes.

c) *Documents multi-scripts* : ce sont des documents écrits par plusieurs personnes différentes, conduisant à des écritures ou scripts différents, éventuellement sous différents langues. Dans le passé, des personnes collaboraient pour écrire un même document ou un même ouvrage. Les cartes de vœux sont un exemple de documents multi-scripts modernes.

d) *Documents hétérogènes* : ce type contient du texte et des images ou illustrations. Beaucoup de documents cartographiques, mécaniques ou architecturaux sont de ce type.

La figure 1.1 montre un exemple de chacune de ces catégories de documents.

[1] Professionnel chargé de la reproduction de documents écrits.



(a) (b) (c) (d)

Fig. 1.1 : Exemples des quatre catégories de documents manuscrits :

(a) Français mono-orientés[Dis], (b) Arabes multi-orientés[IMMa 04],

(c) multi-scripts (Quatre différentes langues)[NJ 04] et (d) hétérogènes[IMMb 04].

Le vieillissement des documents manuscrits les expose aux dégradations. Ajoutés à cela d'autres facteurs de leur contenu comme les trous, les taches, l'apparence des écritures du coté verso dans le coté recto, les parasites, le contact entre les lignes de texte et les styles d'écritures. Ainsi la procédure d'extraction de métadonnées devient encore plus difficile [PYR 15].

2.3.Modèles de dégradation d'images de documents :

Nous présentons les dégradations les plus couramment observées dans les documents manuscrits. Cette étude permet de montrer l'impact de ces dégradations sur le bon fonctionnement d'algorithmes de traitement et d'analyse de documents [Kie 14].

2.3.1. Catégories de dégradations :

La dégradation du document peut intervenir à n'importe quel moment, depuis la conception du document original jusqu'à la phase de numérisation. Elle peut être classée en deux catégories [Kie 14]:

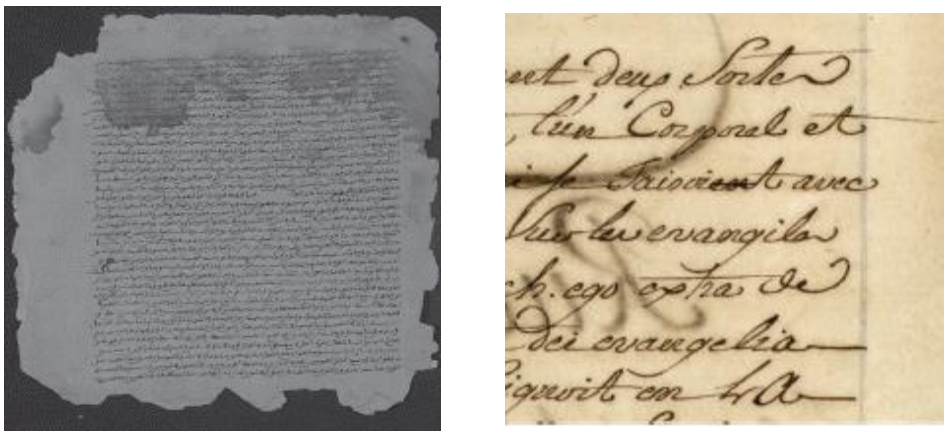
- La dégradation propre au document et à sa condition de conservation ;
- La dégradation due à la numérisation.

2.3.1.1.Les dégradations propres au document et à sa condition de conservation :

Le document ou l'ouvrage peut être altéré lors :

- De sa création (qualité du papier, de l'encre ou de la presse) ;

- De son utilisation (transport, lecture, ajout d'*annotations*[2]) ;
- De sa conservation (conditions de stockage et de manipulation).



(a) (b)

Fig. 1.2 : Exemple de documents dégradés [Kie 14] :

(a) taches d'humidité en haut de la page ; (b) transparence de l'encre

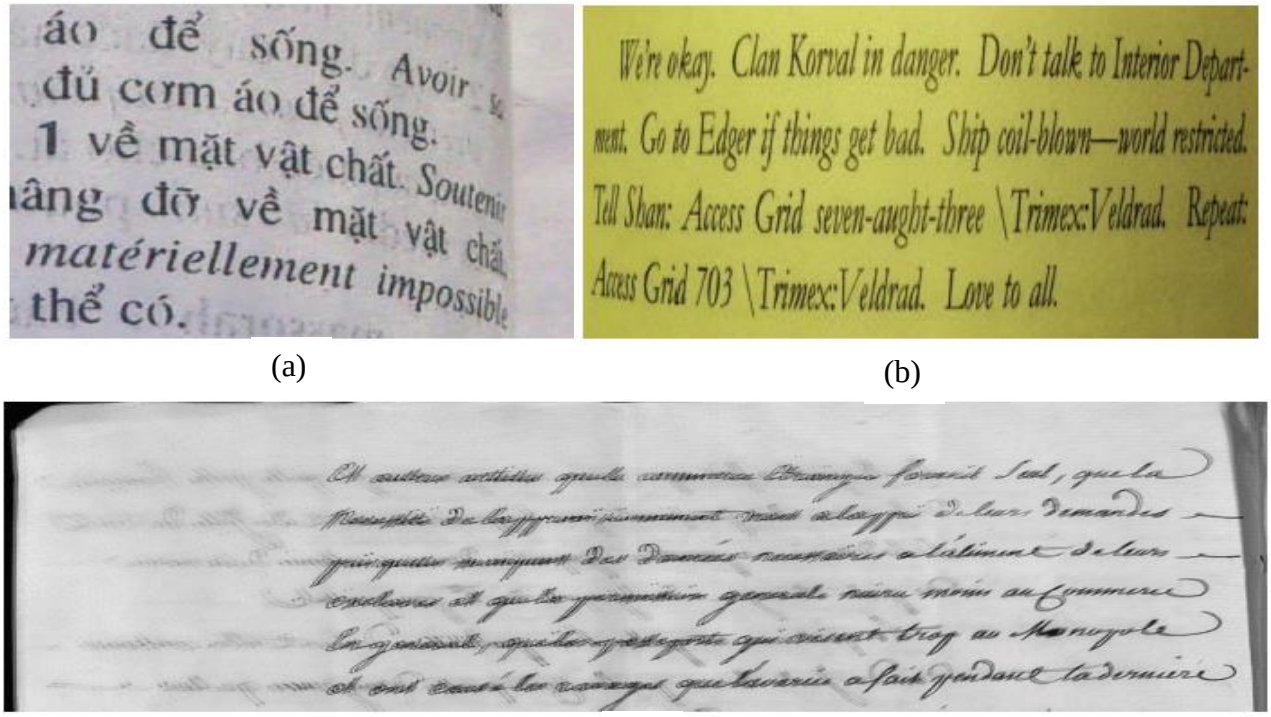
2.3.1.2.Les dégradations dues à la numérisation :

Les défauts principaux dus au processus de numérisation apparaissent dans les situations suivantes :

- un mauvais choix de la technologie du scanner pour numériser un document avecreliure épaisse (impression visuelle de courbure de la page);
- la distorsion de perspective (la surface de la page n'est pas totalement plane)
- une mauvaise gestion de la luminosité ambiante (éclairage autour du scanner, distorsion importante du papier) ;
- le flou causé par la mauvaise gestion du scanner ou le mouvement du document pendant sa numérisation ;
- caractéristiques techniques et physiques du scanner ajoutant un bruit ou parasite. Par exemple le bruit thermique lié à la température de capteurs.
- Le mauvais positionnement (translation ou rotation de la page) ;
- Page pliée suite à une mauvaise manipulation ;
- Oubli d'un objet sur l'ouvrage (exemple : marque page);

[2] Action de faire des remarques sur un texte pour l'expliquer ou le commenter [Lar].

- Présence de doigts.



(a) (b)

(c)

Fig. 1.3 : Quelques défauts liés au processus de numérisation [Kie 14]: (a) mauvaise gestion de reliure épaisse ; (b) distorsion de perspective liée au système optique ; (c) flou.

2.3.2. Niveaux de dégradations :

Selon leur forme, leur localisation et leur rendu visuel, les dégradations peuvent être [Kie 14]:

- Globales : ces dégradations se trouvent sur l'ensemble des éléments du document (distorsion globale du papier, luminosité non uniforme, etc.) ;
- Locales : elles affectent quelques éléments du document (déformation des caractères, trous dans l'encre, taches, etc.) ;
- Diffuses : elles se rapportent plus à une impression visuelle (taches d'humidité, apparition du verso sur le recto par transparence, etc.).

2.4.Difficultés structurelles présentées par les manuscrits anciens:

Les manuscrits anciens soulèvent trois types de difficultés [Bou] :

- La mise en page de ces documents peut être complexe et présenter plusieurscolonnes de taille de corps et d'interlignes différents ;
- L'inévitable courbure des lignes de texte produite par la reliure des livres ;

- Les faibles espaces entre les lignes qui entraînent de nombreux contacts entre les caractères appartenant à de lignes différentes.

3. L'extraction de métadonnées :

3.1.La métadonnée :

Une métadonnée est une donnée qui fournit une information sur une autre donnée. Le titre d'un mémoire peut être un exemple de métadonnée. Il informe sur le sujet traité dans le mémoire. Dans le domaine des métadonnées, on parle de données sur une ressource.

La métadonnée est utilisée pour présenter une donnée à l'utilisateur d'une manière significative. Elle est bien souvent utilisée pour indexer des objets numériques qu'elle décrit [Ton 09]. Elle devient un pont inévitable pour accéder à des informations contenu dans les ressources numériques sans forcément avoir recours directement au contenu de celles-ci [BLA].

Il existe trois types de métadonnées [NISO 04] :

a) Les métadonnées descriptives : elles décrivent une ressource à des fins telles que la découverte et l'identification.

   Exemples : titre d'un ouvrage, résumé, auteur, mots-clés ;

b) Les métadonnées structurelles : elles indiquent comment les objets composés sont mis ensemble.

   Exemple : l'ordre des pages pour former des chapitres ;

c) Les métadonnées administratives : elles fournissent des informations sur la gestion de la ressource.

   Exemples : type de fichier, date de création, type d'accès au fichier.

3.2.L'extraction de métadonnées :

L'extraction de métadonnées est le processus de traduction d'informations dans un formalisme clair pour la machine [Jou 09]. Ces informations sont issues de l'analyse d'une ressource. Dans notre cas la ressource est un document numérisé.

Le processus d'extraction de métadonnées à partir des images de manuscrits anciens numérisés utilise des outils d'analyse d'image pour alimenter une ressource textuelle facilement exploitable par la machine. Le principe d'extraction de métadonnées est souvent utilisé pour alimenter un formulaire. Dans notre cas les informations extraites seront exploitées pour alimenter diverses ressources telles que le catalogue ou l'index.

Un modèle d'extraction de métadonnées est mis en place en incorporant un ensemble de règles de recherche. On détermine ces règles à partir d'une analyse manuelle d'un échantillon de documents [MCT 09].

Il existe deux concepts de cette extraction [Jou 09]:

3.2.1. Extraction manuelle :

Elle concerne les informations convenues et systématiques qu'on retrouve dans les ouvrages (titre, auteur, année, …). Concrètement, cette extraction se fait via le remplissage de champs prédéfinis (Formulaire). Ce choix provient de l'existence des champs textuels (l'origine, l'auteur, la nature du document, . . .) et de l'absence d'informations extractibles directement dans l'image des documents.

L'indexation manuelle a deux problèmes [Jou 09]:

- Elle est dépendante de la langue ;
- Elle est longue à mettre en œuvre.

3.2.2. Extraction automatique :

Elle est utilisée pour les métadonnées basées uniquement sur les informations que l'on peut extraire des images de documents. Une version ASCII d'un texte manuscrit obtenue par un système de reconnaissance optique de caractères (OCR) est un exemple de métadonnées extraites automatiquement.

4. Conclusion :

Les manuscrits anciens constituent des ressources très recherchées pour divers aspects. Ainsi, la mise à disposition de ces ressources aux experts et érudits devient une nécessité absolue. Pour ce faire, nous avons proposé une approche basée sur l'extraction automatique des métadonnées par l'analyse d'images de manuscrits numérisés. Cette tâche est rendue délicate au vue de la nature des manuscrits qui peut présenter plusieurs anomalies qui entravent la phase d'analyse. Dans ce qui suivra, nous présentons l'apport du traitement d'image pour outrepasser ces inconvénients et proposer des solutions à mettre en œuvre. Dans le chapitre suivant nous citerons certaines méthodes du traitement d'image pour améliorer les images de documents, réduire leur dégradation et les préparer pour un traitement ultérieur.

Chapitre II : Traitements d’images de manuscrits anciens numérisés

1. Introduction :

Les dégradations que subissent les documents rendent les tâches de segmentation et de reconnaissance difficiles, voire quasi-impossible. Il existe des méthodes en traitement d'images destinées à réduire ces dégradations et améliorer les images.

Les méthodes et outils les plus courants en traitement d'images sont employées dans l'analyse de documents: suivi de contours, extraction de connexités, séparation texte/graphique, analyse de textures,… Elles permettent d'extraire les descripteurs du document. Un descripteur est un ensemble de caractéristiques des structures (illustrations, éléments graphiques et textuels, lignes de texte)constituant le document [Jou 09].

Dans ce chapitre nous étudierons les traitements basiques d'images. De la numérisation à la segmentation, en passant par les prétraitements, nous présenterons quelques méthodes utilisées dans les images de documents.

2. Numérisation de documents anciens :

La numérisation des documents anciens est devenue une nécessité absolue pour les services d'archives et les bibliothèques afin de répondre aux besoins toujours croissantsdes historiens et des chercheurs. Les possibilités de manipulation, de visualisation et de recherche d'information qui en découlent sont importantes.

La numérisation (dématérialisation ou digitalisation), consiste à créer une image du document (un tableau de points ou de *pixels*[3]) à l'aide d'une caméra numérique ou d'un scanneur. Une haute *résolution*[4] est souvent nécessaire pour restituer les éléments les plus fins de l'écriture et des graphismes. L'image obtenue est en couleur, en niveaux de gris (cf. page 13) ou binaire(cf. page 13) suivant les possibilités du capteur et les choix de numérisation. Le choix du format de stockage dépend de l'application visée et de la taille du support de conservation. Les possibilités de choix étant de conserver l'image brute ou de la comprimer avec ou sans perte d'information [Bot 00].

L'image du document obtenue n'est pas structurée et se présente comme un simple fichier que l'on peut visualiser.L'objectif est de faire appel aux méthodes de traitement d'images pour parvenir à associer aux données « image » des données textuelles (termes d'indexation, transcription, métadonnées) structurées ou non et sur lesquelles des recherches informatisées sont possibles.

[3]Contraction de l'anglais *pictureelement.*

[4] Nombre de pixels par unité de longueur. Elle s'exprime en PPP (point par pouce).

Les méthodes de traitement d'image permettent de restaurer ou nettoyer les images, de rechercher des informations directement dans les images et d'extraire les différentes structures du document (illustrations, éléments graphiques et textuels, lignes de texte) [Mou 06][Lik 03].

3. Représentation d'une image numérique :

L'image numérique est représentée par un tableau à deux dimensions, dont chaque case est appelée pixel. Chaque pixel est composé de trois couches RVB (Rouge, Vert, Bleu). Chaque valeur de pixel comporte une intensité de chacune des trois couches. L'intensité varie généralement dans l'intervalle [0 ; 255].

On peut définir trois classes d'images numériques :

- L'image en couleur (RVB) (voir Fig. 2.1);
- L'image en niveaux de gris (ou monochrome, les trois couleurs RVB sont de même intensité) (voir Fig. 2.2);
- L'image binaire (les trois couleurs RVB sont de même intensité et comportent deux valeurs : noir ou blanc).

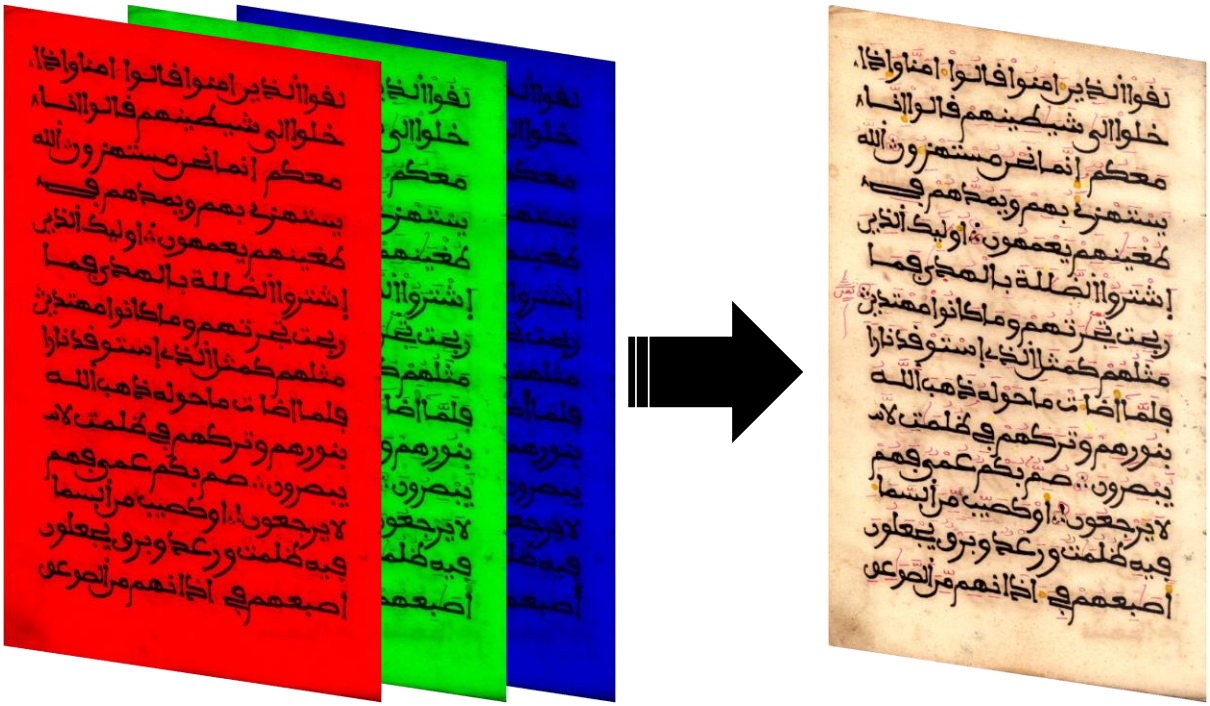


Fig. 2.1 : Représentation d'une image en couleurs RVB :

A gauche : les trois couches RVB ; à droite l'image résultante.

4. Les prétraitements :

Les prétraitements sont les premiers traitements sur l'image numérisée. Ils consistent à améliorer la qualité des images en éliminant les défauts (cf. page 6) dus à l'éclairage et au processus d'acquisition. Ils peuvent être effectués en vue d'une utilisation ultérieure de structuration ou de reconnaissance [Lik 03]. Ils sont utilisés pour :

- L'amélioration : atténuation ou suppression d'une dégradation présente sur l'image ;
- La simplification : suppression de l'information inutile pour l'analyse ultérieure de l'image.

Les principaux prétraitements qu'on applique sur les images de documents sont :

- Les modifications d'histogramme ;
- Les filtrages ;
- Les opérations morpho mathématiques ;

Le tableau suivant présente les prétraitements usuels pour certains types de problèmes rencontrés :

Défaut	Prétraitement
Luminosité trop sombre ou trop claire	Modification de l'histogramme [BEL 92]
Taches	Filtrage passe-haut [FEL 00]
Points parasites	Filtrage passe-bas [BEL 92] Filtrages morphologiques [MEN 00]
Rotation légère de l'image	Calcul de l'angle par projection [BEL 92] Redressement par ré-échantillonnage [DEB 00]
Courbure de l'écriture sur un bord de l'image	Calcul de la courbure locale, Ré-échantillonnage [DEB 00]
Ecriture fragmentée	Filtrages (passe-haut, morphologiques, passe-bas) [FEL 00][DEB 00]
Contours de l'écriture flous	Filtrage passe-haut, Filtrages morphologiques [LAM 96]
Ecriture du verso apparaissant sur le recto	Combinaison des images recto et verso [LAM 96][LIN 94][TAN 02]

Tab. 2.1 : Prétraitements courants sur les images de documents anciens [Lik 03].

4.1.Modification d'histogramme :

Cette opération consiste à affecter de nouvelles valeurs aux pixels de l'image suivant unetransformation linéaire ou non des valeurs de l'image d'origine [Lik 03].

Cette modification n'altère pas les informations contenues dans l'image mais les rend plus ou moins visibles.

Parmi plusieurs types de modification d'histogramme nous citons :

- L'étalement d'histogramme ;
- L'égalisation d'histogramme.

4.1.1. L'histogramme d'une image :

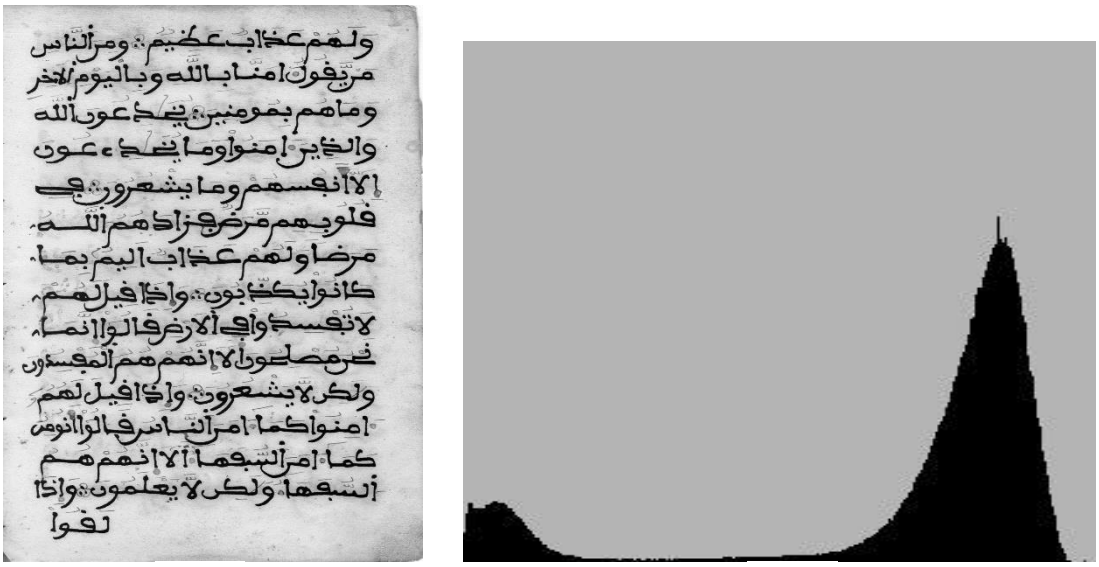
L'histogramme d'une image (voir Fig. 2.2b) est une fonction qui associe à chaque valeur d'intensité le nombre de pixels dans l'image ayant cette valeur.

L'histogramme ne contient aucune information relative à l'emplacement des pixels ni sur la proximité relative de deux pixels. Par contre, l'information qu'il contient peut concerner notamment la brillance apparente et le contraste d'une image, et il est utilisé en traitement d'images pour manipuler les caractéristiques de l'image.

L'image peut avoir un seul histogramme si elle est en niveaux de gris. Elle peut avoir trois histogrammes si elle est en couleur (chacune des couches RVB a son propre histogramme).

On peut exploiter l'histogramme d'une image pour :

- Extraire un seuil pour binariser l'image ;
- Rehausser ou améliorer le *contraste*[5] de l'image.



(a) (b)

Fig. 2.2 : Exemple d'un histogramme d'une image :

(a) image en niveaux de gris ; (b) histogramme de l'image.

[5] Propriété intrinsèque d'une image qui quantifie la différence de luminosité entre les parties claires et sombres d'une image.

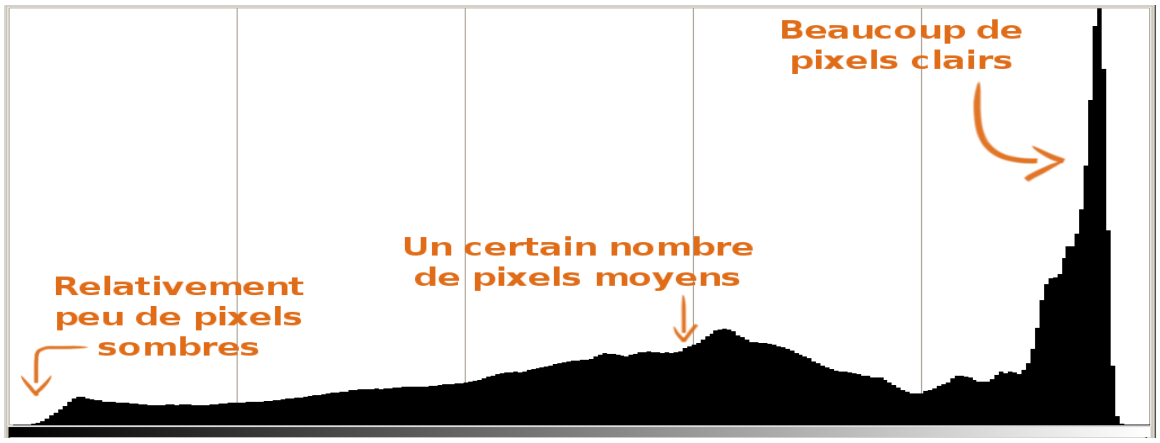


Fig. 2.3 : Interprétation d'un histogramme.

4.1.2. L'étalement d'histogramme :

C'est la forme la plus simple de modification d'histogramme. Encore appelé expansion ou linéarisation, elle consiste à effectuer une dilatation artificielle mais linéaire de l'échelle de gris de l'image.Elle étale l'histogramme d'une image contenu dans l'intervalle [a ; b] sur tous les niveaux de gris disponibles (généralement [0 ; 255]) (voir. Fig. 2.4). L'objectif étant d'augmenter le contraste de l'image.

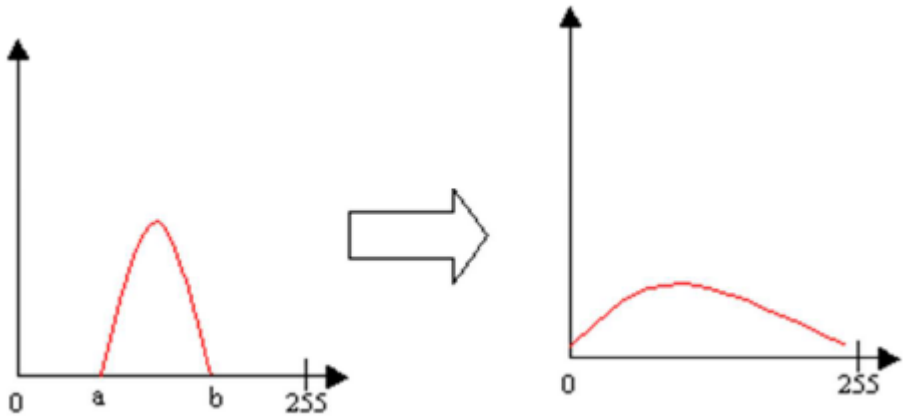


Fig. 2.4 : étalement d'histogramme.

On effectue l'étalement d'histogramme avec la transformation linéaire suivante :

$$I'(x,y) = arrondit(M\frac{I(x,y) - I_{min}}{I_{max} - I_{min}})$$

Où :

$I'(x, y)$: Niveau de gris au point (x, y) de l'image obtenue ;

$I(x, y)$: Niveau de gris au point (x, y) de l'image d'origine ;

I_{min} : Plus petite valeur de niveaux de gris dans l'image d'origine ;

I_{max} : Plus grande valeur de niveaux de gris dans l'image d'origine ;

M : Amplitude d'arrivée (en général M=255).

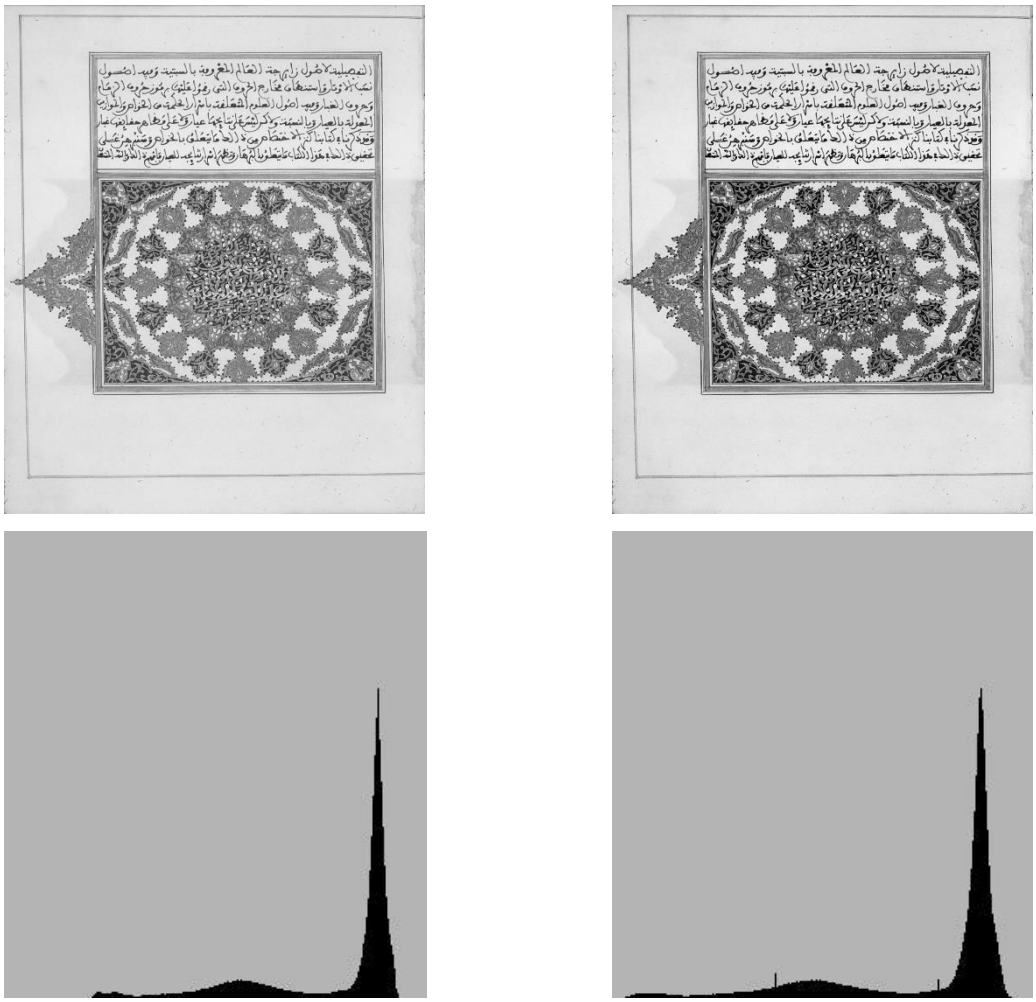


Fig. 2.5 : étalement d'histogramme :

A gauche : image en niveaux de gris [IMMb 04] avec en dessous son histogramme ;

A droite : image obtenue avec l'étalement d'histogramme.

4.1.3. Egalisation d'histogramme :

L'égalisation d'histogramme est une méthode d'ajustement du contraste d'une image numérique. Elle impose sur l'image une répartition homogène des niveaux de gris. Elle force tous les niveaux de gris à être équiprobables.

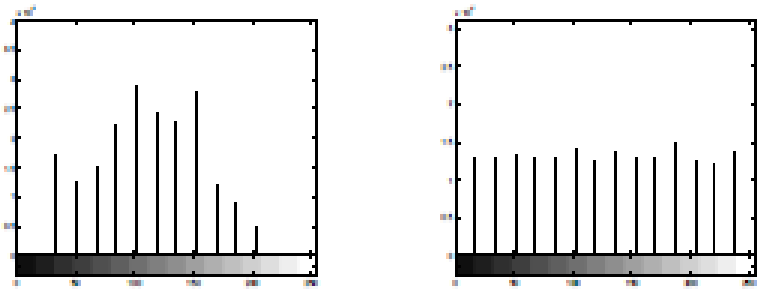


Fig. 2.6 : Egalisation d'histogramme [Ort 04].

Elle consiste à appliquer une transformation sur chaque pixel de l'image, et ainsi obtenir une nouvelle image à partir d'une opération indépendante sur chacun des pixels. Cette transformation est construite à partir de l'histogramme cumulé de l'image de départ.

Cette opération vise à augmenter les nuances dans l'image. Elle est intéressante pour les images dont la totalité, ou seulement une partie, est de faible contraste.

L'algorithme d'égalisation est le suivant :

a) Calculer l'histogramme $H(i)$ de l'image (i est la valeur du niveau de gris);

b) Calculer l'histogramme cumulé $H_C(i)$ de l'image :

$$H_C(i) = \sum_{k=0}^{i} H(k)$$

c) Transformation des niveaux de gris de l'image :

$$I'(x,y) = \frac{H_C\big(I(x,y)\big) * M}{N}$$

Où :

$I'(x,y)$: Niveau de gris au point (x, y) de l'image obtenue ;

$I(x,y)$: Niveau de gris au point (x, y) de l'image d'origine ;

M : Amplitude d'arrivée (en général M=255) ;

N : Nombre total de pixels dans l'image d'origine.

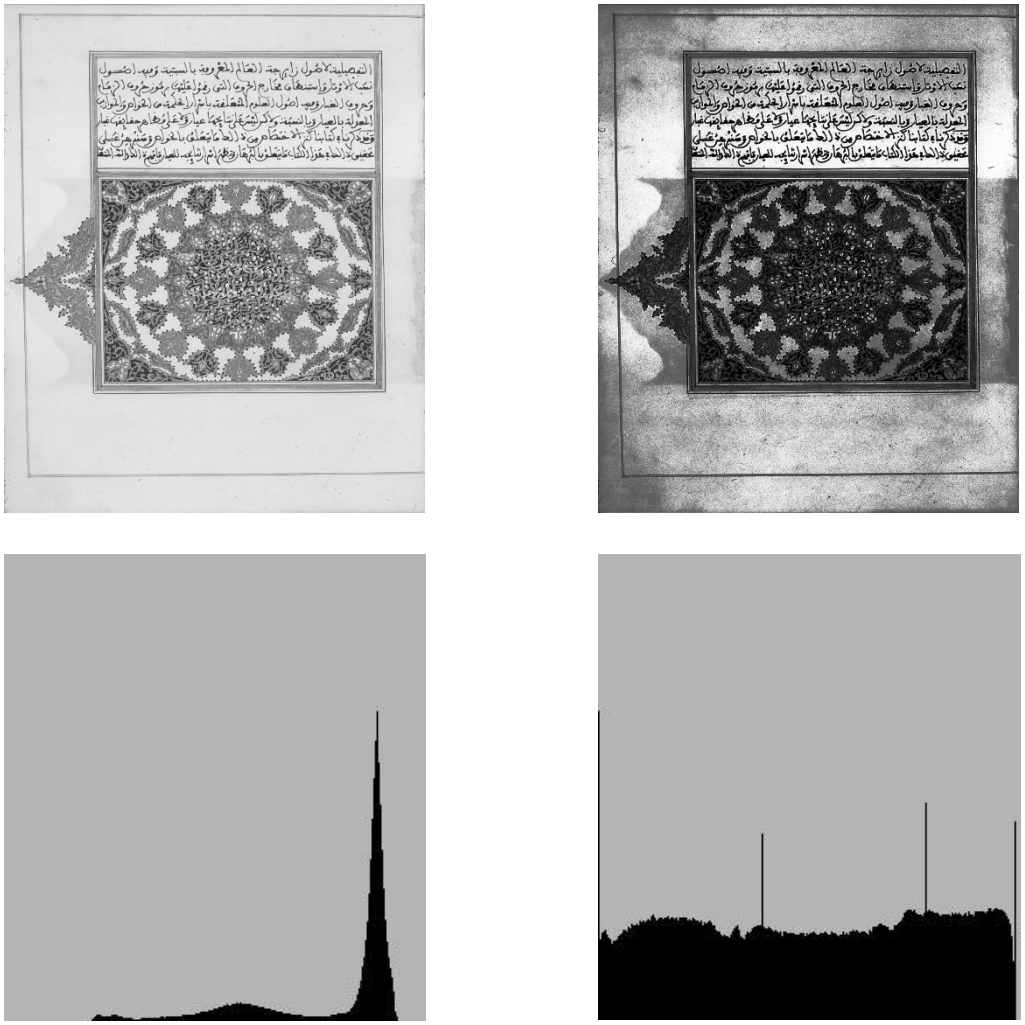


Fig. 2.7 : égalisation d'histogramme :

A gauche : image en niveaux de gris [IMMb 04] avec en dessous son histogramme ;

A droite : image obtenue avec l'égalisation d'histogramme.

Remarque : la modification d'histogramme est souvent appelée *filtre spectral*. Le spectre d'une image étant la courbe de répartition des couleurs d'une image, communément appelé histogramme d'une image [Jan 11].

4.2.Les filtrages :

Le filtrage est l'opération consistant à remplacer la valeur de chaque pixel de l'image par une valeur dépendant de celle des pixels appartenant à son voisinage [Lik 03].Dans la plupart des cas, Il consiste à balayer l'image par une fenêtre d'analyse de taille finie (une matrice de taille impaire). Le calcul du nouveau niveau de gris du pixel considéré ne prend en compte que

les plus proches voisins de celui-ci. Le filtrage sert à retrouver le maximum d'informations sous une image bruitée [Ort 04].

*Remarque :*Dans la suite de notre travail, nous nous baserons sur le voisinage 8-connexité, c'est-à-dire les 8 voisins qui entourent le pixel.

On peut catégoriser trois types de filtrages [Ort 04] :

- Filtrages linéaires ;
- Filtrages linéaires itérés ;
- Filtrages non linéaires.

4.2.1. Filtrages linéaires :

Un filtre F est dit linéaire s'il respecte les deux propriétés suivantes :

- ✓ $F(A+B) = F(A) + F(B)$ (Principe de superposition);
- ✓ $F(k.A) = k.F(A)$ (Principe d'homogénéité) où k est une constante.

Le filtrage linéaire repose sur l'utilisation de masques de convolution. Dans le domaine de traitement d'images, on parle de corrélation.

Soient I(x,y) une image de coordonnées (x,y) et H(x,y) le masque de convolution du filtre (appelé aussi *noyau* de convolution) de dimension (d_1, d_2). La convolution F de I et H est définie par l'équation suivante :

$$F(x,y) = (I * H)(x,y) = \sum_{i=-(\frac{d_1-1}{2})}^{\frac{d_1-1}{2}} \sum_{j=-(\frac{d_2-1}{2})}^{\frac{d_2-1}{2}} I(x-i, y-j).H(i,j)$$

Le résultat de la convolution est divisé par la somme des coefficients du masque.

Les filtres linéaires usuels dans le traitement d'images de documents anciens sont :

- Le filtre passe-bas : il diminue le bruit mais atténue les détails de l'image (augmentation du flou). Parmi les filtres passe-bas : le filtre moyenneur et le filtre gaussien.
- Le filtre passe-haut : utilisé pour la détection de contours et pour augmenter la netteté et le contraste. Le filtre Laplacien et le filtre de Sobel sont des exemples de filtres passe-haut.

a) Le filtre moyenneur :

Le filtre moyenneur permet de lisser l'image. Ce lissage rend l'image floue. Chaque pixel est remplacé par la valeur moyenne de ses voisins. La taille du noyau dépend de l'intensité du bruit et de la taille des détails significatifs de l'image traitée. En conséquence, plus les

dimensions du noyau seront importantes, plus le bruit sera éliminé ; mais en contrepartie, les détails fins seront eux-aussi effacés et les contours étalés.

$\frac{1}{9}$

1	1	1
1	1	1
1	1	1

(a)

$\frac{1}{25}$

1	1	1	1	1
1	1	1	1	1
1	1	1	1	1
1	1	1	1	1
1	1	1	1	1

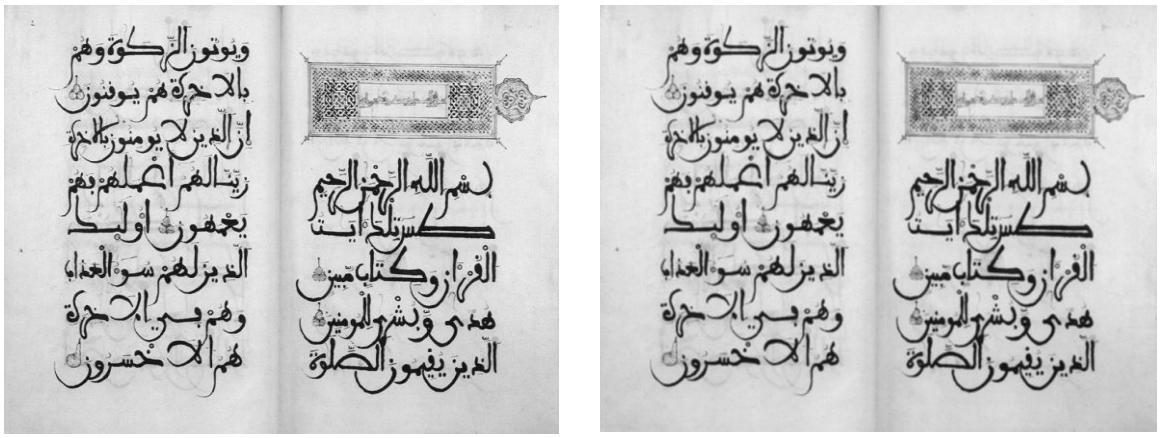
(b)

$\frac{1}{49}$

1	1	1	1	1	1	1
1	1	1	1	1	1	1
1	1	1	1	1	1	1
1	1	1	1	1	1	1
1	1	1	1	1	1	1
1	1	1	1	1	1	1
1	1	1	1	1	1	1

(c)

Tab. 2.2 : Exemple de filtres moyenneurs : (a) 3x3 ; (b) 5x5 ; (c) 7x7.



(a) (b)

(c) (d)

Fig. 2.8 : Filtrage moyen : (a) Image originale ;

(b) filtrage moyen 3x3 ; (c) filtrage moyen 5x5 ; (d) filtrage moyen 7x7.

b) Le filtre Gaussien :

Le filtre gaussien est aussi un filtre de lissage comme le filtre moyenneur. Cependant il donne un meilleur lissage, une meilleure réduction de bruit, les contours et les détails fins sont mieux conservés qu'avec le filtre moyenneur.

Un filtre gaussien est donné par la discrétisation de la fonction gaussienne [Ber 15] :

$$G(x,y) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}}$$, sur un voisinage de (0,0).

σest l'écart type.

Si par exemple $\sigma = 0{,}8$, on a le filtre 3x3 suivant :

G (-1,-1)	G(0,-1)	G(1,-1)
G (-1,0)	G(0,0)	G(1,0)
G (-1,1)	G(0,1)	G(1,1)

$\approx \frac{1}{16}$

1	2	1
2	4	2
1	2	1

Si $\sigma = 1$, on a le filtre 5x5 suivant :

$\frac{1}{273}$

1	4	7	4	1
4	16	26	16	4
7	26	41	26	7
4	16	26	16	4
1	4	7	4	1

La taille du filtre gaussien est gouvernée par σ. En général un filtre gaussien avec σ < 1 est utilisé pour réduire le bruit. Plus σ est grand, plus le flou appliqué à l'image sera important[Ber 15] . Idéalement, on devrait prévoir un filtre de taille (6σ +1)×(6σ +1) [Ber 10].

(a) (b)

Fig. 2.9 : filtrage gaussien :

(a) image originale ; (b) filtrage gaussien 5x5 avec $\sigma = 1$.

c) Le filtre Laplacien :

Le filtre Laplacien permet la détection de contours. Un contour est définit comme une discontinuité locale de l'intensité lumineuse.

On peut trouver différents opérateurs Laplaciens [Ber 10] :

0	1	0
1	-4	1
0	1	0

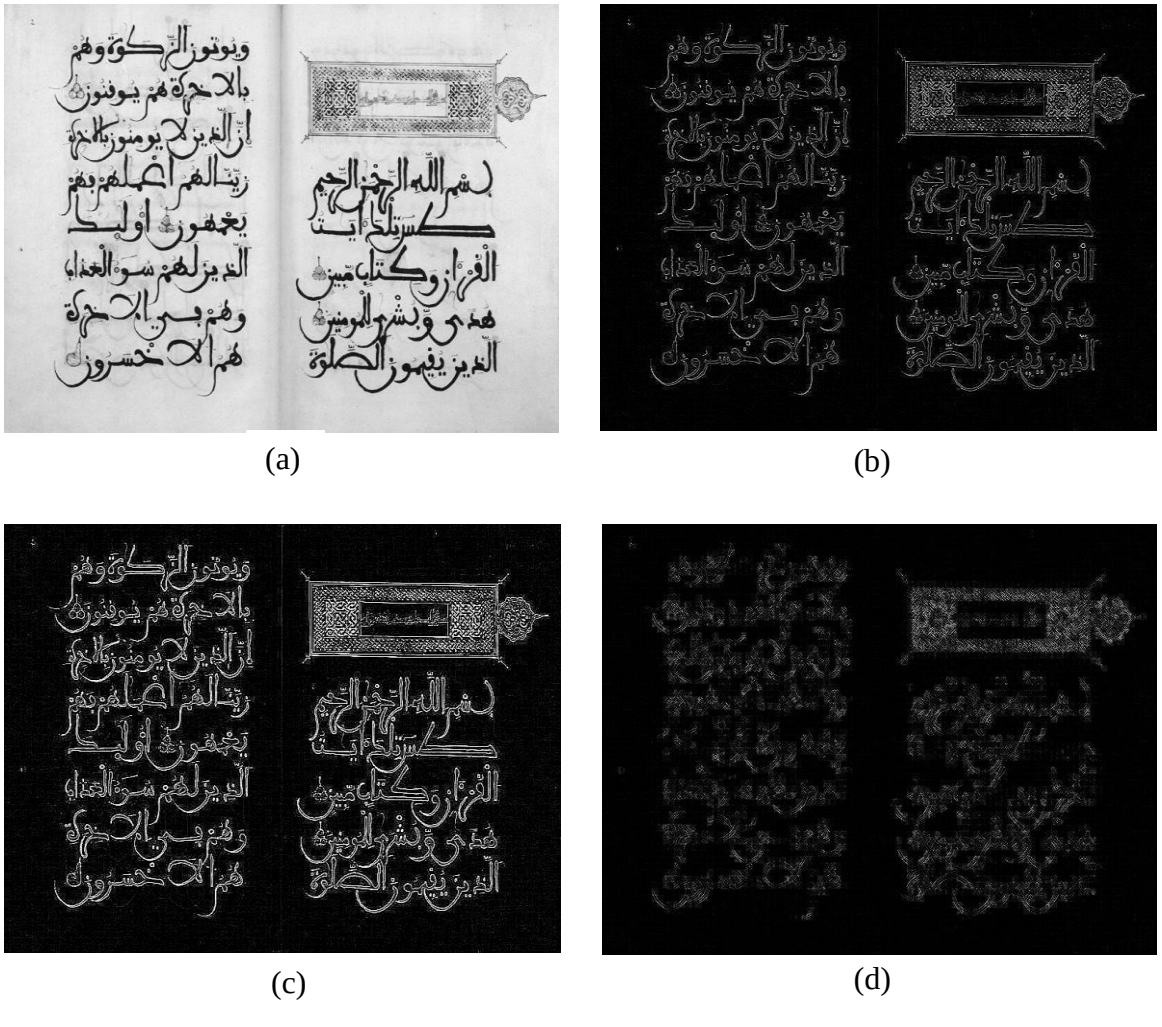
Laplacien 4-connexe

1	1	1
1	-8	1
1	1	1

Laplacien 8-connexe

1	-2	1
-2	4	-2
1	-2	1

Laplacien de Robinson



(a) (b)

(c) (d)

Fig. 2.10 : Filtrage Laplacien : (a) Image originale ;

(b) Laplacien 4-connexe ; (c) Laplacien 8-connexe ; (d) Laplacien de Robinson.

d) Le filtre de Sobel :

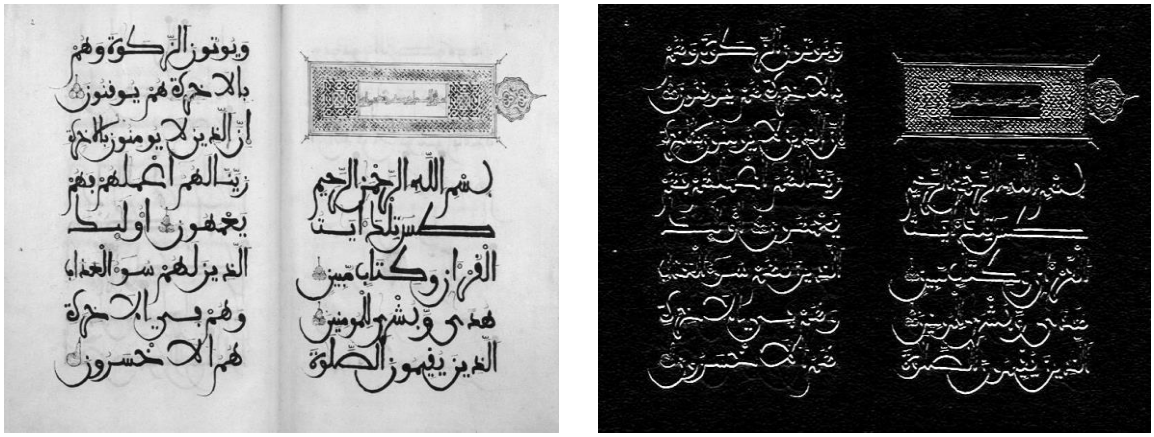
Le filtre de Sobel est un autre détecteur de contours. Il comporte un masque horizontal et un autre vertical :

1	0	-1
2	0	-2
1	0	-1

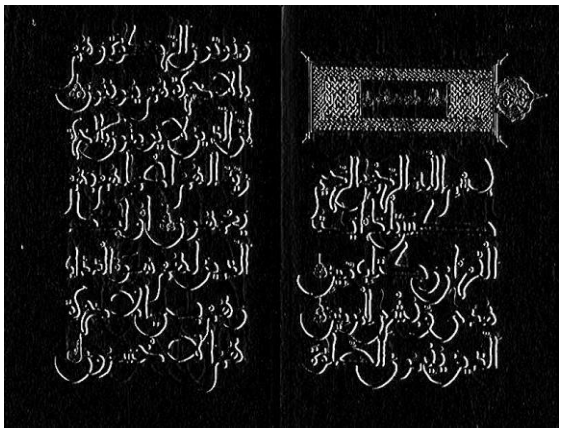
Filtre horizontal

1	2	1
0	0	0
-1	-2	-1

Filtre vertical



(a) (b)



(c)

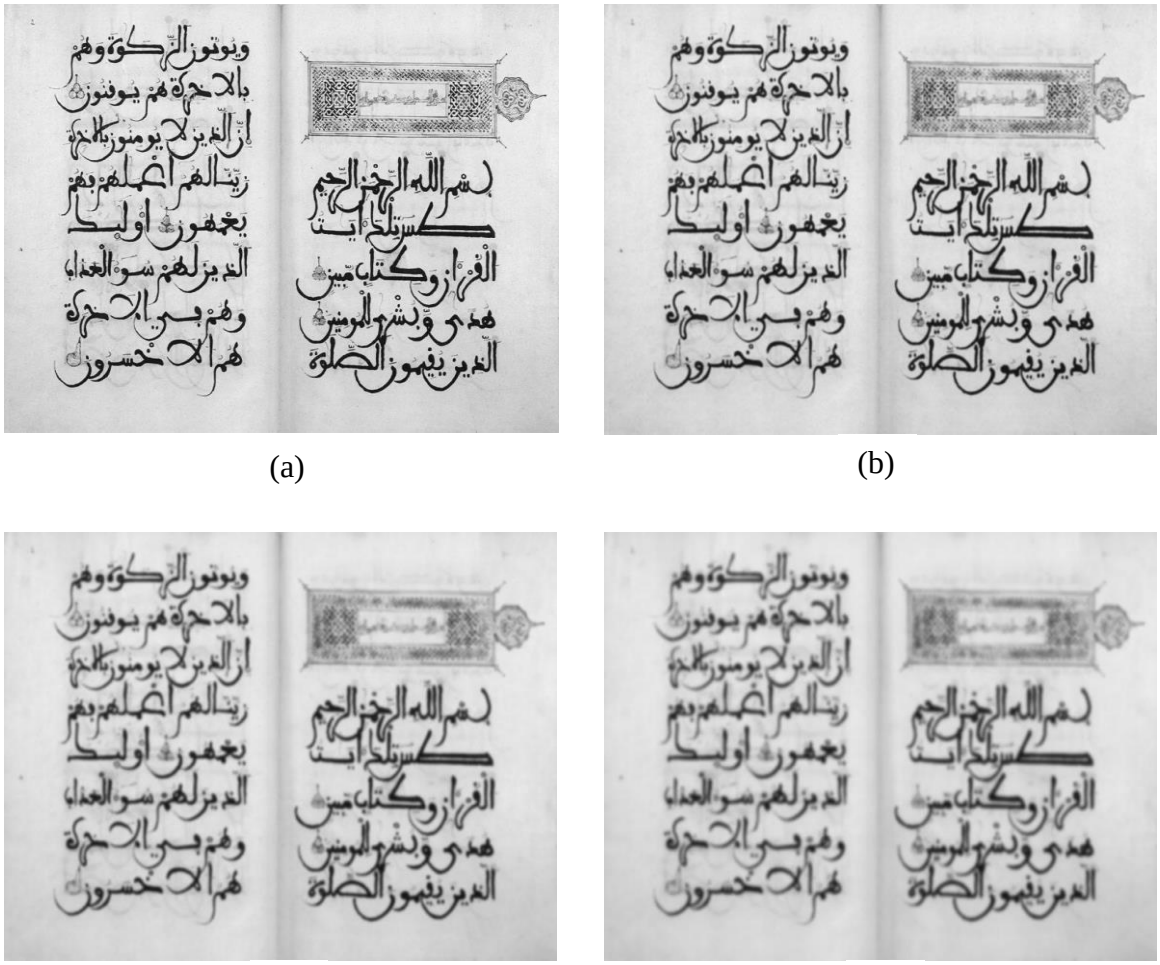
Fig. 2.11 : Filtrage de Sobel : (a) Image originale ;

(b) filtrage vertical ; (c) filtrage horizontal.

4.2.2. Filtrages linéaires itérés :

Il s'agit du filtrage linéaire amélioré soit :

- En cherchant les meilleurs valeurs du masque ;
- En augmentant la taille du masque ;
- En appliquant plusieurs fois de suite le même masque (voir Fig 2.12).



(a) (b)

(c) (d)

Fig. 2.12 : Application du filtre gaussien 5x5 avec $\sigma = 1$:

(a) Image originale ; (b) 1 itération ; (c) 3 itérations ; (d) 5 itérations.

4.2.3 Filtrages non linéaires :

Le filtre non linéaire est un filtre qui ne peut pas s'implémenter comme un produit de convolution car il ne respecte pas les deux propriétés de linéarité citées ci-dessus.

Les filtres non linéaires usuels sont le filtre médian, le filtre MIN et le filtre MAX.

a) Le filtre médian :

Le filtre médian ne s'implémente pas comme un produit de convolution. Il est utilisé pour nettoyer les bruits dans l'image. Il remplace la valeur du pixel par la valeur médiane dans son voisinage. Ce filtre est plus performant que le filtre moyenneur ou le filtre gaussien, surtout pour éliminer l'effet « poivre et sel », c'est-à-dire les faux points noirs et points blancs dans l'image.

Par exemple, si on prend une image I(x, y) :

17	85	50
22	255	48
30	25	25

Alors le filtre

Fmed(x, y)= médiane{I(x, y)}=médiane{17,22,25,25,**30**,48,50,85,255}=30.

b) Le filtre MIN :

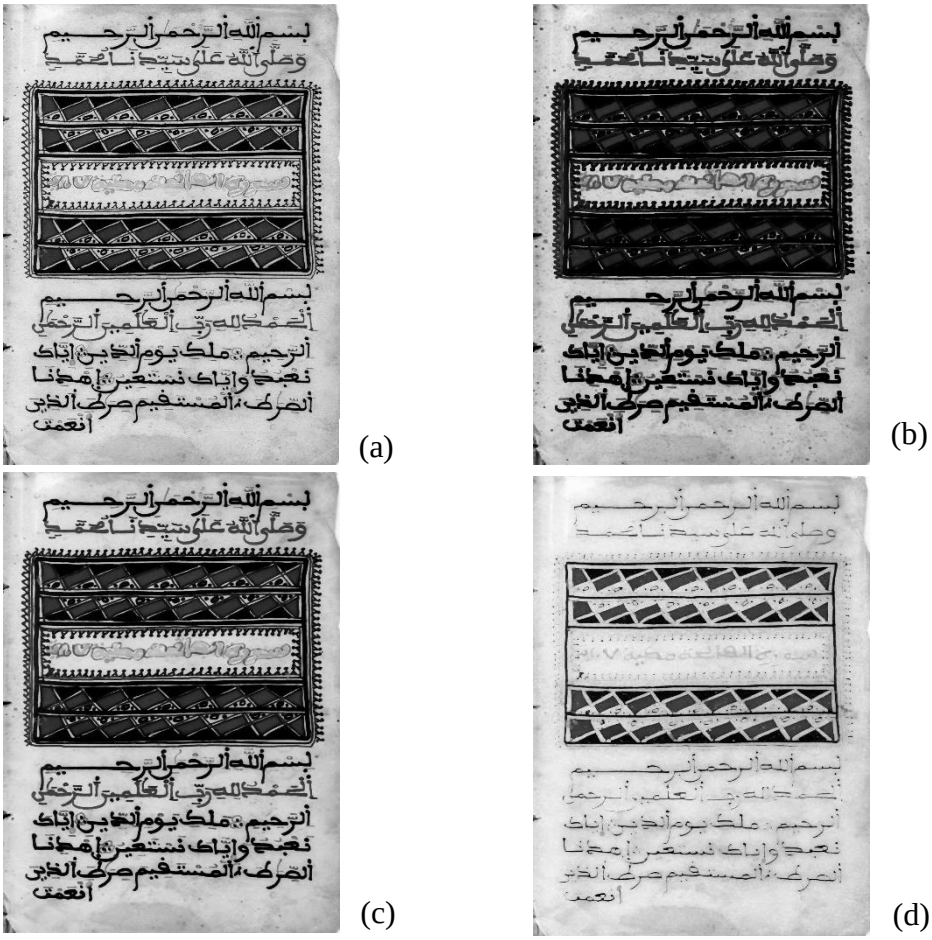
Le filtre MIN remplace la valeur du pixel par la valeur minimale dans son voisinage. Il permet de dilater les pixels sombres de l'image.

Si on prend l'exemple précédent, Fmin=17.

c) Le filtre MAX :

Le filtre MAX remplace la valeur du pixel par la valeur maximale dans son voisinage. Il permet de dilater les pixels clairs de l'image.

Si on prend l'exemple précédent, Fmax=255.



(a) (b) (c) (d)

Fig. 2.13 : Filtrages non linéaires :

(a) Image originale ; (b) application d'un filtre MIN 5x5 ; (c) application d'un filtre médian 5x5; (d) application d'un filtre MAX 5x5.

4.3.Les opérateurs morpho mathématiques (ou morphologiques):

La morphologie mathématique est une théorie et technique mathématique et informatique d'analyse de structures.Elle est généralement appliquée sur des images binaires. Son but est de détecter et remplir les trous, lisser les bords, segmenter l'image et mettre en évidence des caractéristiques de l'image[FPW 04]. Elle repose sur l'utilisation d'un élément structurant [Ort 04].Un élément structurant est un masque binaire (constitué de pixels blancs et noirs) muni d'un point d'ancrage.Le point d'ancrage est appelé origine et il est situé dans son centre.

0	1	0
1	**1**	1
0	1	0

(a)

1	1	1
1	**1**	1
1	1	1

(b)

Tab. 2.3 : Exemples d'éléments structurants :

(a) Elément 4-connexité ; (b) élément 8-connexité.

Le principe des opérateurs morphologiques est le suivant :

- Balayer l'image avec un élément structurant ;
- Décider de l'opération à effectuer sur chaque pixel en le comparant à l'élément structurant.

Les deux principales opérations morphologiques sont :

- La dilatation, et
- l'érosion.

On peut trouver d'autres opérations comme :

- L'ouverture : c'est une érosion suivie d'une dilatation. Elle élimine les éléments fins et modifie les contours;
- La fermeture : c'est une dilatation suivie d'une érosion. elle permet de remplir les petits trous et de lisser les contours;
- La squelettisation : c'est la représentation la plus fine possible des objets.

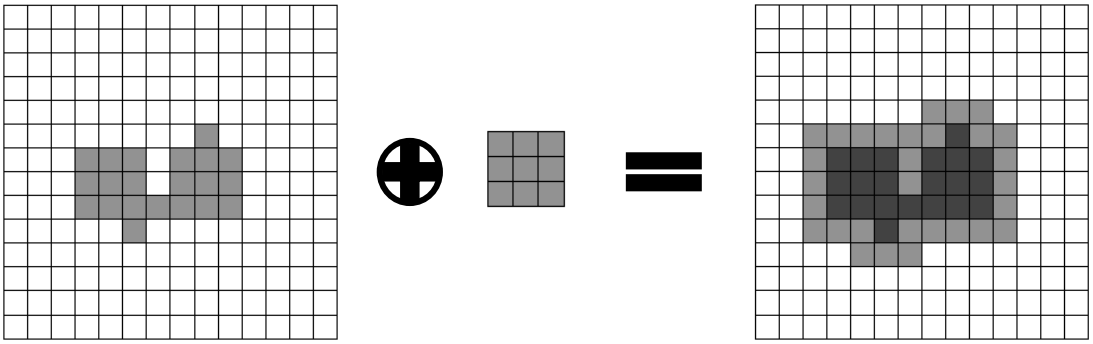
L'ouverture et la fermeture ne sont pas inverses car la dilatation et l'érosion ne sont pas inverses.

4.3.1. La dilatation :

Comme son nom l'indique, la dilatation consiste à élargir les frontières des objets. La taille des objets augmente et les trous présents entre eux diminuent[FPW 04].

Soit un objet A et l'élément structurant B. La dilatation de A par B est définie comme suit :

Si l'origine de B touche en partie un élément de A alors implanter B dans ce lieu.



(A) (B) (C)

Fig. 2.14 : Dilatation de l'image :

(A) image originale ; (B) élément structurant ; (C) image dilatée ;

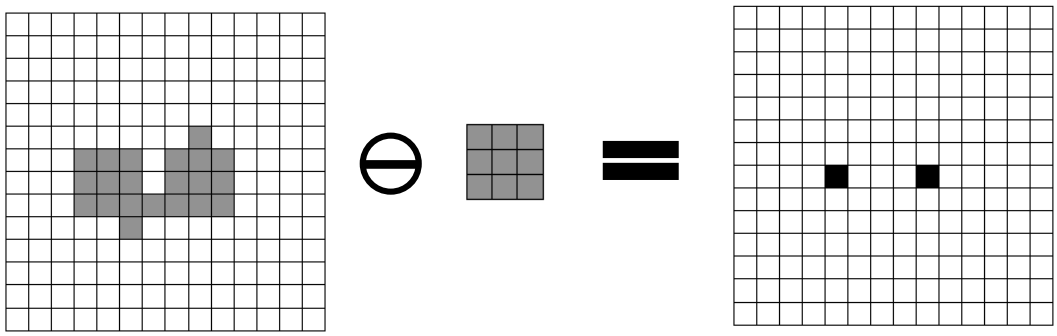
$\oplus$: Opérateur de dilatation.

4.3.2. L'érosion :

Elle consiste à rétrécir la taille des objets. Les objets avec concavités ou trous peuvent être divisés et les petits détails disparaissent.

Soit l'objet A et l'élément structurant B. L'érosion de A par B est définie comme suit :

Si B est complétement inclus dans A alors conserver le pixel de A coïncidant avec l'origine de B. les autres pixels sont effacés.



(A) (B) (C)

Fig. 2.15 : Erosion de l'image :

(A) image originale ; (B) élément structurant ; (C) image érodée ;

$\ominus$: Opérateur d'érosion.

Remarques :

- Les opérateurs morpho mathématiques sont des filtres non linéaires ;

- Appliquer la dilatation (respectivement L'érosion) sur une image en niveaux de gris revient à appliquer le filtre MAX (respectivement le filtre MIN).

5. La segmentation :

La segmentation de l'image est le cœur de tout système d'analyse de document et éventuellement de vision artificielle. C'est une étape importante dans le processus d'analyse d'image.

L'objectif de la segmentation est de fournir une description des objets contenus dans l'image à travers l'extraction de diverses indications visuelles telles que les contours des objets ou les régions homogènes.

La littérature avance qu'il n'y a pas de méthode universelle pour la segmentation d'images. Toute technique est efficace uniquement pour un type d'image donné, pour un type d'application donné, et dans un contexte donné [ISPS 11].

Il existe différentes techniques de segmentation [Lel 07], nous citons entre autre :

- La binarisation ;
- La segmentation ascendante ;
- La segmentation descendante ;
- La segmentation par texture.

5.1.La binarisation :

Binariser une image revient à segmenter l'image en deux classes : le fond et l'objet [Lel 07]. La binarisation transforme une image (couleurs ou niveaux de gris) en une image noir et blanc.

La binarisation est une étape très importante dans tout processus de traitement et d'analyse d'image. Elle a comme but de diminuer la quantité d'informations présentes dans l'image, et de ne garder que les informations pertinentes.

Il existe deux types de binarisation :

- Binarisation locale : Dans ce type de méthode les valeurs des seuils sont déterminées localement, pixel par pixel ou bien région par région. Le seuil est calculé automatiquement pour chaque pixel de l'image. Les méthodes de Sauvola ou de Niblack sont utilisées pour la binarisation locale.
- Binarisation globale : elle consiste à calculer un seuil pour toute l'image. Elle subdivise l'histogramme de l'image en deux classes C_0 et C_1. C_0 représente la classe des pixels noirs. C_1 représente la classe des pixels blancs.

Le seuil de binarisation T permettra de déterminer les classes C0 et C1 en respectant l'algorithme suivant :

$$I(x, y) = \begin{cases} 0 & \text{Si } I(x, y) <= T \quad \text{(Noir)} \\ 255 & \text{Sinon} \quad \text{(blanc)} \end{cases}$$

Où I(x, y) un pixel d'une image I de coordonnées (x, y).

La binarisation globale peut être manuelle ou automatique. La méthode manuelle présente un inconvénient car le seuil optimal dépend d'une image à une autre. La méthode automatique est meilleure mais elle peut échouer si l'image est dégradée (mauvais éclairage, bruits…).

5.1.1. Mise en niveaux de gris :

Généralement la binarisation s'opère sur des images en niveaux de gris. Pour passer d'une image couleurs à une image en niveaux de gris on peut appliquer ces formules :

(a) Formule standard :

$$\text{Gris} = \frac{\text{Rouge} + \text{Vert} + \text{Bleu}}{3}$$

(b) formule proposée par la C.I.E. (Commission Internationale de l'éclairage) dans sa recommandation 709, qui concerne les couleurs « vraies » ou naturelles :

$$\text{Gris} = 0.2125 * \text{rouge} + 0{,}7154 * \text{Vert} + 0{,}0721 * \text{Bleu}$$

Le tableau suivant regroupe les différentes méthodes de binarisations avec leurs limites :

Nom	Année	Type	Principe	Inconvénients
Otsu [Ots 79]	1979	Seuillage global	Cherche à minimise la variance intra-classe à partir de l'histogramme normalisé.	Problèmes pour les documents mal éclairés.
Wu [WM 98]	1998	Seuillage global	« Floute » l'image pour mieux séparer l'histogramme et utilise une méthode de seuillage global [Ots 79].	Problèmes lorsqu'il n'y a pas deux modes distincts sur l'histogramme.
Bernsen [Ber 86]	1986	Seuillage Local	Estime la valeur du seuil en faisant la moyenne de la plus haute et la plus basse valeur de la fenêtre.	Le seuil est trop bas lorsque la fenêtre est centrée sur du fond.
Niblack [Nib 86]	1986	Seuillage Local	Amélioration de [Ber 86] : prise en compte de la variance et de la moyenne.	Même problème que[Ber 86] : apparition de bruit sur les zones uniformes.
Sauvola [SP 00]	2000	Seuillage Local	Insère des constantes dans la méthode de [Nib86] afin d'améliorer la méthode sur les zones uniformes.	Les constantes à ajuster empêchent la méthode de traiter parfaitement des documents non uniformes.

Kim [KJP02]	2002	Seuillage Local	Utilise la valeur des pixels comme courbes de niveaux pour simuler une montée des eaux.	Difficultés à adapter le débit de l'eau ou le nombre d'itérations au document.
Wolf [WD02]	2002	Seuillage Local	Utilise les champs de Markov pour savoir où se trouvent les caractères.	L'utilisation de [SP00] rend la technique victime des mêmes limitations que pour Sauvola.
Garain [GPH05]	2005	Seuillage Local	Utilise les composantes connexes pour créer un graphe d'adjacence qui est ensuite réduit.	Marche mal si l'image à traiter contient des illustrations.
Gatos [GPP06]	2006	Seuillage Local	Cherche à estimer le fond pour ensuite faire un seuillage sur la différence entre le fond et l'image d'origine.	Très bonnes performances.

Tab. 2.4 : Différentes techniques de binarisation[Lel 07].

5.2.La segmentation ascendante :

Cette catégorie est caractérisée par le fait que l'analyse part des éléments de bas niveau (comme les pixels) pour essayer de les fusionner [Lel 07]. Deux des méthodes les plus connues de cette catégorie sont le RLSA (RunLengthSmoothingAlgorithm) et la transformée de Hough.

Nom	Année	Principe	Inconvénients
Wong [WCW82]	1982	Noirci les espaces blancs entre deux pixels noirs verticalement et horizontalement puis fait un « ET logique » entre les deux images noircies (RLSA).	Nécessite une orientation horizontale du texte.
Antonacopoulos [Ant98]	1998	Utilise des tuiles sur le fond pour estimer les interlignes.	Très sensible au bruit de fond.
Lienhart [LE00]	2000	Accroissement de régions où les frontières se déplacent en fonction du gradient.	plus adapté à la segmentation de vidéos.
Wang [WPH06]	2006	Utilisation d'un vecteur de 69 caractéristiques, réduction à 23 par un algorithme de classification pour ensuite identifier le type de la boîte.	L'intérêt de la méthode n'est pas de segmenter mais de classifier.
Nicolas [NPH06]	2006	Utilisation des champs de Markov pour caractériser le texte.	L'apprentissage ne rend la méthode valable que pour un type de document à la fois.
Caponetti [CCG07]	2007	Utilise deux réseaux de neurones flous pour segmenter une image.	La phase d'apprentissage est très lourde à mettre en place en raison du type de réseau.

Tab. 2.5 : Quelques approches ascendantes [Lel 07].

5.3.La segmentation descendante :

La famille de techniques de segmentation descendante essaie d'avoir une approche globale pour affiner les régions. Ces méthodes sont apparues après les méthodes descendantes. Elles n'ont pas connu un grand développement malgré les très bonnes performances obtenues[Lel 07]. Dans cette méthode nous trouvons la technique de projection.

Nom	Année	Principe	Inconvénients
Horowitz [HP72]	1972	Commence par découper l'image en quatre, récursivement puis fusionne les zones de caractéristiques proches.	Nécessite une organisation horizontale de l'image.
Nagy [NS84]	1984	Découpe l'image horizontalement puis verticalement, récursivement. Le découpage se fait dans le creux des projections.	Il existe des documents impossibles à segmenter.
Kim [Kim96]	1996	Utilise les pics de l'histogramme pour sélectionner les pixels de couleurs proches. Regroupe ensuite les composants proches et utilise des heuristiques pour classer les composantes.	L'utilisation de l'histogramme est trop générale pour donner de bons résultats.
Kim [Kim96]	1996	Réduit le nombre de couleurs en fonction des couleurs les plus proches. Regroupe ensuite les composantes de couleurs proches et utilise des heuristiques pour classer les composantes.	Pas de prise en compte spatiale des couleurs.

Tab. 2.6 : Quelques approches descendantes [Lel 07].

5.4.La segmentation par texture :

L'approche par segmentation sur la texture regroupe beaucoup de techniques différentes. Le but de ces approches est de trouver les caractéristiques de texture qui sont propres au texte. De nombreux filtres sont alors utilisés pour transformer l'image en une représentation mettant en avant ces caractéristiques [Lel 07]. Une des approches de cette catégorie est le filtrage passe-haut (cf. page 20).

Nom	Principe	Avantages	Inconvénients
Fourier [ZZJ00]	Passe l'image dans le domaine fréquentiel.	Rapide, utilisé dans les images JPEG.	Perte de la localisation. Nécessite une fenêtre d'analyse.
Dérivée [WMR99] [SKHS98]	Met en avant les variationsdans l'image.	Rapide, permet de localiser les contours.	Nécessite une taille adaptée de la fenêtre d'analyse.
Autocorrélation [Jou06]	Met en avant l'orientation générale et la périodicité de la texture.	Permet de créer une rose des directions.	Très coûteux en calculs. La taille de la fenêtre d'analyse est un paramètre critique.

Tab. 2.7 : Quelques méthodes pour l'approche texturelle[Lel 07].

6. Conclusion :

Ce chapitre nous a fait initier au domaine du traitement d'images. Nous avons vu l'utilité des prétraitements dans l'amélioration et la simplification de l'image, et l'importance de la segmentation pour mettre en avant-plan et extraire les informations utiles à partir de l'image.

Dans le chapitre suivant nous verrons en détail quelques méthodes de segmentation d'images de documents manuscrits. Particulièrement nous étudierons les techniques de segmentation

Chapitre III : Segmentation d'images de documents manuscrits

1. Introduction :

La segmentation d'images de documents est une étape indispensable pour un système d'extraction de métadonnées. Par exemple pour extraire les lignes de texte, la segmentation partitionne le texte en lignes, mais encore une ligne en mots et un mot en caractères [GKG 14]. Pour notre cas d'étude, la segmentation vise à binariser l'image en conservant uniquement le texte et le graphique, à séparer les lignes de texte pour pouvoir les compter, et à distinguer l'élément graphique de l'élément textuel.

Dans ce chapitre nous donnerons quelques méthodes de segmentation d'images de documents. Nous commençons par la binarisation avec la méthode d'Otsu. Ensuite nous définirons une méthode de segmentation texte/graphique basée sur la morphologie mathématique. Enfin nous présenterons deux méthodes de segmentation des lignes de texte : celle basée sur la projection, et la segmentation RLSA.

2. La binarisation :

La binarisation est parfois nécessaire pour distinguer le fond de l'image des autres éléments (textuels ou graphiques) [Lik 03]. Le choix du type d'algorithme influe sur le résultat de la binarisation.

2.1.La méthode d'Otsu :

La méthode d'Otsu [Ots 79] est un algorithme de binarisation globale (cf. page 30). Pour une image, les données sont les niveaux de gris des pixels de l'image (compris entre 0 et 255). La méthode d'Otsu travaille avec des images en niveaux de gris mais elle nécessite au préalable le calcul d'histogramme normalisé de l'image. On obtient l'histogramme normalisé en divisant l'histogramme sur taille de l'image.

Le principe consiste à trouver le seuil qui minimise la variance intra-classe pondérée des pixels.

$$\sigma_w{}^2(t) = q_1(t)\sigma_1{}^2(t) + q_2(t)\sigma_2{}^2(t),$$

Avec :

- $t \in [0; 255]$ le seuil de séparation des deux classes.
- Les quantités $q_1(t)$ et $q_2(t)$ représentent les proportions relatives des deux classes ($q_2(t) = 1 - q_1(t)$)

$$q_1(t) = \sum_{i=0}^{t} P(i)\,, q_2(t) = \sum_{i=t+1}^{255} P(i),$$

Et où

– $\mu_1(t)$ et ${\sigma_1}^2(t)$ représentent la moyenne et la variance de la classe constituée des pixels de niveau de gris compris dans l'intervalle [0; t], et symétriquement

– $\mu_2(t)$ et ${\sigma_2}^2(t)$ représentent la moyenne et la variance de la classe constituée des pixels de niveau de gris compris dans l'intervalle [t+1; 255],

Avec

$$\mu_1(t) = \frac{1}{q_1(t)} \sum_{i=0}^{t} i * P(i)\,, \mu_2(t) = \frac{1}{q_2(t)} \sum_{i=t+1}^{255} i * P(i),$$

Et

$${\sigma_1}^2(t) = \frac{1}{q_1(t)} \sum_{i=0}^{t} (i - \mu_1(t))^2 * P(i) = \frac{1}{q_1(t)} \sum_{i=0}^{t} i^2 * P(i) - {\mu_1}^2(t),$$

$${\sigma_2}^2(t) = \frac{1}{q_2(t)} \sum_{i=t+1}^{255} (i - \mu_2(t))^2 * P(i) = \frac{1}{q_2(t)} \sum_{i=t+1}^{255} i^2 * P(i) - {\mu_2}^2(t),$$

- **Algorithme d'Otsu :**

– Calculer pour chaque valeur de t dans [0; 255], la variance intra-classe pondérée.

– Sélectionner la valeur t qui minimise la variance intra-classe pondérée.

L'algorithme se programme uniquement à partir de l'histogramme normalisé de l'image et non de l'image elle-même. Pour accélérer encore les calculs, la relation entre les variances intra-classe et inter classe peut être utilisée.

La moyenne totale de l'image μ s'écrit, quelque soit t dans [0; 255],

$$\mu = q_1(t)\mu_1(t) + q_2(t)\mu_2(t),$$

Et la variance totale de l'image σ^2

$$\sigma^2 = \sum_{i=0}^{255} i^2 * P(i) - \mu^2 = \sum_{i=0}^{t} i^2 * P(i) + \sum_{i=t+1}^{255} i^2 * P(i) - (q_1\mu_1 + q_2\mu_2)^2$$
$$= q_1{\sigma_1}^2 + q_2{\sigma_2}^2 + {\mu_1}^2(q_1 - {q_1}^2) + {\mu_2}^2(q_2 - {q_2}^2) - 2q_1q_2\mu_1\mu_2$$

(Avec ${\sigma_W}^2 = q_1{\sigma_1}^2 + q_2{\sigma_2}^2$)

Or, comme $q_2 = 1 - q_1 et\ (q_1 - {q_1}^2) = (q_2 - {q_2}^2)$, nous pouvons écrire :

$$\sigma^2 = {\sigma_w}^2 + {\sigma_b}^2,$$

Avec ${\sigma_b}^2 = q_1(1 - q_1)(\mu_1 - \mu_2)^2$, la variance inter classe (’b’ = between).

Comme la somme est constante et indépendante de t, changer le seuil a pour effet de changer en sens inverse les contributions respectives de chacun des deux termes. Ainsi, minimiser la variance intra-classe revient à maximiser la variance inter classe. Or cette variance inter classe peut se calculer progressivement, alors que t augmente de 0 à 255 :

- t=0: $q_1(0) = P(0), \quad \mu_1(0) = 0.$

- t=1..254 : $q_1(t) = q_1(t-1) + P(t),$

$$\mu_1(t) = \frac{q_1(t-1)\mu_1(t-1) + t * P(t)}{q_1(t)},$$

$$\mu_2(t) = \frac{\mu - q_1(t)\mu_1(t)}{1 - q_1(t)}.$$

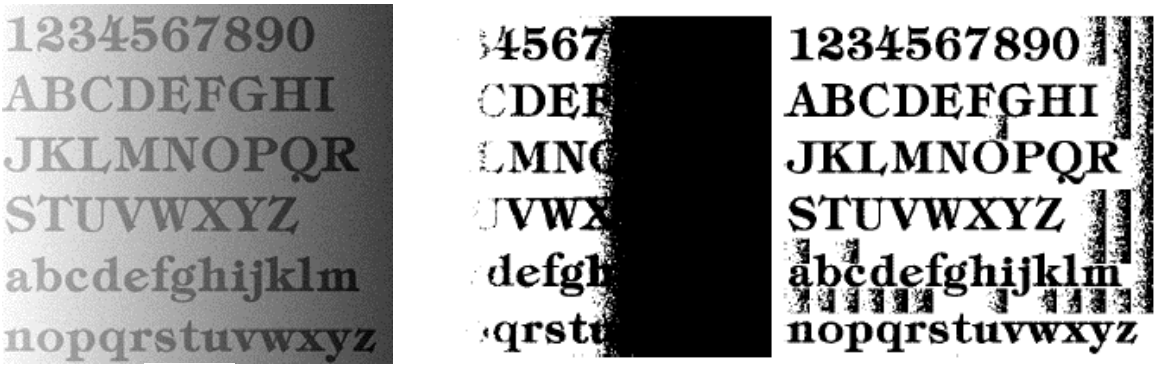
- t=255 : $q_1(255) = 1,$

$$\mu_1(255) = \mu,$$

$$\mu_2(255) = 0.$$

2.2.Limites de la méthode d’Otsu :

La méthode d’Otsu n’est pas efficace lorsque l’image présente un éclairage non uniforme ou insuffisant. Pour pallier à ce problème, dans [GJG 07][NS][ZWU 11], l’image est d’abord divisée en petits blocs nxn. On applique ensuite la méthode d’Otsu localement (pour chaque bloc) (voir Fig.3.1). Différents prétraitements comme la modification d’histogramme (cf. page 14) peuvent améliorer la luminosité de l’image. Ainsi, le problème d’éclairage aura moins d’impact sur la méthode d’Otsu.



(a) (b) (c)

Fig. 3.1 : Binarisationd'Otsu [ZWU 11].

(a) Image originale ; (b) méthode d'Otsu globale ; (c) méthode d'Otsu adaptative.

3. La segmentation texte/graphique :

Une fois l'image binarisée, la segmentation de documents se produit sur deux niveaux [OGK 09]:

- Au premier niveau, si le document contient à la fois du texte et du graphique, ceux-ci sont séparés pour un traitement ultérieur par des méthodes différentes ;
- Au deuxième niveau,la segmentation est effectuée sur le texte en localisant les colonnes, les paragraphes, les mots, les titres et les légendes (d'une figure par exemple). Elle est aussi appliquée sur les graphiquesgénéralement pour la séparation de symboles et de composants géométriques (lignes, cercles…).

La segmentation Texte/graphique a de nombreuses applications et reste la première étape incontournable pour l'interprétation et l'indexation des images de documents [Mou 06].

[Gra 00] ont effectué la segmentation des images en niveaux de gris et l'identification des blocs de texte et des figures de livres anciens à l'aide de la morphologie mathématique (cf. page 27). Les composants à détecter (texte, lignes et figures) sont considérés comme des « vallées » ou « sillons » découpés au-dessus d'un fond complètement plat. Pour les mettre en valeur, ils sont complétés et soustraits à l'image initiale. Cette opération conduit à la production d'un masque, le plus précis possible, qui aide à l'émergence de tous les éléments de type texte et figures. Le masque est conçu en exploitant l'information morphologique du gradient. Ceci permet d'éliminer tous les objets qui ont des contours plus doux que les lignes (comme les taches).

4. La segmentation des lignes de texte :

4.1.Difficulté de segmentation :

Plusieurs facteurs rendent difficile la segmentation des lignes de texte, notamment pour les textes manuscrits. Parmi ces facteurs [GKG 14]:

- Chevauchement et contact entre les lignes de texte. Il devient très difficile d'identifier les limites d'une ligne de texte.
- Le style d'écriture de l'auteur : l'orientation des lignes et l'espacement entre les lignes ne sont pas uniformes ou régulières.
- La mauvaise qualité des documents due à leur dégradation ;
- Régularité faible et longueurs différentes des lignes [Mou 06].

4.2.Méthodes de segmentation des lignes de texte:

La présence de traits d'écriture est généralement détectée par un filtre passe-hautrecherchant les pixels correspondants au passage écriture/fond (voir Fig. 2.10, Fig. 2.11), donc à des transitions rapides. Cela permet de repérer les traits plus clairs de l'écriture et de différencier l'écriture des taches [Lik 03].

Parmi plusieurs méthodes de segmentation des lignes, nous citons entre autres :

- Méthodes basées sur la projection (méthode descendante) ;
- Le RLSA (méthode ascendante) ;

4.2.1. Méthode basée sur la projection des pixels :

Un profil de projection horizontale (respectivement vertical) est obtenu en additionnant les valeurs des pixels sur l'axe horizontal (respectivement vertical). Les lignes de texte peuvent être déterminées en cherchant deux vides consécutifs dans la projection horizontale [Ala 14] [Ouw 10].

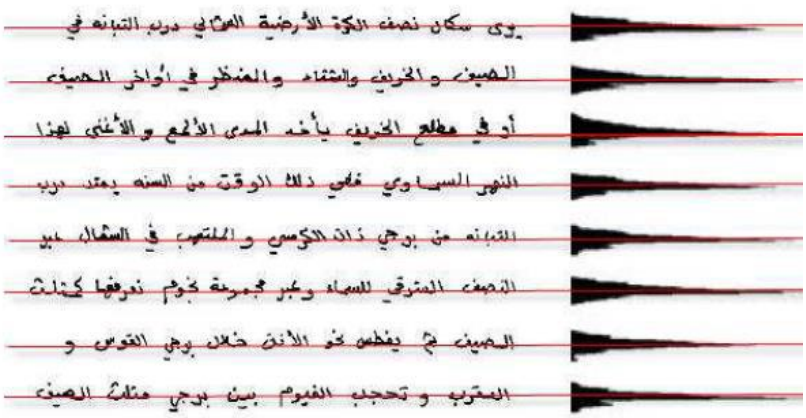


Fig. 3.2 : profil de projection horizontale. Les lignes rouges sont tracées à partir des pics du profil [Ala 14].

Certains documents peuvent présenter une sinuosité ou une mal orientation des lignes. Dans [Ven], l'image est divisée en colonnes. Une projection horizontale est appliquée pour chaque colonne. Les valeurs minimales (ou les vides) de chaque colonne représentent une séparation possible entre deux lignes. Un trait horizontal est tracé entre deux lignes de chaque colonne.

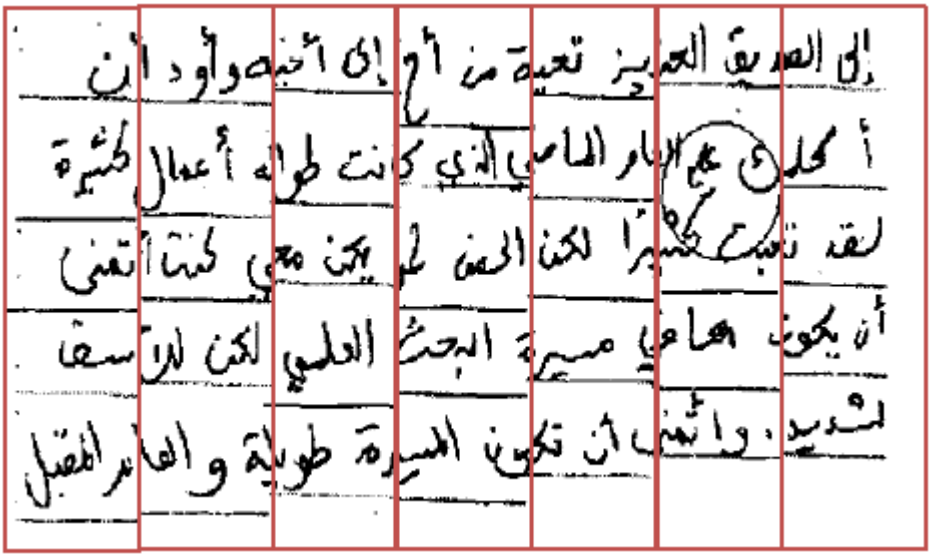


Fig. 3.3 : Division de l'image en plusieurs colonnes. Les traits horizontaux séparent entre deux lignes. L'image montre aussi un chevauchement entre deux lignes [Ala 14].

4.2.2. La segmentation RLSA (RunLengthSmoothingAlgorithm) :

C'est une méthode itérative basée sur des opérations morphologiques de traitement d'images. Le principe de cette méthode est de noircir toute séquence de pixels blancs comprise entre deux pixels noirs, de longueur inférieure à un seuil donné. En pratique l'algorithme est appliqué horizontalement et verticalement sur l'image binaire originale avec des seuils éventuellement différents pour l'horizontale et la verticale, puis une opération «ET logique » est appliquée entre les deux images lissées obtenues. L'extraction des composantes connexes de l'image résultante permet d'obtenir les entités de la structure physique sur un niveau hiérarchique donné. On peut ainsi, en répétant la procédure avec des seuils de lissage horizontal et vertical différents, extraire itérativement les blocs de l'image, puis les lignes de texte et les mots. Ces seuils de lissage sont les seuls paramètres de l'algorithme RLSA. Ils contrôlent la manière dont les composantes sont fusionnées sur un niveau de segmentation prédéterminé. Par exemple pour segmenter des lignes de texte horizontales, droites et bien espacées, on pourra utiliser un seuil de lissage verticalnul et un seuil de lissage horizontal suffisamment grand pour combler les espaces inter-lettres etinter-mots[Lel 07].

Cet algorithme a l'avantage, en plus d'être très simple à mettre en œuvre, de ne requérir qu'un nombre limite d'itérations pour atteindre la segmentation complète du document. La principale difficulté dans l'utilisation de cet algorithme est le réglage des seuils de lissage qui est délicat. Il est nécessaire de déterminer les seuils adéquats pour chaque niveau de segmentation. De tels seuils ne peuvent être déterminés qu'empiriquement. De plus, pour une

itération donnée, ces seuils sont constants. Il faut donc que pour le niveau de segmentation considéré, les espaces entre les composantes de l'image à fusionner soient aussi constants[Lel 07].

(a) (b) (c) (d)

Fig. 3.4 : Segmentation RLSA [Lel 07] :

(a) image originale ; (b) lissage horizontal ; (c) lissage vertical ; (d) résultat RLSA.

5. Conclusion :

Nous avons présenté quelques méthodes de segmentation d'images de documents. Nous avons vu qu'elles ont des avantages et des inconvénients. La méthode d'Otsu donne de bons résultats si l'image est en bonnes conditions. Dans le cas contraire un prétraitement est nécessaire pour palier à ce problème. La méthode du RLSA est simple à mettre en œuvre mais vu que les manuscrits anciens présentent des irrégularités, le paramètre de lissage introduit par l'utilisateur rend la méthode très exposée à des erreurs de segmentation. Dans la partie suivante nous allons modifier la méthode du RLSA de manière à ce qu'elle soit semi-automatique. Le résultat de cette segmentation va être utilisé pour extraire le nombre de lignes du texte avec la projection horizontale. Pour extraire les figures à partir d'un document manuscrit ancien, nous utiliserons les opérations morpho mathématiques, notamment la dilatation et l'érosion.

Partie 2 : Conception et réalisation

Chapitre IV : Extraction du nombre de lignes et extraction de figures

1. Introduction :

Précédemment, nous avons présenté les concepts de prétraitements et de segmentation. Dans le présent chapitre, ces différents outils seront utilisés afin de concevoir un système d'extraction automatique de métadonnées à partir d'images de manuscrits anciens numérisés.

Deux aspects fondamentaux seront passés en revue :

- Le comptage du nombre de lignes d'une page quelconque d'un manuscrit. En effet, le *catalogueur*[6] rencontre d'énormes difficultés pour extraire manuellement le nombre de lignes. Pour ce faire, nous modifierons la méthode du RLSA pour la segmentation des lignes de texte puis nous utiliserons l'histogramme de projection horizontal pour extraire le nombre de lignes. Notre corpus d'images sera constitué de documents mono-orientés (cf. page 5) de manière horizontale;
- La présence de figures dans les images de documents manuscrits. Nous effectuerons cette extraction avec un filtre alterné séquentiel (cf. page 54).

Nous donnerons d'abord un schéma général de notre système, ensuite nous expliquerons l'algorithme d'extraction de chacune des métadonnées suscitées.

2. Description générale du système d'extraction de métadonnées :

Le schéma suivant (voir Fig. 4.1) représente le système global pour l'extraction de métadonnées.

[6] Personne chargée de créer des notices descriptives des documents.

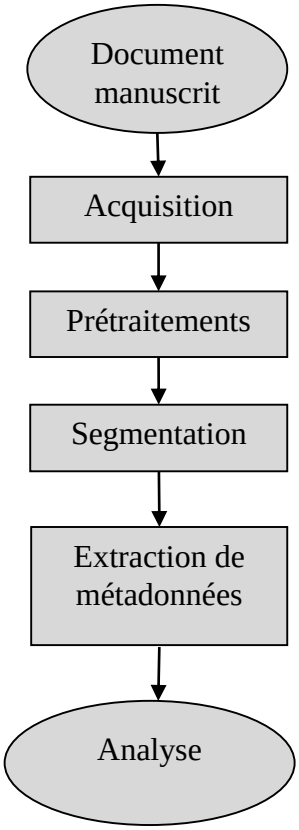


Fig. 4.1 : Schéma général du système d'extraction de métadonnées.

2.1.Acquisition :

C'est la première étape du système. Le document est numérisé avec un scanner ou un appareil photo numérique. Il faut veiller à ne pas lui causer des dégradations (cf. page 6).

2.2.Prétraitements :

Cette étape consiste à préparer l'image pour la segmentation. L'image est rehaussée par la méthode de modification d'histogramme et des outils de filtrage. Les deux actions ont pour but de :

- simplifier l'image en mettant en avant les informations utiles (texte ou graphique);
- améliorer l'image en réduisant au mieux les bruits et les dégradations (taches, luminosité…).

2.3.Segmentation :

Elle consiste à découper l'image en un ensemble de segments. Ces segments peuvent être du texte ou du graphique dans le cas du document manuscrit.

2.4.Extraction :

Cette phase permet de détecter et d'extraire les métadonnées à partir de l'image segmentée.

2.5.Analyse :

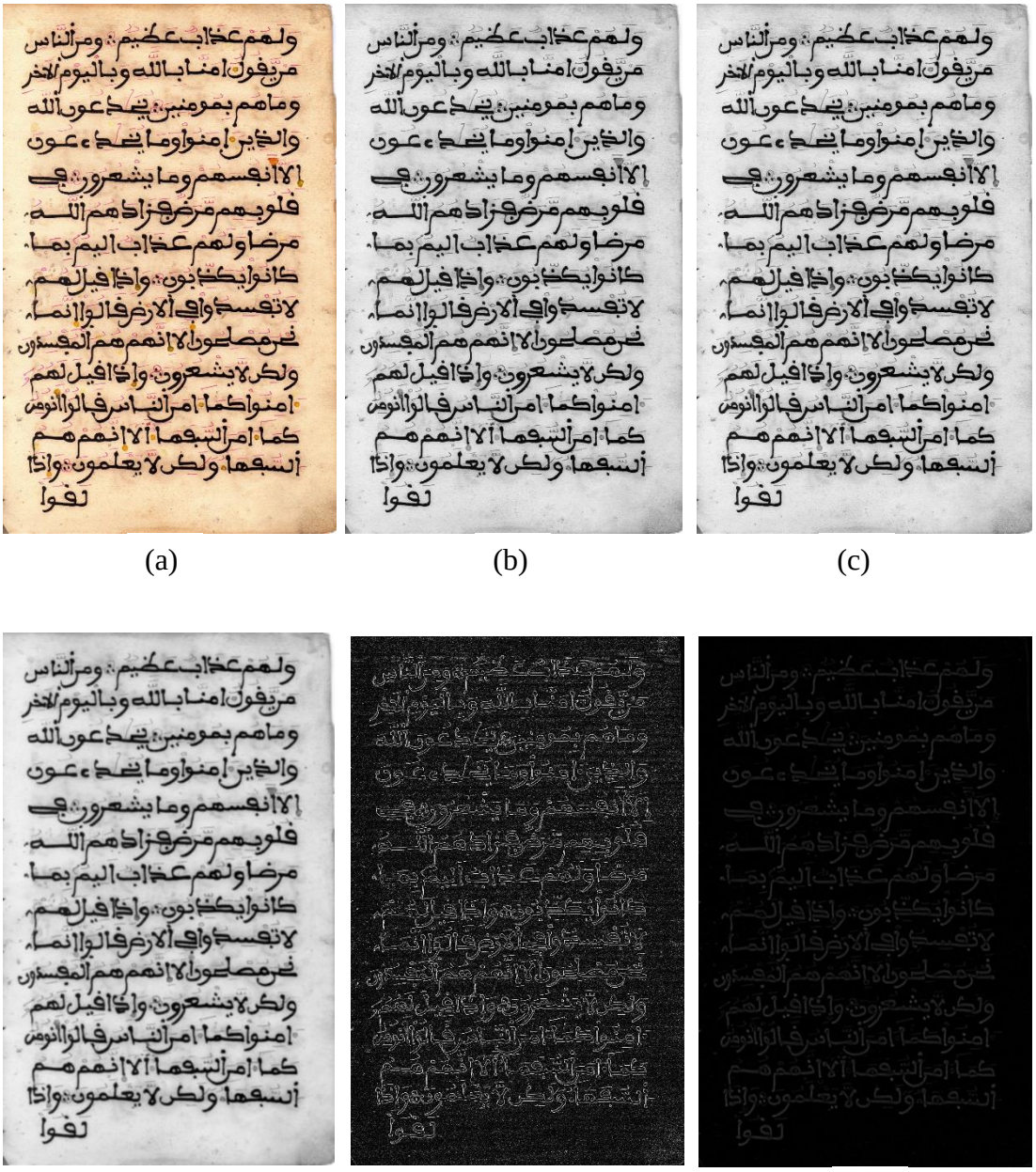
La phase d'analyse permet de statuer sur la nature de la métadonnée extraite et émettre un avis sur la structure du manuscrit tels que sa composition, sa complétude, présence d'ornementations, …etc.

3. Algorithme d'extraction du nombre de lignes :

3.1.Prétraitements :

- Mise en niveaux de gris de l'image (cf. page 31) ;
- Etalement (cf. page 16) de l'image en niveaux de gris ;
- Application d'un filtre gaussien (cf. page 22) itéré 3 fois sur l'image étalé avec une fenêtre 5x5 et un paramètre $\sigma = 1$. Cette itération sert à réduire les bruits et à enlever les petits détails dans l'image comme les ponctuations dans un document manuscrit arabe ;
- Application d'un filtre Laplacien 8-connexe (cf. page 23) à l'image obtenue pour détecter les contours et ignorer certains défauts comme les taches ;
- Application d'un filtre gaussien sur l'image obtenue avec une fenêtre 5x5 et un paramètre $\sigma = 5$. Puisque l'image contient uniquement les contours des objets, notamment les contours des caractères, cette opération sert à flouter plus l'image pour pouvoir remplir l'intérieur des objets;

Cette procédure est décrite dans la figure 4.3 :



(a) (b) (c)

(d) (e) (f)

Fig. 4.2 : Phase de prétraitements dans l'algorithme d'extraction du nombre de lignes : (a) Image originale ; (b) image en niveaux de gris ; (c) image étalée ; (d) filtrage gaussien 5x5 avec $\sigma = 1$itéré 3 fois ; (e) filtrage Laplacien 8-connexe ; (f) filtrage gaussien 5x5 avec $\sigma = 5$. (f) représente l'image prétraitée.

3.2.Segmentation :

L'image prétraitée va être segmentée sur deux niveaux :

- Dans un premier temps, l'avant-plan (le texte) est isolé de l'arrière-plan (le fond). Pour cela, nous utilisons la méthode de binarisation d'Otsu (cf. page 36).L'avant plan de l'image binarisée est représenté par la couleur blanche et l'arrière-plan est représenté par la couleur noire ;

- Dans le second temps, la méthode du RLSA (cf. page 41) modifiée est appliquée pour la segmentation des lignes du texte.

L'extraction est appliquée sur une image binaire. Pour pouvoir compter le nombre de lignes, il faut d'abord segmenter le texte en lignes afin de les séparer. Pour ce, nous modifionsl'algorithme du RLSA (cf. page 48).

3.2.1. Modification du RLSA :

La méthode du RLSA est une méthode manuelle, c'est-à-dire que l'utilisateur doit fournir un paramètre (ou un seuil) pour appliquer le RLSA. Ce seuil ne peut pas être le même pour toutes les images. Dans la méthode proposée, nous modifions le RLSA pour qu'il soit semi-automatique. Lorsque l'algorithme parcours l'image horizontalement ou verticalement, arrivé à un pixel blanc donné, il calcule le nombre de pixels consécutifs de la même couleur. Ce nombre est ensuite multiplié par un facteur. Le seuil étant le résultat de cette multiplication. Par conséquent, au lieu de donner un seuil, c'est un facteur qui est fourni.

Par exemple : Si (Séquence continue de pixels=10)

Alors Seuil RLSA = 10 x facteur

Cette idée vient du fait que les caractères ont une longueur horizontal assez grande pour combler le vide entre le caractère courant et le caractère suivant. Ce processus vise à réduire la sensibilité de l'algorithme à la variation de la taille des écritures et des images.

Pour mieux afficher les lignes de texte, nous conservons uniquement les pixels auxquels on applique le RLSA (ceux de l'image originale et ceux ajoutés). Cette conservation vise à éliminer les contacts ou chevauchements entre deux caractères de lignes distinctes. Par ailleurs, pour plus de connexion sur chaque ligne, le RLSA va parcourir l'image de gauche à droite, et de droite à gauche. Ensuite, les deux images obtenues sont fusionnées avec un « OU » logique.

3.2.2. Algorithme de segmentation en lignes :

L'algorithme de segmentation en lignes est le suivant :

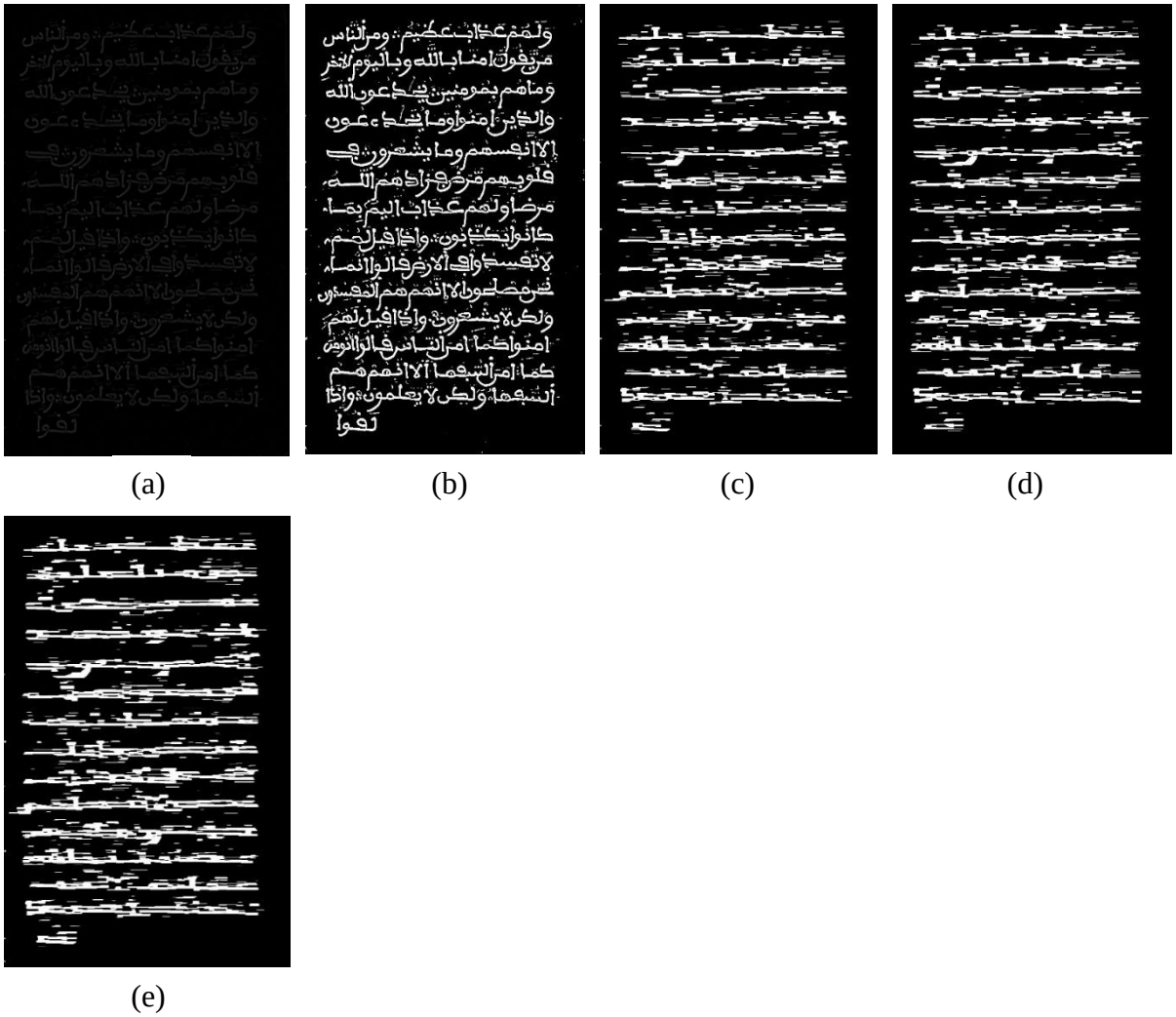
- Binariser l'image prétraitée (voir Fig. 4.2f et Fig. 4.3a) avec la méthode d'Otsu (voir Fig. 4.3b) ;

- Appliquer le RLSA horizontal semi-automatique (cf. page 48) sur l'image binaire de droite à gauche (voir Fig. 4.3c) avec un facteur f=1,5. Ce facteur est défini empiriquement et il donne de bons résultats ;

- Appliquer le RLSA horizontal semi-automatique sur l'image binaire de gauche à droite (voir Fig. 4.3d) avec un facteur f=1,5 ;

- Fusionner les deux images obtenues avec un « OU » logique (voir Fig. 4.3e).

Cette procédure est mise en évidence par la figure 4.3:



(a) (b) (c) (d)

(e)

Fig. 4.3 : Phase de segmentation dans l'algorithme d'extraction du nombre de lignes : (a) image prétraitée ; (b) image binarisée ; (c) RLSA semi-automatique de 'b' de droite à gauche avec un facteur f=1.5 ; (d) RLSA semi-automatique de 'b' de gauche à droite avec un facteur f=1.5 ; (e) OU logique entre 'c' et 'd'.

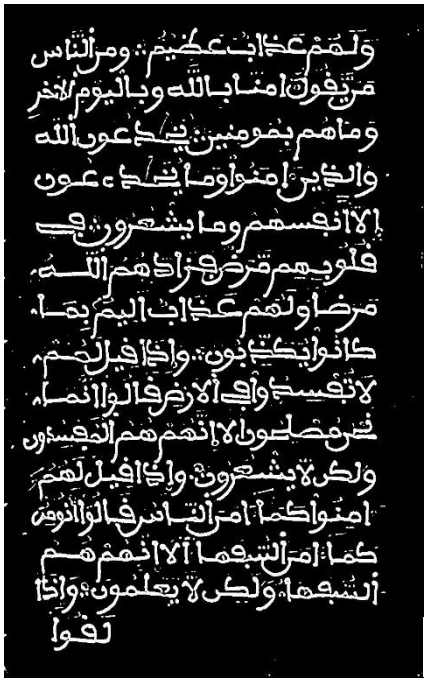
3.3.Extraction :

A partir de l'image segmentée, les métadonnées sont extraites. Pour ce faire, nous utilisons la projection horizontale (cf. page 40), ce qui aura pour effet d'extraire le nombre de lignes d'un texte.

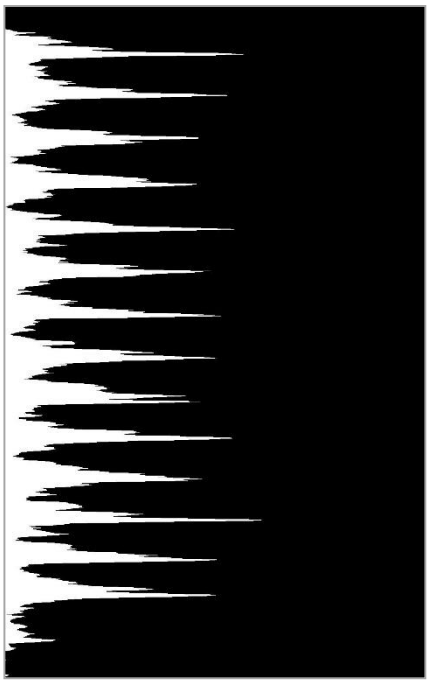
Les lignes du texte peuvent avoir un contact entre elles (connexité verticale). Toutefois, la méthode du RLSA semi-automatique permet d'extraire ces linges tout en éliminant les chevauchements entre ces lignes. De ce fait, les lignes du texte seront séparées entre elle comme le montre la figure 4.4. Par conséquent, les lignes du texte seront comptabilisées selon la procédure suivante :

- Compter les transitions entre le blanc et le noir dans la projection en prenant en considération uniquement les blocs supérieurs à un certain seuil dans l'histogramme.

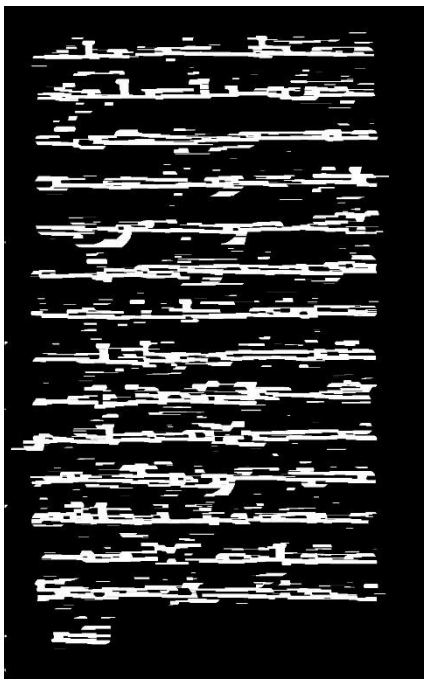
Nous avons testé l'extraction avec différentes seuils, parmi lesquels (3/5) x (profondeur de l'image) est le plus adéquat. Les lignes inférieures à ce seuil ne sont pas comptabilisées comme étant des lignes de texte. Par exemple dans la figure 4.4, la dernière ligne n'est pas comptabilisée. Par contre cette ligne n'est pas considérée comme une ligne textuelle car il s'agit d'une *réclame*[7].



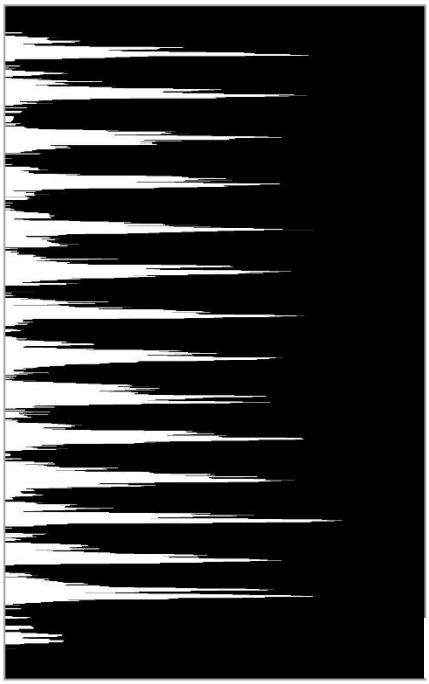
(a)



(b)



(c)



(d)

[7] Mention sur une page du premier mot de la page suivante.

Fig. 4.4 : Comparaison entre l'histogramme de projection d'une image binaire avec celui de sa segmentation RLSA semi-automatique : (a) image binaire ; (b) projection horizontale de 'a' ; (c) RLSA semi-automatique de 'a' ; (d) projection horizontale de 'c'.

La figure 4.5 résume le système d'extraction du nombre de lignes conçu.

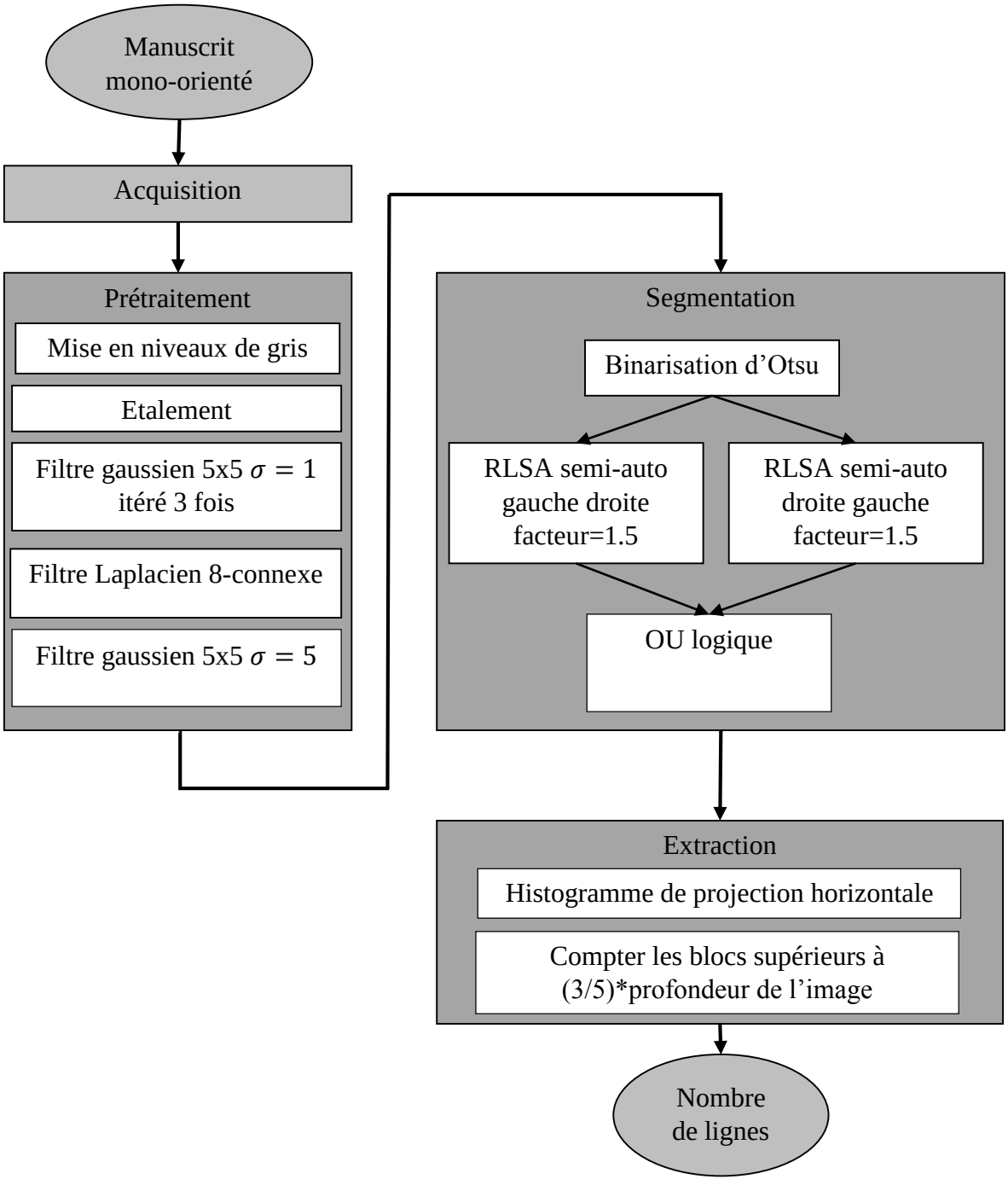
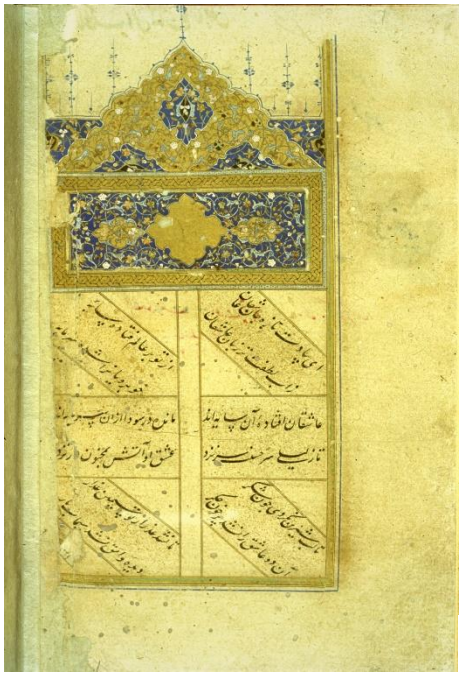


Fig. 4.5 : Algorithme d'extraction du nombre de lignes.

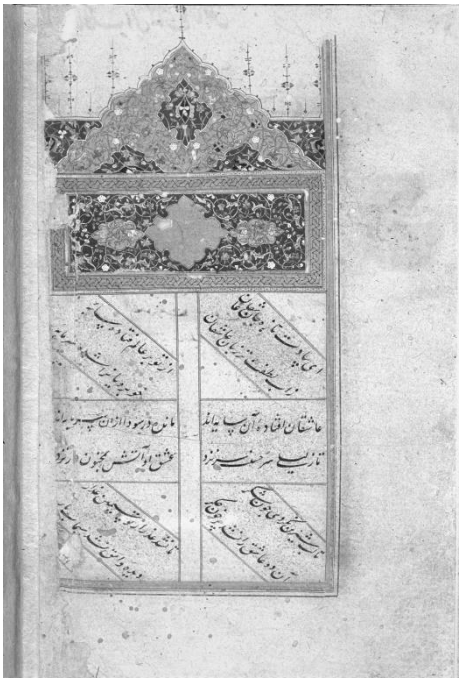
4. Algorithme d'extraction de figures :

4.1.Prétraitements :

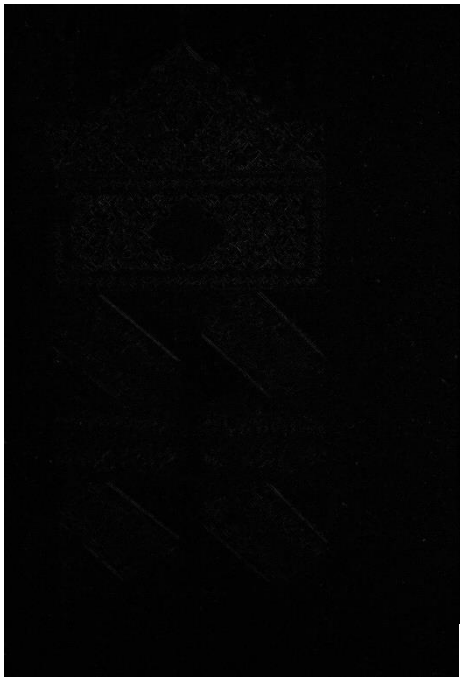
- Mise en niveaux de gris de l'image ;
- Etalement de l'image en niveaux de gris ;
- Application du filtre Laplacien de Robinson (cf. page 23) à l'image obtenue pour détecter les contours et ignorer certains défauts comme les taches. Les contours détectés sont discontinus (voir Fig. 2.10d). Ce filtre a été utilisé dans le but de transformer l'image en un nuage de points. Le nuage de points est un indice frappant pour la présence de figures, de dessins ou d'ornementations. Les figures seront repérées dans les zones où les points sont plus denses ;
- Application d'un filtre gaussien sur l'image obtenue avec une fenêtre 5x5 et un paramètre $\sigma = 1$. Puisque l'image contient uniquement les contours des objets, notamment les contours des caractères, cette opération sert à flouter plus l'image pour pouvoir remplir l'intérieur des objets ;



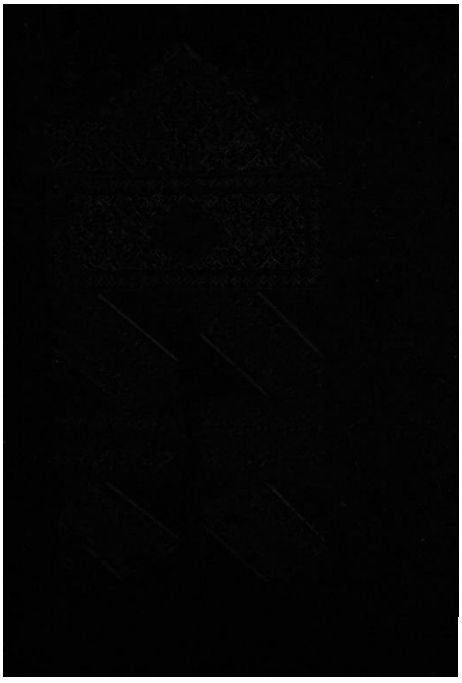
(a)



(b)



(c)



(d)

Fig. 4.6 : Phase de prétraitements dans l'algorithme d'extraction de figures :

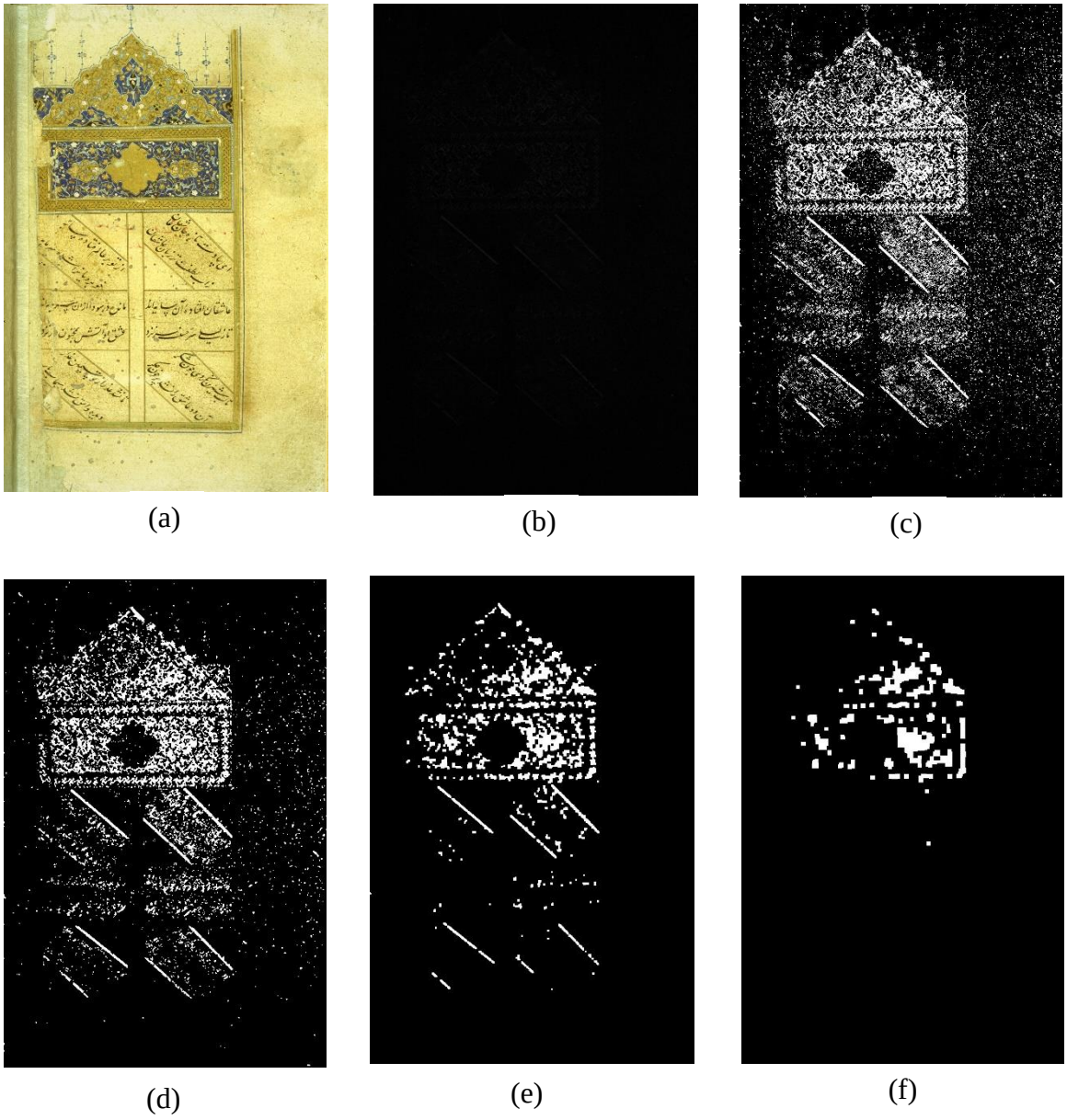
(a) Image originale ; (b) image en niveaux de gris étalée; (c) filtrage Laplacien de Robinson ; (d) filtrage gaussien 5x5 avec $\sigma = 1$; (d) représente l'image prétraitée.

4.2.Segmentation :

- Binariser l'image prétraitée avec la méthode d'Otsu.

4.3.Extraction :

- Appliquer un filtre alterné séquentiel sur l'image binarisée. Il s'agit d'une succession d'ouvertures et de fermetures (cf. page 28) utilisant des éléments structurants (cf. page 27)croissants [BTM]. Les éléments structurants sont de type 8-connexe (cf. page 23) et varie d'une taille 3x3 jusqu'à 13x13. Ce qui veut dire que ce filtre s'exécutera six fois. Nous utilisons ce filtre pour regrouper les points qui sont très proches entre eux. Les points isolés seront effacés.



(a) (b) (c)

(d) (e) (f)

Fig. 4.7 : Phase de segmentation et d'extraction dans l'algorithme d'extraction de figures : (a) image originale ; (b) image prétraitée ; (c) image binarisée ; (d) Filtre séquentiel alterné après 1 itération ; (e) Filtre séquentiel alterné après 3 itération ; (f) Filtre séquentiel alterné après 6 itération ;

La figure 4.7f montre que le graphique de l'image 4.7a est détecté partiellement. Nous pouvons utiliser cet algorithme pour détecter la présence de figures dans une image de manuscrit ancien.

La figure 4.9 résume le système d'extraction de figures conçu.

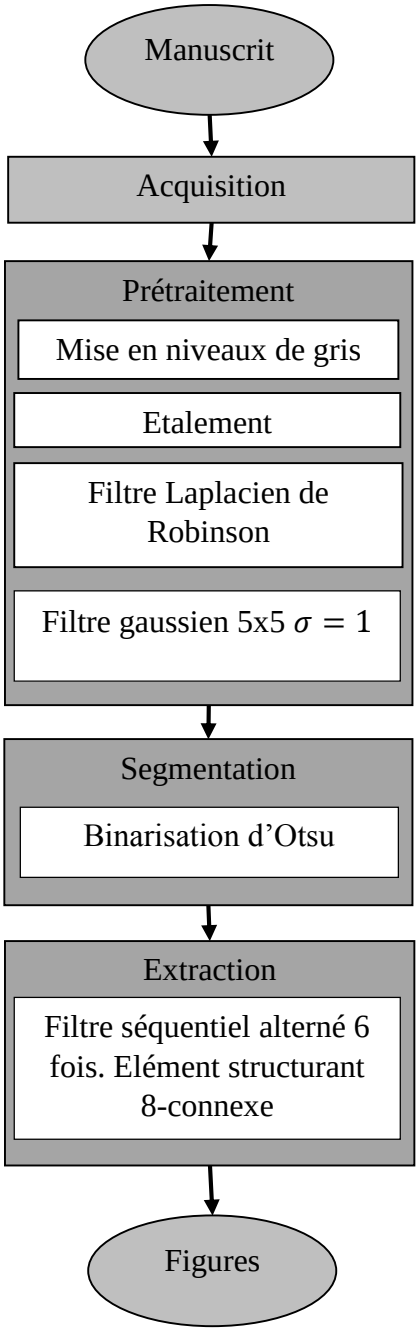


Fig. 4.8 : Algorithme d'extraction de figures.

5. Conclusion :

Nous avons élaboré un algorithme d'extraction du nombre de lignes d'un texte manuscrit mono-orienté. Pour ce faire, nous avons d'abord effectué un prétraitement sur l'image pour faire ressortir uniquement le texte et éviter les dégradations comme les taches ou le défaut de luminosité. Ensuite nous avons effectué une binarisation sur l'image prétraitée pour afficher uniquement le texte. Nous avons segmentée l'image binarisée avec à la méthode RLSA que nous avons au préalable modifié. L'image résultante de cette segmentation nous permet de déterminer le nombre de lignes dans une page de manuscrit en exploitant son histogramme de projection horizontal. La détermination du nombre de ligne pourrait renseigner les usagers sur plusieurs caractéristiques du manuscrit, telles que l'hétérogénéité du manuscrit la présence de réglure.

Nous avons aussi conçu un algorithme de détection de présence de graphique dans un manuscrit ancien. Nous avons d'abord appliqué un filtre qui nous donne un nuage de points. Les points les plus denses sont regroupés grâce au filtre alterné séquentiel.

Notre approche montre la possibilité d'utiliser les outils de traitement d'image pour assister les usagers, tel que le catalogueur dans la définition de certains aspects du manuscrit qui pourrait se révéler être fastidieux manuellement.

Dans le chapitre suivant, nous présenterons l'application de mise en œuvre des différents aspects présentés dans ce chapitre. Elle l'implémentation et l'évaluation des deux algorithmes conçus.

Chapitre V : Réalisation

1. Introduction :

Dans le chapitre précédent nous avons présenté le concept général pour l'extraction d'une métadonnée à partir d'images de manuscrits anciens. Nous avons conçu et détaillé l'algorithme d'extraction du nombre de lignes d'un texte manuscrit ainsi que celui des figures.

Dans ce chapitre nous présenterons les outils utilisés pour la réalisation de notre système d'extraction des deux métadonnées. Après implémentation, nous établirons un ensemble de tests pour évaluer la performance de nos algorithmes.

2. Présentation des outils utilisés :

2.1.Environnement matériel :

L'application a été développée sur un ordinateur portable LENOVO G510 qui se caractérise par :

- Processeur : intel® core™ i7 4702MQ @ 2.20 GHz ;
- Mémoire installée (RAM) : 8.00 Go ;
- Système: système d'exploitation 64 bits, processeur x64 ;
- Ecran : 15 pouces.

2.2.Environnement logiciel :

2.2.1. Système d'exploitation :

- Microsoft Windows 8.1 Professionnel.

2.2.2. Langage de programmation :

L'application a été développée en JAVA car il est de plus en plus utilisé dans le monde de la recherche scientifique ainsi que dans l'industrie. En effet, ce langage de programmation présente un large avantage grâce à sa portabilité. Les programmes java peuvent être exécutés sur différentes plateformes.

Quelques chiffres et faits à propos de Java en 2011 [Dou]:

- 97% des machines d'entreprises ont une JVM installée ;
- Java est téléchargé plus d'un milliards de fois chaque année ;
- Il y a plus de 9 millions développeur java dans le monde ;
- Java est un des langages les plus utilisés dans le monde ;

- Tous les lecteurs de Blue-Ray[8] utilisent Java ;
- Plus de 3 milliards d'appareils Java mettent en œuvre Java.

2.2.3. Outil de programmation :

L'outil utilisé est eclipse version 4.2.2. C'est un IDE (EDI : environnement de développement intégré), qui simplifie la programmation en proposant un certain nombre de raccourcis et d'aide à la programmation. Il est développé par IBM, est gratuit et disponible pour la plupart des systèmes d'exploitation.

3. Evaluation du système d'extraction du nombre de lignes :

Nous avons testé le système d'extraction du nombre de lignes conçu sur un *corpus*[9]d'images de manuscrits numérisés, issu de la mosquée de DBK (Draâ Ben Khedda). Il comporte trente images, la première contient un graphique et les 29 autres sont mono-orientées (sans graphique). Nous avons évalué notre système d'extraction du nombre de lignes avec ces 29 images. Les résultats d'évaluation sont les suivants :

Nombre total de documents	29
Nombre de documents avec une extraction complète	5
Nombre de documents avec une extraction incomplète	24
Taux de bonne extraction	**17%**

Tab. 5.1 : Taux de bonne extraction du nombre de lignes.

Nombre total de lignes des 29 documents	421
Nombre total de lignes extraits automatiquement	373
Taux de lignes extraites	**88.6%**

Tab. 5.2 : Taux de lignes extraites sur tous les documents.

Nous remarquons que le nombre de documents avec une extraction complète du nombre de lignes est très faible (voir Tab. 5.1). Par ailleurs, le taux d'extraction de lignes (voir Tab. 5.2) est assez élevé. Ces deux résultats peuvent être interprétés comme suit :

[8]Format de disque numérique breveté et commercialisé par l'industriel japonais Sony permettant de stocker et restituer des vidéogrammes en haute définition

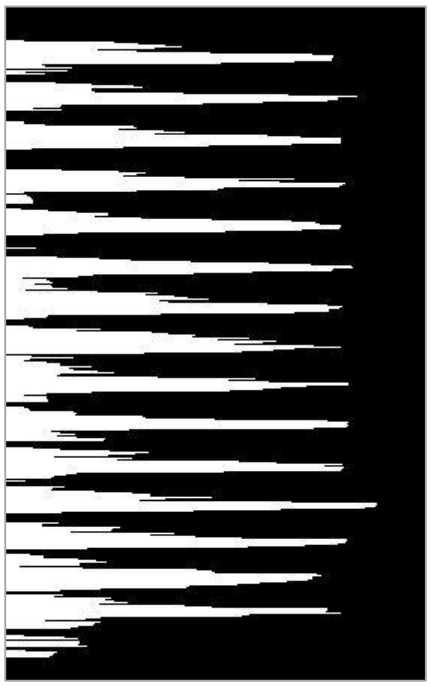
[9] Ensemble de documents regroupés en vue de leur conservation.

- Sur chaque document avec une extraction incomplète, le nombre de lignes extrait est inférieur au nombre réel de lignes. Cela veut dire que l'algorithme d'extraction peut considérer deux (ou plus) lignes comme étant une seule ligne (voir Fig. 5.1). L'algorithme du RLSA modifié ne réussit pas toujours à séparer les lignes.
- Le taux de lignes extraites est élevé. Ce résultat est très intéressant car l'algorithme fonctionne bien si les lignes sont bien séparées avec la méthode du RLSA modifié. En perspective, l'algorithme pourra être amélioré par la suite. L'objectif est de pouvoir séparer encore plus de lignes entre elles et détecter aussi les fins de lignes (souvent de petite taille).

Le corpus que nous avons testé comporte des images avec des réclames. Sur tous les documents l'algorithme ne les comptabilise pas.

La réclame n'étant pas considérée comme une ligne de texte, nous pouvons dire que le système d'extraction du nombre de lignes ne comptabilise pas les réclames comme des lignes de texte.

(a)



(b)

Fig. 5.1 : Erreur d'extraction du nombre de lignes : (a) image originale avec 15 lignes de texte ; (b) image segmentée avec extraction automatique de 14 lignes. Nous pouvons remarquer que les lignes 6 et 7 dans 'b' ne sont pas séparées, l'algorithme les comptabilise comme une seule ligne.

4. Evaluation du système d'extraction de figures :

Nous avons testé le système d'extraction de figures conçu sur 20 images de documents manuscrits, 15 d'entre eux contiennent des figures, et les 5 autres sont des textes manuscrits sans figures. Nous présentons les résultats d'évaluation dans le tableau suivant :

Documents sans figures	5
Nombre de fausses extractions	**2**

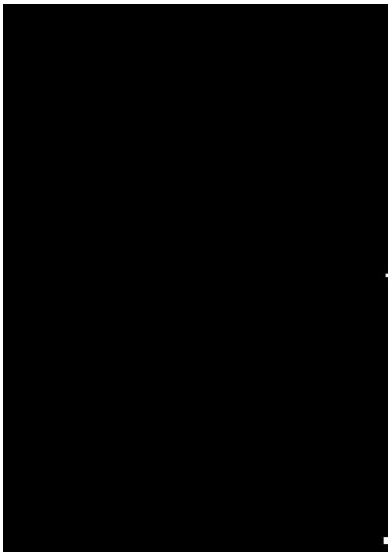
Tab. 5.3 : Test de fausse extraction sur des documents sans figures.

Nombre total de documents contenant des figures	15
Nombre de documents avec une extraction complète de figures	0
Nombre de documents avec une extraction partielle de figures	4
Nombre de documents avec une extraction incomplète de figures	4
Fausse extraction	7

Tab. 5.4 : Evaluation de la qualité d'extraction sur les documents contenant des figures.

D'après le tableau 5.3, le module d'extraction de figures peut s'avérer inefficace et extraire une figure qui n'existe pas dans un document ne contenant pas de figure. En analysant les images de documents sans figures, nous avons remarqué que le système d'extraction de figures extrait de fausses figures sur des images de petite dimension (voir Fig.5.2).

(a1)



(a2)

(b1)

(b2)

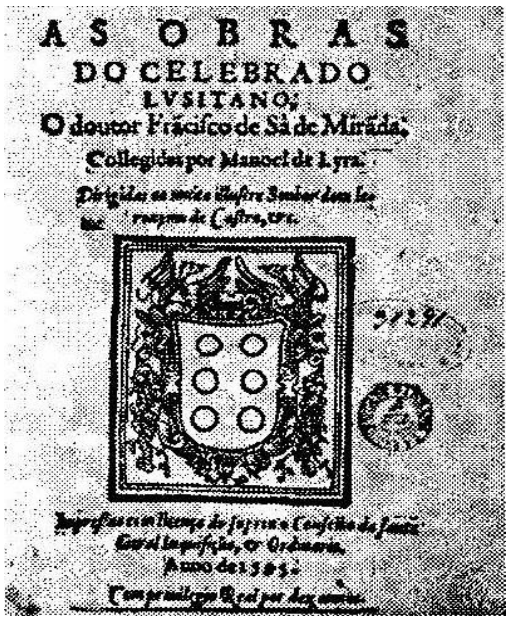
Fig. 5.2 : Résultat d'extraction de figures sur des images ne contenant aucune figure :

(a1) Image originale avec une grande dimension (1200x1792) ; (a2) résultat d'extraction de figure ; (b1) image originale avec une petite dimension(135x199) ; (b2) résultat de l'extraction de figure. C'est une fausse extraction. La dimension de l'image influe sur le résultat d'extraction.

D'après le tableau 5.4, le système ne peut pas extraire de figure entière. En analysant les images testés avec les résultats du tableau, nous pouvons déduire que l'extraction de figures échoue lorsque:

- La dimension de l'image est petite (voir Fig. 5.3);
- Les figures sont simples comme les cadres ou les lignes (voir Fig. 5.4).

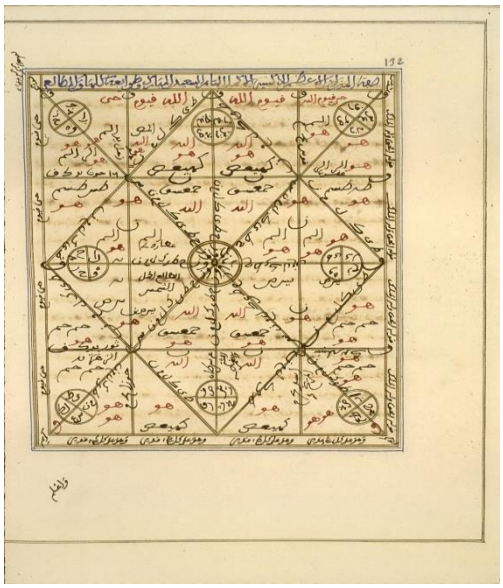
Par contre elle réussit partiellement à extraire certaines figures comme les *ornements*[10] ou les formes *kaléidoscopiques*[11] (voir Fig. 5.5).



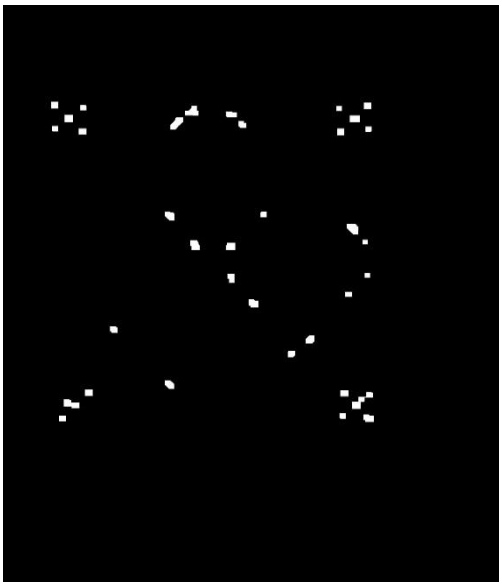
(a)

(b)

Fig. 5.3 : Une fausse extraction de figure sur une image de petite dimension : (a) image originale avec une dimension 400x536 ; (b) résultat d'extraction de figures.



(a)

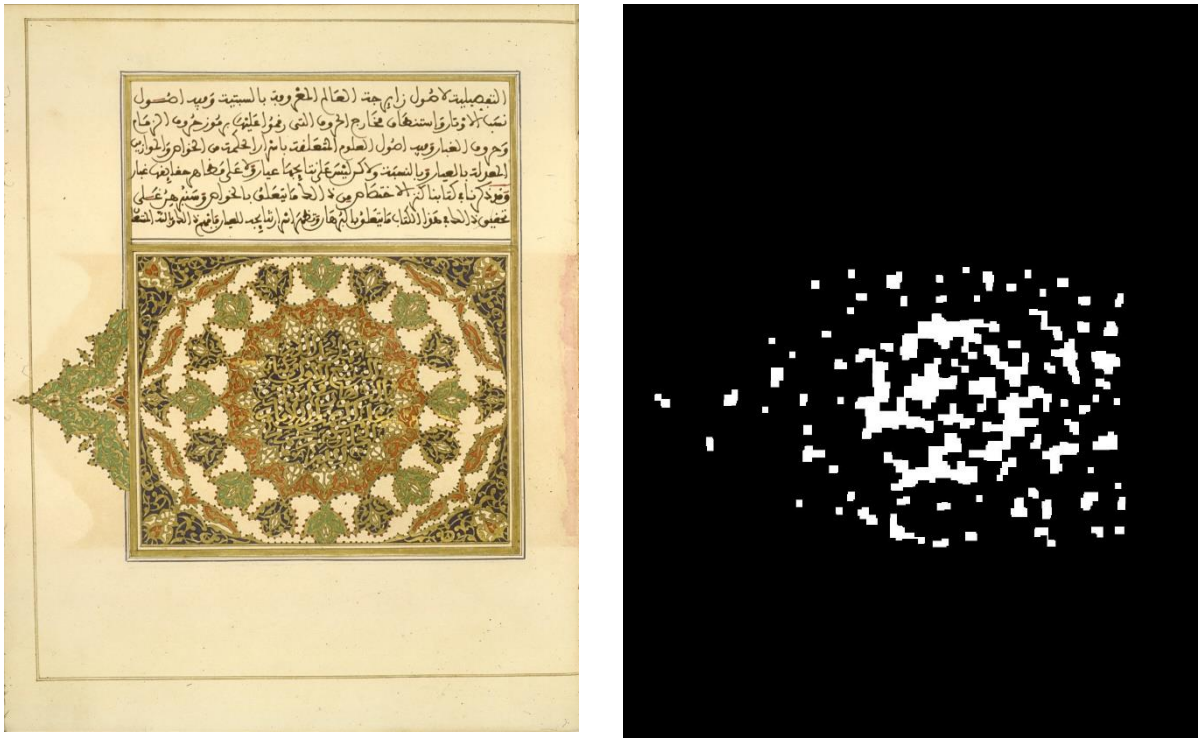


(b)

Fig. 5.4 : Une extraction incomplète de figure : (a) image originale ; (b) résultat d'extraction de figures.

[10]Figure de fantaisie

[11]Un kaléidoscope est un ensemble de formes et de couleurs variées donnant un effet de symétrie à la figure.



(a) (b)

Fig. 5.5 : Une extraction partielle de figure : (a) image originale ; (b) résultat d'extraction de figures.

5. Conclusion :

Nous avons exposé dans ce chapitre les résultats d'évaluation de notre système d'extraction du nombre de ligne et de figures. L'algorithme d'extraction du nombre de ligne ne donne pas des résultats concluant mais permet de donner un nombre très proche de la réalité avec taux d'erreur de 12%.

Le module donne de bons résultats en l'absence de bruit. Toutefois, l'algorithme d'extraction de figures échoue si l'image est de petite taille ou si les figures sont simples.

Par contre il extraie partiellement les ornements. En l'améliorant, il peut être utile pour la détection de présence de figures. Les méthodes utilisées pour le comptage du nombre de lignes et le test de la présence de figures dans le manuscrit ont montrés leurs limites dans certains cas particuliers, mais la technique d'extraction automatique de métadonnées à partir d'images de manuscrits numérisés semble être un outil prometteur, qu'il faudrait améliorer afin qu'il puisse apporter une assistance réelle dans le domaine de catalogage de manuscrits.

Conclusion générale et perspectives :

L’objet de notre travail est de pouvoir assister le catalogueur dans ses tâches répétitives, qui lui sont pénibles et fastidieuses telles que le comptage du nombre de lignes dans chaque pas du manuscrit, la détection de la présence d'objets autres que le texte dans la page d'un manuscrit, la détection de l'homogénéité du manuscrit et bien d'autres caractéristiques qui font appel plus à un aspect mécanique qu'intellectuel. Pour ce faire, nous avons fait appel à des outils de traitement d'image pour extraire automatiquement des métadonnées.

Dans le présent mémoire, nous nous sommes limités principalement, à l’extraction de deux métadonnées :

- Le nombre de lignes, et
- Les figures présentes dans une image de manuscrit ancien numérisé.

La mise en œuvre d'un tel objectif a fait appel pour des concepts de traitement d’images: les prétraitements (passage au niveau de gris, filtrage, rehaussement d'images, …etc.) et la segmentation (binarisation et RLSA).

Dans notre travail, nous ne sommes pas limités à des applications simples de concepts et d'outils de traitement d'images, mais nous avons pu adapter la méthode de segmentation RLSA, en effectuant des modifications, afin d'optimiser le processus d'extraction de lignes.

D'une manière générale, nous pouvons résumer notre tâche à deux processus majeurs, à savoir:

- Dans l’extraction du nombre de lignes, nous avons utilisé le filtrage linéaire pour améliorer la qualité de l’image en vue d’une bonne segmentation. Nous avons modifié la méthode du RLSA afin de lier les caractères d’une seule ligne entre eux, de séparer les lignes de texte et d’éliminer les chevauchements entre les lignes. L’extraction est réalisée grâce à l’histogramme de projection de l’image segmentée.
- Dans l’extraction de figures, nous avons utilisé le filtrage linéaire pour améliorer la qualité de l’image et transformer l’image en un nuage de points. Le filtrage non linéaire a eu pour rôle de fusionner les nuages les plus denses et effacer ceux qui le sont moins.

Après la mise en œuvre de ces algorithmes d’extraction, nous sommes parvenus à des résultats encourageants à savoir :

- L’extraction du nombre de lignes réussit dans le cas où les lignes sont bien alignées ou séparées entre elles. Le système ne comptabilise pas les réclames comme des lignes. Dans le cas contraire, plusieurs lignes peuvent être prises comme une seule ligne, ce qui constitue une limite de l'algorithme du RLSA modifié, car il ne réussit pas dans tous les cas à séparer les lignes entre elles.
- L’extraction de figures réussit partiellement lorsque l’image est de grande dimension et les figures sont de forme ornementale ou kaléidoscopique.

Pour conclure, nous pouvons affirmer, suite au modeste travail que nous avons effectué, que le traitement d'image reste un outil puissant et prometteur qui pourrait venir en aide d'une manière efficiente au catalogueur dans ses diverses tâches de catalogage quotidienne.

Perspectives :

Il serait intéressant de pouvoir améliorer les résultats actuels de notre travail. Pour ce faire, nous proposons quelques pistes qui nous semblent concluantes à explorer. Nous citons entre autre :

Extraction du nombre de lignes :

- Effectuer un décalage dans l'histogramme de projection horizontal, c'est-à-dire enlever la partie qui contient le moins de transitions entre la couleur blanche et noire. Cela a pour but d'éviter les contacts entre les blocs de l'histogramme ;
- Répéter l'opération de segmentation et étudier l'éventualité d'une minimisation de contacts entre les lignes.

Extraction de figures

- Faire une reconstruction à partir des résultats d'extraction, c'est-à-dire faire accroitre le volume des éléments extraits de l'image et effectuer « ET » logique entre ces éléments et l'image originale jusqu'à avoir toujours le même résultat.

Bibliographies:

[AGR 09] : M. Agrawal and D. S. Doermann. Voronoi++: A dynamic page segmentation approach based on voronoi and docstrum features. In ICDAR, 2009 (.in [MSVJ 12]).

[Ala 14] : Omar Alaql, « TEXT LINE EXTRACTION FOR HISTORICAL DOCUMENT IMAGES USING LOCAL CONNECTIVITY MAP », thèse de Master, Kent State University, Etats Unis d'Amérique, 2014.

[Ant 98] : A. Antonacopoulos. Page segmentation using the description of the background. Computer Vision and Image Understanding, 70(3) :350–369, 1998 (.in [Lel 07]).

[BAI 90] : H. S. Baird, S. E. jones, and S. J. Fortune. Image segmentation by shape-directed covers. In ICPR, 1990 (.in [MSVJ 12]).

[BEL 92] : Belaid A., Belaid Y., Reconnaissance de formes : méthodes et applications, Interéditions, 1992 (.in [Lik 03])

[Ber 10] : MaïtineBergounioux, « Quelques méthodes de filtrage en Traitement d'Image. Cours donnée dans le cadre d'une école CIMPA - en attente de publication dans les actes. 2010.

[Ber 15] : M. Bergounioux, « Débruitage par filtrage linéaire », Introduction au traitement mathématique des images - méthodes déterministes, Mathématiques et Applications, 2015.

[Ber 86] : J. Bernsen. Dynamic thresholding of grey-level images. In Proc. Eighth Int 'l Conf. on Pattern Recognition, pages 1251–1255, 1986 (.in [Lel 07]).

[BLA] : AbdourahmaneBaldé, Yves Lechevallier, Marie-Aude Aufaure, « Extraction de métadonnées sur les prototypes issus de la classification d'objets », INRIA Rocquencourt, France.

[Bot 00] : Bottou L., Haffner P., LeCun Y., Horward P., Vincent P., Riemers B., « DjVu : un système de compression d'images pour la distribution réticulaire de documents numérisés », Actes de CIFED'2000, Lyon, juillet 2000, p. 453-462.

[Bou] : F. Le Bourgeois, H. Emptoz, E. Trinh, F. Muge, C. Pinto et I. Granado, "Wp4.3-4 Numérisation, Traitement et Interprétation des Images de Documents Anciens ", Project DEBORA Telematics Applications Programme n° 5608 – (.in [Mou 06]).

[BTI] :Costin-Anton Boiangiu, Mihai Cristian Tanase, RaduIoanitescu, « TEXT LINE SEGMENTATION IN HANDWRITTEN DOCUMENTS BASED ON DYNAMIC WEIGHTS ».

[BTM] : Isabelle Bloch, Florence Tupin, Antoine Manzanera, « TERI : Traitement et reconnaissance d'images », Université PIERRE ET MARIE CURIE.

[CCG 07] : L. Caponetti, C. Castiello, and P. Gorecki. Document page segmentation using neurofuzzy approach. Applied Soft Computing, In Press, Corrected Proof : –, 2007 (.in [Lel 07]).

[DEB 00] : Debora, «Présentation du projet européen Debora, projet no LB 5608/A» R. Bouché (coordonnateur) document distribué lors de CIFED'2000, Colloque International Francophone sur l'Ecrit et le Document, Lyon, juillet 2000. (.in [Lik 03])

[DKS] : Markus Diem, Florian Kleber, Robert Sablatnig, « Text Line Detection for Heterogeneous Documents », Computer Vision Lab, Vienna University of Technology.

[FEI 09] :Fei Yin, Cheng-LinLiu. Handwritten Chinese text line segmentation by clustering with distance metric learning Pattern Recognition (2009) pp. 3146 – 3157 (.in [KPKB 12]).

[FEL 00] :Feldbach M., « Generierungeinersemantischenreprasentationausabbildungenhandschritlicherkirchenbuchaufzeichnungen, Diplomarbeit, Otto von Guericke Universitat Magdeburg, juillet 2000. (.in [Lik 03])

[GJG 07] : Maya R. Gupta, Nathaniel P. Jacobson, Eric K. Garcia, « OCR binarization and image pre-processing for searching historical documents », Pattern Recognition 40, pages 389 – 397, 2007.

[GKG 14] : Rahul Garg, Naresh Kumar Garg, « Problems and Review of Line Segmentation of Handwritten Text Document », International Journal of Advanced Research in Computer Science and Software Engineering, Volume 4, issue 4, pages 1036-1038 ; 2014.

[GPH 05] : U. Garain, T. Paquet, and L. Heutte. On foreground-background separation in low quality color document images. In ICDAR '05 : Proceedings of the Eighth International Conference on Document Analysis and Recognition, pages 585–589, 2005 (.in [Lel 07]).

[GPP 06] : B. Gatos, I. Pratikakis, and S. J. Perantonis. Adaptive degraded document image binarization. Pattern Recogn., 39(3) :317–327, 2006 (.in [Lel 07]).

[Gra 00] : Granado, I., Mengucci, M., Muge, F.: Extraction de textes et de figures dans les livres anciens à l'aide de la morphologie mathématique. In: Actes de CIFED'2000, Colloque International Francophone sur l'Ecrit et le Document, Lyon, pp. 81–90, 2000.

[HP 72] : S. L. Horowitz and T. Pavlidis. Picture segmentation by a traversal algorithm. Comput. Graphics Image Process., 1 :360–372, 1972 (.in [Lel 07]).

[ISPS 11] : « Image Segmentation Using Continuous Cellular Automata », 2011.

[Jou 06] : N. Journet. Analyse d'images de documents anciens : Catégorisation de contenus par approche texture. PhD thesis, Université de La Rochelle, 2006 (.in [Lel 07]).

[Jou 09] : Guillaume Joutel, « Analyse multirésolution des images de documents manuscrits, Application à l'analyse de l'écriture », thèse de doctorat, Institut National des Sciences Appliquées de Lyon, France, 2009.

[Kie 14] : Van CuongKieu, « Modèle de dégradation d'images de documents anciens pour la génération de données semi-synthétiques », Traitement des images, Université de La Rochelle, France, 2014.

[Kim 96] : H.K. Kim. Efficient automatic text location method and content-based indexing and structuring of video database. Journal of Visual Communication and Image Representation, 7(4) :336–344, 1996 (.in [Lel 07]).

[KJP02] : I.-K. Kim, D.-W. Jung, and R.-H. Park. Document image binarization based on topographic analysis using a water flow model. Pattern Recognition, 35 :265–277, 2002(.in [Lel 07]).

[KPKB 12] : M.Ravi Kumar, R. Pradeep, B.S.Puneeth Kumar, Prasad Babu, « A Simple Text-line segmentation Method for Handwritten Documents », International Journal of Computer Applications (0975 – 8878) on National Conference on Advanced Computing and Communications – NCACC, page 46-51, 2012.

[LAM 96] : Lamouche I., Bellissant C., «Séparation recto/verso d'images de manuscrits anciens », Actes de CNED'96, Colloque National sur l'Ecrit et le Document, Nantes, juillet 1996, pp. 199- 206. (.in [Lik 03])

[LE 00] : R. Lienhart and W. Effelsberg. Automatic text segmentation and text recognition for video indexing. MultimediaSystems, 8 :69–81, 2000 (.in [Lel 07]).

[Lel 07] : Thibault LELORE, « Segmentation d'image, Application aux documents anciens », Mémoire de Master de recherche, LABORATOIRE DES SCIENCES DE L'INFORMATION ET DES SYSTÈMES, Université de Nantes, France, 2007.

[Lik 03] : Laurence Likforman-Sulem, « Apport du traitement des images à la numérisation des documents manuscrits anciens », Ecole Nationale Supérieure des Télécommunications, 2003.

[LIN 94] :Lins R.D., GuimaraesNeto M., FrançaNeto L., Galdino Rosa L., «An Environment for Processing Images of Historical Documents », Microprocessing and Microprogramming, 40 (1994), pp. 939-942. (.in [Lik 03])

[LOU 08] : G. Louloudisa, B.Gatosb, I.Pratikakisb, C.Halatsisa, Text line and word segmentation of handwritten documents. Pattern Recognition (2008) pp. 3169 – 3183 (.in [KPKB 12])

[MCT 09] : DharitriMisra, Siyuan Chen, George R. Thoma, « A System for Automated Extraction of Metadata from Scanned Documents using Layout Recognition and String Pattern Search Models », National Library of Medicine, 1509STP: 107–112, Bethesda, Maryland, Etats Unis d'Amérique, 2009.

[MEN 00] : Mengucci M., Granado I., «Morphological Segmentation of text and figures in Renaissance books (XVI Century) », in Mathematical Morphology and its applications to image processing, J. Goutsias, L. Vincent, D. Bloomberg (eds), Kluwer, 2000, pp. 397-404. (.in [Lik 03])

[Mou 06] : Kamel MOUATS, « Segmentation d'Images de Documents Anciens par Approche Texture - APPLICATION du filtre de Gabor », Mémoire de Master, LABORATOIRE L3I – INFORMATIQUE IMAGE INTERACTION, Université de La Rochelle, France, 2006.

[MSVJ 12] :Anand Mishra, Naveen Sankaran, VireshRanjan, C. V. Jawahar, « Automatic Localization and Correction of Line Segmentation Errors », Center for Visual Information Technology, IIIT Hyderabad, India, 2012.

[NAG 92] : G. Nagy, S. C. Seth, and M. Viswanathan. A prototype document image analysis system for technical journals. IEEE Computer, 1992 (.in [MSVJ 12]).

[Nib 86] : W. Niblack. An introduction to digital image processing. Prentice Hall (July 1986), 1986 (.in [Lel 07]).

[NISO 04] : National Information Standards Organization, « Understanding Metadata », Bethesda, Maryland, Etats Unis d'Amérique2004.

[NPH 06] : S. Nicolas, T. Paquet, and L. Heutte. Extraction de la structure de documents manuscrits complexes à l'aide de champs markoviens. In Actes du 9ème Colloque International Francophone sur l'Ecrit et le Document, pages 13–18, 2006 (.in [Lel 07]).

[NS 84] : G. Nagy and S. Seth. Hierarchical representation of optically scanned documents. International conference on Pattern Recognition, 7 :347–349, 1984 (.in [Lel 07]).

[NS] :MahuaNandy (Pal), SatadalSaha, « An Analytical Study of different Document Image Binarization Methods », IEEE National Conference on Computing and Communication Systems (COCOSYS-09) CS24, page 71-76.

[OGK 09] : Lawrence O'Gorman, RangacharKasturi, « Document Image Analysis », IEEE Computer Society Executive Briefings, 2009.

[Ort 04] : Mathias Ortner, « Algorithmes pour le traitement de l'image », 2004.

[Ots 79] : N. Otsu, A threshold selection method from grey scale histogram, IEEE Trans. On SMC, Vol. 1, pp. 62-66, 1979.

[Ouw 10] : Nazih OUWAYED, « Segmentation en lignes de documents anciens : application aux documents arabes », thèse de doctorat, Ecole doctorale IAEM Lorraine, Université Nancy 2, France, 2010.

[PYR 15] :RuggeroPintus, Ying Yang, Holly Rushmeier, « ATHENA: Automatic text height extraction for the analysis of text lines in old handwritten manuscripts », ACM J. Comput. Cult. Herit. 8, 1, Article 1, 25 pages, 2015.

[SKHS 98] : T. Sato, T. Kanade, E. K. Hughes, and M. A. Smith. Video ocr for digital news archive. In International Workshop on Content-Based Access of Image and Video Databases (CAIVD '98), pages 52–60, 1998 (.in [Lel 07]).

[SP 00] : J. Sauvola and M. Pietikäinen. Adaptive document image binarization. Pattern Recognition, 33(2) :225 – 236, 2000 (.in [Lel 07]).

[TAN 02] : Tan C. L., Cao R., Shen P., « Restoration of archival documents using a wavelet technique », IEEE PAMI, Vol 24, no 10, octobre 2002, pp. 1399-1404. (.in [Lik 03])

[Ton 09] : Emma Tonkin, « MetRe: Supporting the Metadata Revision Process », Bath, RoyaumeUni, 2009.

[Ven] :F.Venturelli, "A Successful Technique For Unconstrained Hand-Written Line Segmentation," Progress in Handwriting Recognition, pp. 563-568.

[WCW 82] : K.Y. Wong, R.G. Casey, and F.M. Wahl. Document analysis system. IBM Journal of Research and Development, 26(6) :647–656, 1982 (.in [Lel 07]).

[WD02] : C. Wolf and D. Doermann. Binarization of low quality text using a markov random field model. Proceedings of the 16 th International Conference on Pattern Recognition (ICPR'02), 3 :30160, 2002 (.in [Lel 07]).

[WM 98] :V.Wu and R. Manmatha. Document image clean-up and binarization. Proceedings of IS&T/SPIE Symposium on Electronic Imaging, 3305 :263–273, 1998. (.in [Lel 07]).

[WMR 99] :V.Wu, R. Manmatha, and E. M. Riseman. Textfinder : An automatic system to detect and recognize text in images. IEEE Transactions on Pattern Analysis and Machine Intelligence, 21(11) :1224–1229, 1999 (.in [Lel 07]).

[WON 82] : K. Y. Wong, R. G. Casey, and F. M. Wahl. Document analysis system. IBM Journal of research and development, 1982 (.in [MSVJ 12]).

[WPH 06] :Y.Wang, I. T. Phillips, and R. M. Haralick. Document zone content classification and its performance evaluation. Pattern Recognition, 39(1) :57–73, 2006 (.in [Lel 07]).

[ZWU 11] : Yudong ZHANG , Lenan WU, « Fast Document Image Binarization Based on an Improved Adaptive Otsu's Method and Destination Word Accumulation », Journal of Computational Information Systems 7: 6, pages 1886-1892, 2011.

[ZZJ 00] : Y. Zhong, H. Zhang, and A. K. Jain. Automatic caption localisation in compressed video. IEEE Trans. Pattern Anal. Mach. Intell., 22(4) :385 – 392, 2000 (.in [Lel 07]).

Webographie:

[Dis] : Jean-Marc Dissaux, Un exemple de contrat d'arrentement, (En ligne) archivespasdecalais.fr, Les archives du Pas-de-Calais, consulté le 28/07/2016, disponible sur internet :

http://www.archivespasdecalais.fr/Activites-culturelles/Un-document-a-l-honneur/Un-exemple-de-contrat-d-arrentement

[Dou] : Jean-Michel DOUDOUX., Développons en Java, Présentation de Java. (En ligne) jmdoudoux.fr. Jean-Michel DOUDOUX, consulté le 26 juin 2014, disponible sur internet:

http://www.jmdoudoux.fr/java/dej/chap-presentation.htm

[FPW 04] : R. Fisher, S. Perkins, A. Walker, E. Wolfart, « Dilatation », Image Processing Learning Resources, (En ligne) inf.ed.ac.uk, 2004, consulté le 22/08/2016, disponible sur internet :

http://homepages.inf.ed.ac.uk/rbf/HIPR2/dilate.htm

[IMMa 04] : Islamic Medical Manuscripts at the National Library of Medicine, Catalogue: Dietetics and Regimen - Gallery, (Enligne) nlm.nih.gov, U.S. National Library of Medicine, 2004, consulté le 28/07/2016, disponiblesur internet :

https://www.nlm.nih.gov/hmd/arabic/diet_gallery.html

[IMMb 04] : Islamic Medical Manuscripts at the National Library of Medicine, Catalogue: Alchemy - Gallery, (Enligne) nlm.nih.gov, U.S. National Library of Medicine, 2004, consulté le 28/07/2016, disponiblesur internet :

https://www.nlm.nih.gov/hmd/arabic/alchemy_gallery.html

[Jan 11] : Nicolas JANEY, « Le traitement d'images », Infographie, (En ligne) univ-fcomte.fr, UFR Sciences et Techniques, Université de Besançon, 2011, consulté le 22/08/2016, disponible sur internet :

http://raphaello.univ-fcomte.fr/IG/TraitementImages/TraitementImages.htm

[Lar] : Larousse, Dictionnaire de français, annotation – définition, (En ligne) larousse.fr, Larousse, consulté le 29/07/2016, disponible sur internet :

http://www.larousse.fr/dictionnaires/francais/annotation/3676

[NJ 04] :Anoop M Namboodiri, Anil K Jain, « Online handwritten script recognition », (En ligne) researchgate.net, ResearchGate, 2004, consulté le 28/07/2016, disponible sur internet :

https://www.researchgate.net/figure/8331691_fig1_Fig-1-A-multiscript-online-document-containing-Cyrillic-Hebrew-Roman-Arabic