République Algérienne Démocratique et Populaire
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique
Université M'hamed Bougara - Boumerdès

Faculté des Sciences
Département d'Informatique

Domaine : Mathématiques Informatique
Filière : Informatique
Spécialité : Ingénierie des logiciels et Traitement de L'information

Mémoire de fin d'études en vu de l'obtention du Diplôme de master académique

Thème

TRAITEMENT ET RECONNAISSANCE DES CARACTERES

Présenté par :

Baka Abdeladim
Fillali Hicham

Stage Pratique réalisé au ***:*** *boite de développement «* Protid Systems ».
Sous l'encadrement de Mr : *Mr Mirad Sofian.*

Soutenu le 21/06/2016. Devant le jury composé de

Mme	Hadjidj Drifa	: Président
Mr	Gaceb Djamel	: Membre
Mr	Mohamed Tahar Bennai	: Promoteur

Promotion: 2015-2016

Table des matières

Table des figures

Remerciement

Nous remercions piètrement Allah tous puissant pour la volonté, la santé, et la puissance qu'il nous a donné durant toute ces années.

Nous tenons à exprimer nos remerciements à notre directeur de projet Monsieur Mirad Sofiane pour avoir dirigé ce travail, pour nous avoir soutenu

tous au long de notre projet et pour ces précieux conseils.

Nous tenons à adresser nos plus vifs et plus sincères remerciement à notre promoteur Monsieur Bennai Mohamed pour son aide et son suivi permanent.

On remercie également les membres de jury pour avoir accepté d'évaluer ce travail : Monsieur Gaceb Djamel et Madame Hadjidj .

En, à tous nos amis qui nous ont apporté un soutient moral durant cette année.

Dédicaces

Je dédie ce modeste travail à mes chers parents qui m'ont soutenu et encouragé tout au long de mes études :

à ma soeur. à ma grande famille, oncles et tantes. à tous mes cousins et mes cousines .

à tous mes chers amis et mon binôme de ce mémoire FILLALI HICHAM.

à tous ceux que j'aime et qui m'aiment sans exception.

Abdeladim

Dédicaces

Je dédie ce modeste travail à mes chers parents qui m'ont soutenu et encouragé tout au long de mes études :

à mes frères et ma soeur. à ma grande famille, oncles et tantes. à tous mes cousins et mes cousines.

à tous mes chers amis et mon binôme de ce mémoire Baka Abdeladim.

à tous ceux que j'aime et qui m'aiment sans exception.

Hicham.

Résumé

Un système de reconnaissance optique des caractères analyse optiquement un texte et en produit une version informatique, sous forme d'un fichier texte, comme s'il avait été saisi sur un ordinateur. On utilise également l'acronyme OCR du terme anglais Optical Character Recognition. L'OCR est évidemment une technique utile, mais il faut en connaître les limites et en tenir compte, en prévoyant une ou plusieurs lectures personnelles du document. Parfois les documents à traiter peuvent être dégradé physiquement ou lors de leurs acquisition pour cela l'étape de prétraitement est donc indispensable afin de rendre fiable l'étape de conversion de l'image vers un texte, Les filtres linéaires pour le traitement du bruit, tel que le filtre gaussien, moyen ...etc. permettent de lisser l'image et ainsi diminuer le bruit qui pourrait impacter négativement sur le résultat de reconnaissance sans garantir la conversion des contours. Pour améliorer les résultats de l'OCR, nous allons utiliser quelques prétraitements pour améliorer la performance de l'OCR.

Mots-Clés : Prétraitement des images, Binarisation, Filtres, OCR, Tesseract.

Introduction générale

L'écriture dans ses différentes formes, imprimée et manuscrite a toujours été un outil essentiel dans la communication humaine, elle est aussi présente dans la majorité des secteurs et des activités. Elle est utilisée pour conserver et archiver le savoir. De ce fait, l'homme a toujours développé des techniques visant sa pérennité à travers les générations. En effet, avec l'apparition des nouvelles technologies d'information : l'électronique et l'informatique, et l'augmentation de la puissance des machines, l'automatisation des traitements (la lecture, la recherche et l'archivage...) apparaît incontournable.

Ce besoin d'automatisation a donnée naissance au domaine de la reconnaissance automatique d'écriture. C'est ainsi que les recherches concernant cette discipline ont été lancées. Depuis le début des années 2000, la numérisation s'accompagne d'un processus de conversion en mode texte, dit OCR pour Optical Character Recognition. Pourquoi faire cette conversion ?

Elle permet avant tout l'accès au contenu "plein texte", c'est-à-dire au lien qu'il existe entre un mot et la liste des documents qui le contiennent. En ce sens, l'OCR-isation" des documents numérisés permet l'indexation des documents "par le contenu". La conversion peut aussi être utile dans d'autres contextes comme la production d'Ebooks, la citation des parties de texte, la généalogie, la poétisation. Plusieurs facteurs agissent sur la qualité finale des résultats de conversion de l'OCR :

- Les caractéristiques de l'œuvre (contenu textuel, illustrations, présence de formules mathématiques,...) et de son édition (éditeur, date d'édition).
- Les caractéristiques du papier, de l'impression, de la qualité de conservation, de l'encre, de l'encrage, de la fonte.
- Les caractéristiques de la numérisation (qualité du scan, paramètres de numériseur) et de l'image numérique.

Le prétraitement automatique des images permet de manipuler une image dans l'objectif d'améliorer sa qualité ou de le conditionner pour la phase d'extraction de ses informations avec une méthode efficace. Pour notre projet de master, nous somme dirigé vers la société « Protid système » pour la première expérience professionnelle. Nous allons donc aborder dans le premier chapitre une présentation de notre organisme d'accueil. Le thème proposé

par ce dernier fut donc la réalisation d'une application fonctionnelle pour la reconnaissance de caractère utilisant Tesseract_ocr.

Pour mener à bien ce projet, nous avons d'abord effectué une étude bibliographique sur l'imagerie numérique et ses différents traitement que nous présentant dans le chapitre deux. En suite nous abordons la notion de système de reconnaissance de caractère ainsi que ses domaines d'application que nous détaillerons dans le chapitre trois. Dans le chapitre quatre nous allons présenter un ensemble de prés traitement utiliser dans l'imagerie pour améliorer la qualité des images. Cela dans le but d'augmenter les performances de notre future application. Une fois les connaissances théorique acquises, nous nous somme concentrés sur la conception de notre application, nous présentons les résultats de cette conception dans le cinquième chapitre. Le sixième chapitre se focalisera sur l'implémentation de notre application et les résultats obtenus sur quelques images tests avant de conclure par une conclusion générale.

Chapitre 1

Présentation de l'organisme d'accueil

1.1 Introduction

Dans ce chapitre nous allons présenter le contexte du projet à savoir l'organisme d'accueil ainsi que le projet du stage de fin d'études.Nous détaillerons aussi, les objectifs généraux du projet et méthodologie suivie pour la réalisation de ces objectifs.

1.2 Profil de la société Protid Systems

Créée fin 2007, Protid Systems se positionne comme un acteur novateur sur le marché Algérien.

Elle est l'un des premiers prestataires de services spécialisés dans l'infogérance, le développement, l'intégration des applications de gestion et les services sur mesure, qui permettent d'apporter de la valeur ajoutée aux systèmes d'informations de ses clients.

Son métier est d'accompagner ses clients sur des missions stratégiques et d'apporter des solutions efficaces et rentables adaptées à leurs problématiques en prenant en charge le projet depuis la définition des besoins et de l'architecture technique jusqu'à la mise en production et la maintenance.

Protid Systems déploie ses compétences couvrant les principales solutions de gestion (GRH, ERP, GPAO, GMAO, CRM ...), afin de répondre aux exigences métiers des grandes administrations, télécoms, tourisme, industrie et permettant une gestion efficiente de l'information pour toute l'organisation de l'entreprise.

Sa capacité unique de concevoir des applications n-tiers modulables, flexibles (multi-SGBD, Multi-langues et Multi-plate-formes) lui a permis de participer à la réalisation de plusieurs grands projets tels que MARA (Modernisation et Assistance aux Réformes Administrative) pour le Ministère des finances et AMECO (Appui au Management de l'Économie) pour le compte du ministère des travaux publics.

FIGURE 1.1 – Logo de l'organisme d'accueil "Protid systems"

1.3 Avantages des services de Protid Systems

- Ses solution sont multi-langue, Multiplateformes et multi-SGBD.
- Ses solutions sont conformes à la réglementation locale.
- Avec la flexibilité de ses solution, ses client adaptent leurs organisations ou leur processus métiers rapidement et à moindre cout pour rester compétitifs.

1.4 Les activités de l'organisme

- Dévelopement des applications métiers.
- Gestion électronique des documents.
- Dématérialisation des processus métiers.
- Industrialisation du développement .
- Solution d'infrastructures avancées.

1.5 Conclusion

Dans ce chapitre nous avons présenté l'organisme d'accueille avec ses avantages et ses activités, dans le chapitre suivant nous allons présenter une bref présentation sur les images ainsi que les leurs structure et les caractéristiques en générale parmi lesquelles le bruit, pixel, dimension. . .etc.

Chapitre 2

Généralité sur les images

2.1 Introduction

Le stockage physique des documents de plus en plus problématique car ce stockage consomme de l'espace physique (locaux) il nécessite un certain nombre de conditions de conservation (humidité, température) qui peuvent être très couteuse. Les organisation et sociétés se tournent de plus en plus vers le stockage numérique des documents qui est entre autre moins coûteu que son équivalent physique. Pour cela les documents doivent être numérisé et enregistré sous forme d'image numérique, ces images seront ensuite traité par des OCR pour extraire les informations qu'il contiennent. Notre projet traite de différentes notions liées à l'imagerie numérique. Nous allons donc détailler certaines de ces notions dans le chapitre ci-dessous.

2.2 L'image numérique

L'image numérique est une représentation à deux dimensions d'une scène en trois dimensions, divisée en éléments de tailles fixes appelés cellules ou pixels, ayant chacun comme caractéristique un niveau de gris ou de couleur prélevé à l'emplacement correspondant dans l'image réelle, ou calculé à partir d'une description interne de la scène représenter .

La numérisation d'une image est la conversion de celle-ci de son état analogique (distribution continue d'intensités lumineuses dans un plan x * y) en une image numérique représentée par une matrice bidimensionnelle de valeurs numériques f(x, y) où x, y : coordonnées cartésiennes d'un point de l'image.

L'amplitude de f pour chaque paire de coordonnées (x, y) est appelée intensité de l'image au point donné. [1]

2.3 Caractéristique d'une image numérique

L'image est un ensemble structuré d'informations caractérisé par les paramètres suivants:

2.3.1 Pixel

Une image numérique est composée d'une grille de pixels. En appliquant un zoom sur une image affichée à l'écran, ces pixels apparaissent comme autant de petits carrés, porteurs d'une information de couleur élémentaire. le pixel est le plus petit élément de l'image c'est une entité indivisible qui peut recevoir une structure et une quantification si le bit est la plus petites unité d'information qui peut traiter un ordinateur, le pixel est le plus petit élément que peuvent manipuler les matériel et logiciel d'affichage ou d'impression, la quantité d'information que véhicule chaque pixel donne les nuances entre image en niveau de gris et l'image couleur. Dans une image en niveaux de gris, chaque pixel est codé sur un octet. [2]

Dans une image couleur (RVB) un pixel est codé sur trois octets Un octet pour chacune des couleurs : (R) Rouge, (V) Vert, (B) Bleu.

La lettre A, par exemple, peut être affichée comme un groupe de pixels dans la figure ci-dessous :

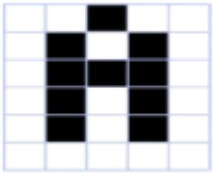


Figure 2.1: Représentation de la lettre A sous la forme d'un groupe de pixels

2.3.2 Dimension

On appelle dimension, le nombre de points (pixel) constituant l'image, c'est à dire sa (dimension informatique). Cette dernière se présente sous forme de matrice dont les éléments sont des valeurs numériques représentatives des intensités lumineuses (pixels). Le nombre de lignes de cette matrice multiplié par le nombre de colonnes nous donne le nombre total de pixels dans une image. [2]

2.3.3 Résolution "la finesse de l'image"

La résolution d'une image composée de points est définie par la densité des points par unité de surface, (1 pouce = 2.54 cm). Une image de 10 ppp (ou 10 dpi) contient 100

points par pouce carré. (10 x 10 = 100). La résolution permet de définir la finesse de l'image. Plus la rèsolution est grande, plus la finesse de l'image est grande. Les points d'une image ont différents noms dépendant du média. Sur les écrans on parle de pixel, les médias imprimés parlent de points où dots. Par conséquent la résolution dans le domaine de l'écran est ppi - pixels per inch (PPP en français : pixels par pouce). La résolution dans le domaine des médias imprimés est dpi (dots per inch). [3]

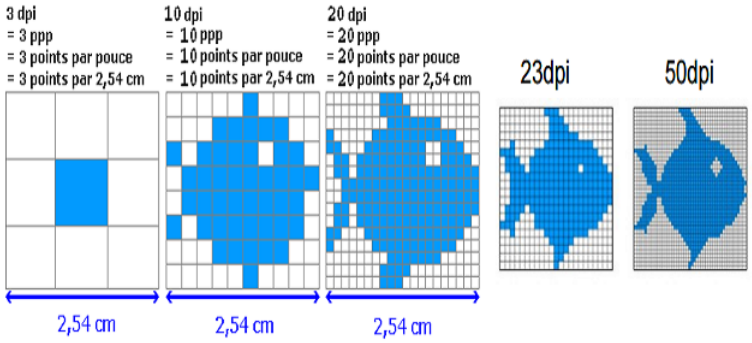


FIGURE 2.2 – Exemple 1 de Resolution d'une image

Figure 2.3: Exemple 2 de Resolution d'une image

2.3.4 Luminance

C'est le degré de luminosité des points de l'image. Elle est définie aussi comme étant le quotient de l'intensité lumineuse d'une surface par l'aire apparente de cette surface, pour un observateur lointain, le mot luminance est substitué au mot brillance, qui correspond à l'éclat d'un objet. Le contraste [6]

2.3.5 Le contraste

Est une propriété intrinsèque d'une image qui désigne et quantifie la différence entre les parties claires et foncées d'une image (elle différencie les couleurs claires des couleurs foncées).

Le contraste est défini en fonction des luminances de deux zones d'images. Si L1 et L2 sont les degrés de luminosité respectivement de deux zones voisines A1et A2 d'une image, le contraste C est défini par le rapport :

C=L1-L2/L1 +L2.[4]

2.4 Type des images

Il existe deux grandes familles d'images numériques matricielles (qui nous intéressent dans le cadre de notre étude) et vectorielles. [5]

2.4.1 Images matricielles (ou images bitmap)

Les images Matricielles (ou image en mode point, en anglais « bitmap ») sont celles que nous utilisons généralement pour restituer des photos numériques. Elles reposent sur une grille de plusieurs pixels formant une image avec une définition bien précise. Lorsqu'on les agrandi trop, on perd de la qualité. Les différents formats les plus répandus associés à ce type d'images sont : BMP, GIF, JPEG, TIFF, PNG... il existe deux type de format bitmap compressés (PNG,JPG) et non compressés (BMP,TIFF).

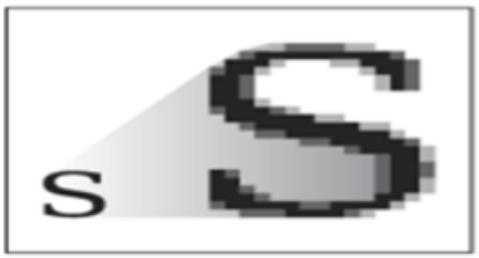


FIGURE 2.4 – Exemple d'une image matricielle

2.4.2 Images vectorielles

Une image vectorielle (ou image en mode trait), en informatique, est une image numérique composée d'objets géométriques individuels (segments de droite, polygones, arcs de cercle, ...etc.) définis chacun par divers attributs de forme, de position, de couleur, ...etc. Elle se différencie de cette manière des images matricielles ou « bitmap », dans

lesquelles on travaille sur des pixels Le principe est de représenter les données de l'image par des formules géométriques qui vont pouvoir être décrites d'un point de vue mathématique. Cela signifie qu'au lieu de mémoriser une mosaïque de points élémentaires, on stocke la succession d'opérations conduisant au tracé. L'avantage de ce type d'image est la possibilité de l'agrandir indéfiniment sans perdre la qualité initiale. [5].

Figure 2.5: Exemple d'une image vectorielle

2.5 Les differents modes Colorimetriques

Il existe différentes catégories d'image selon le nombre de bit Sur lequel est codée la valeur de chaque pixel.[5]

2.5.1 L'image monochrome

Le mode monochrome est le plus simple, chaque pixel y est soit allumé [Blanc], soit éteint [Noir] l'image obtenue n'est pas très nuancée. Alors, pour convertir une image couleur en mode monochrome il faut d' abord passer par le mode niveaux de gris.

Figure 2.6: L'image en mode monochrome

2.5.2 L'image en niveaux du gris

Une image en niveaux de gris est une image dans laquelle on trouve différents nuances de gris et dans chaque élément a une intensité qui varie de 0(noire) à 255(blanc).

Comme le montre la figure suivante :

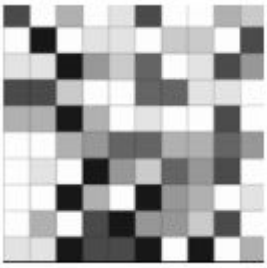


FIGURE 2.7 – Pixels d'une image en niveaux de gris.

Une image en niveaux de gris normale a une profondeur de couleur de 8bits=256 niveaux de gris. La figure suivante illustre les valeurs numérique de la luminance de quatre niveaux du gris :

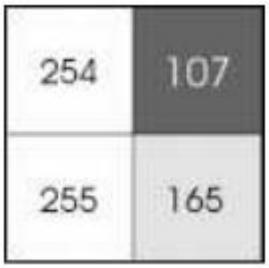


FIGURE 2.8 – Valeurs numériques des niveaux de gris des pixels d'une image.

Certaines images en niveaux de gris ont plus de niveaux de gris, par exemple pour l'image a 16 bits=65536 niveaux. Mais la plupart des systèmes de traitement automatique des images travaillent sur des images de 8 bits pour des raisons de vitesse de traitement.

2.5.3 L'image en couleurs

Même s'il est parfois utile de pouvoir représenter des images en noir et blanc ou en niveau de gris, les applications multimédias utilisent le plus souvent des images en couleurs. La représentation des couleurs s'effectue de la même manière que les images monochromes avec cependant quelques particularités. En effet, il faut tout d'abord choisir un modelé de représentation.

Pour cela on utilise un espace de couleur à plusieurs dimensions qui consiste à donner suffisamment de composantes numériques pour décrire une couleur. Il y a des différentes représentations des images couleur :

- La représentation en couleurs réelle sur 24 bits.
- La représentation en couleurs indexées, on utilise une table appelée palette pour éviter la redondance de couleur.
- Le mode RGB est idéal pour l'affichage sur écran, une image RGB est composée de trois couches Rouge, Vert et Bleu. Chaque pixel est définit par une valeur possible de ces couleurs de [0 à 255]. Une fois combinée ces couches permettent de générer toutes nuances de couleur. Le mode RGB correspond à l'affichage des moniteurs, où chaque point affiché est décomposé d'un mélange de lumière RGB.[6]

2.6 Définition d'une région :

La notion de région dans le traitement d'images, comme évoquée ci-dessus, est de regrouper des zones possédant les mêmes caractéristiques . C'est-adire que si plusieurs pixels adjacents s'avèrent être de couleur identique alors la zone qu'ils forment est une région. Ci-après, une illustration montrant deux régions de pixels différentes :

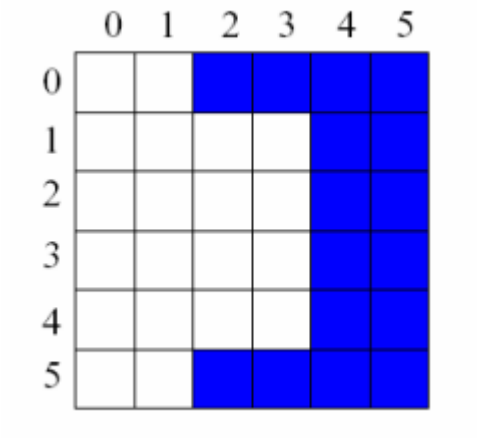


Figure 2.9: Différentes régions d'une image

2.7 Contours

Les contours représentent la frontière entre les objets de l'image, ou la limite entre deux pixels dont les niveaux de gris représentent une différence significative [4].

2.8 Bruit

C'est un signal qui lors de l'acquisition ou la transmission Vient s'ajouter à l'image, Il se matérialise par la présence dans une région homogène des valeurs plus ou moins éloignées de l'intensité de la région. Le bruit est le résultat de certains défauts électroniques du capteur et de la qualité de numérisation.

Figure 2.10: Image avec et sans bruit

Les sources de bruit d'une image sont nombreuses et diverses :
- bruits liés aux conditions de prise de vue (bougé, éclairage de la scène).
- bruits liés aux capteurs (appareil numérique de bas de gamme).
- bruits liés à l'échantillonnage.
- bruits liés à la nature de la scène (poussières, rayures).

2.8.1 Type de Bruit

Le bruit peut être catégorisé selon l'eet qu'il a sur l'image :

le bruit gaussien a un eet dégradant sous forme de grains, le bruit sel et poivre qui apparait généralement dans les images en niveaux de gris (impulsionnel), ce dernier est représenté par des points noir (poivre) et des points blanc (sel) il est généré par des erreurs de conversion ou bien des erreurs de transmission de bit (voir la figure suivante)

FIGURE 2.11 – Type de bruit

2.9 Conclusion

Nous avons donc présenté dans ce chapitre les différents notions de bases concernant les images numériques parmi lesquelles les pixels contours, luminance...etc. Ces notions seront essentielles à la compréhension des chapitres suivants qui traitent du processus automatique de la reconnaissance de caractère dans les l'image numérique.

Chapitre 3

Systèmes de reconnaissance de caractères OCR

3.1 Introduction

Toute information écrite peut être reprise dans une chaine de traitement informatisée à différentes fins : la rédaction et l'édition de rapports, la diffusion de documents dans un système de messagerie ... conduisent à exploiter des informations disponibles seulement sur papier.

La reconnaissance optique des caractères (OCR) est une opération informatique rapide permettant de réaliser la transformation d'un texte écrit sur papier en un texte sous forme d'un fichier informatique en représentation symbolique (par exemple pour les écritures latines, le codage opéré est le code ASCII (American Standard code for information interchange), tandis que pour l'arabe on utilise généralement le code ASMO (Arabic Standard Metrology Organization)

Dans ce chapitre nous allons présenter ce qu' est un OCR ainsi que le processus utilisé et ses domaines d'application de la reconnaissance de caractère .

3.2 Définition

La reconnaissance optique de caractères est un procédé permettant de récupérer les symboles dans les images de textes numérisées. Les images consistent en des matrices de pixels. La tâche d'un OCR consiste à segmenter les images , mots, caractères puis effectuer la reconnaissance des caractères. Il résulte une transcription textuele de l'image par laquelle des traitements automatiques sont possibles : recherche de mots, de noms, résumé. En général, l'OCR concerne le traitement d'un document numérisé.[7]

3.3 Domaine d'application

La technologie de reconnaissance optique des caractères, largement utilisée dans des applications commerciales depuis les années 1970, permet aujourd'hui d'automatiser des tâches telles que le traitement de passeports, le traitement sécurisé de documents (chèques, documents financiers, factures), le suivi postal, l'édition, le conditionnement de biens de consommation (codes de lots, codes de paquets, dates d'expiration) et des applications cliniques. Des lecteurs et logiciels OCR peuvent être utilisés, ainsi que des caméras intelligentes et des systèmes de vision industrielle offrant des capacités supplémentaires comme la lecture des codes à barres et l'inspection de produits.

Les principales applications déduites de la lecture de documents par des machines sont :

- L'aide à la lecture pour les non-voyants : Les systèmes de reconnaissance associés à des synthétiseurs vocaux permettent la compréhension de documents et livres pour les aveugles.
- La saisie automatique de document : La reconnaissance de caractères permet un traitement automatique de pages d'écriture. De nombreux systèmes ont été développés pour : la lecture de cartes ou plans cadastraux, la lecture des fax et l'envoi de courrier électronique.[8]
- La lecture des tickets de transport aérien : Chaque place réservée nécessite trois enregistrements : un auprès de la compagnie, un auprès de l'agence de voyage et un pour le voyageur. Afin de limiter le grand nombre de billets ainsi créés et d'éviter l'attente avant embarquement, de nombreuses compagnies ont recours à un système d'identification automatique qui lit le ticket et compare les indications avec la base de données de chaque vol.
- La lecture de formulaires : De nombreuses enquêtes ou fiches de renseignements utilisent des formulaires préimprimés. L'utilisation d'un système de reconnaissance, capable de lire directement les données dans les zones réservées permet d'effectuer rapidement la saisie de ces documents.[9]
- La lecture des passeports : Certaines douanes sont équipées de lecture de passeports afin d'identifier chaque voyageur. Le système permet de lire le nom, la nationalité, le numéro de passeport et aussi de contrôler directement auprès des bases de données des services d'immigration, l'autorisation de séjour.
- La gestion automatique des chèques bancaires ou postaux [10] : Les chèques sont automatiquement traités grâce à lecture automatique du montant en chiffres et en lettres. L'utilisation. d'un système lisant les deux montants réduit les risques d'erreur.

3.4 Classification de l'ocr

Il n'existe pas de système universel d'OCR qui permet de reconnaître n'importe quel caractère dans n'importe quelle fonte. Tout dépend du type de données traitées et bien évidemment de l'application visée [9].

Il existe plusieurs modes de classification des systèmes OCR parmi lesquels on peut citer :[7]

- Les systèmes « en-ligne » ou « hors-ligne » suivant le mode d'acquisition.
- Les approches globales ou analytiques selon que l'analyse s'opère sur la totalité du mot, ou par segmentation en caractères.

3.4.1 Reconnaissance en-ligne et hors-ligne

Ce sont deux modes différents d'OCR, ayant chacun ses outils propres d'acquisition et ses algorithmes correspondants de reconnaissance.

3.4.1.1 La reconnaissance en-ligne (on-line)

Ce mode de reconnaissance s'opère en temps réel (pendant l'écriture). Les symboles sont reconnus au fur et à mesure qu'ils sont écrits à la main. Ce mode est réservé généralement à l'écriture manuscrite . c'est une approche où la reconnaissance est effectuée sur des données à une dimension . l'écriture est représentée comme un ensemble de points dont les coordonnées sont fonction du temps[11]. La reconnaissance en-ligne présente un avantage majeur c'est la possibilité de correction et de modification de l'écriture de manière interactive vu la réponse en continu du système [12]. L'acquisition de l'écrit est généralement assurée par une tablette graphique munie d'un stylo électronique.

FIGURE 3.1 – Exemple de reconnaissance en ligne

3.4.1.2 La reconnaissance hors-ligne (off-line)

Démarre après l'acquisition. Elle convient aux documents imprimés et les manuscrits déjà rédigés. Ce mode peut être considéré comme le cas le plus général de la reconnaissance de l'écriture. Il se rapproche du mode de la reconnaissance visuelle. L'interprétation de l'information est indépendante de la source de génération [13].

La reconnaissance hors-ligne peut être classée en plusieurs types :

Reconnaissance de texte ou analyse de documents

Dans le premier cas il s'agit de reconnaître un texte de structure limitée à quelques lignes ou mots. La recherche consiste en un simple repérage des mots dans les lignes, puis à un découpage de chaque mot en caractères . Dans le second cas (analyse de document), il s'agit de données bien structurés dont la lecture nécessite la connaissance de la typographie et de la mise en page du document. Ici la démarche n'est plus un simple prétraitement, mais une démarche experte d'analyse de document il y'a localisation des régions, séparation des régions graphiques et photographique, étiquetage sémantique des zones textuelles à partir de modèles, détermination de l'ordre de lecture et de la structure du document [14].

Reconnaissance de l'imprimé ou du manuscrit :

Les approches diffèrent selon qu'il s'agisse de reconnaissance de caractères imprimés ou manuscrits. Les caractères imprimés sont dans le cas général alignés horizontalement et séparés verticalement, ce qui simplifie la phase de lecture [15]. La forme des caractères est définie par un style calligraphique (fonte) qui constitue un modèle pour l'identification. Dans le cas du manuscrit, les caractères sont souvent ligaturés et leur graphisme est inégalement proportionné provenant de la variabilité intra et interscripteurs. Cela nécessite généralement l'emploi de techniques de délimitation spécifiques et souvent des connaissances contextuelles pour guider la lecture [14].

Dans le cas de l'imprimé, la reconnaissance peut être monofonte, multifonte ou omnifonte :[15]

- Un système est dit monofonte s'il ne peut reconnaître qu'une seule fonte à la fois c'est à dire qu'il ne connaît de graphisme que d'une fonte unique. C'est le cas le plus simple de reconnaissance de caractères imprimés .
- Un système est dit multifonte s'il est capable de reconnaître divers types de fontes parmi un ensemble de fontes préalablement apprises.
- Et un système omnifonte est capable de reconnaître toute fonte, généralement sans apprentissage préalable. Cependant ceci est quasiment impossible car il existe des milliers de fontes dont certaines illisible par l'homme .

3.4.2 Reconnaissance globale ou analytique

L'approche globale :

considère le mot comme une seule entité et le décrit indépendamment des caractères qui le constituent. Cette approche présente l'avantage de garder le caractère dans son contexte avoisinant, ce qui permet une modélisation plus efficace des variations de l'écriture et des dégradations qu'elle peut subir. Cependant cette méthode est pénalisante par la taille mémoire, le temps de calcul et la complexité du traitement qui croient linéairement avec la taille du lexique considéré, d'où une limitation du vocabulaire [16].

L'approche analytique :

Contrairement à l'approche globale, le mot est segmenté en caractères ou en fragments morphologiques significatifs inférieurs au caractère appelés graphèmes. La reconnaissance du mot consiste à reconnaître les entités segmentés puis tendre vers une reconnaissance du mot, ce qui constitue une tâche délicate pouvant générer différents types d'erreurs [17], Un processus de reconnaissance selon cette approche est basé sur une alternance entre deux phases : la phase de segmentation et la phase d'identification des segments. Deux solutions sont alors possibles : la segmentation explicite (externe) ou la segmentation implicite (interne).

Par ailleurs, les méthodes analytiques par opposition aux méthodes globales, présentent l'avantage de pouvoir se généraliser à la reconnaissance d'un vocabulaire sans limite à priori, car le nombre de caractères est naturellement fini. De plus l'extraction des primitives est plus aisée sur un caractère que sur une chaîne de caractères [16].

3.5 Critères concernant l'œuvre

3.5.1 Composition du texte

Indiscutablement, les OCRs sont calibrés pour reconnaître du texte, des mots, tous les éléments non purement textuels peuvent perturber la reconnaissance : images, tableaux, formules mathématiques et chimiques, chiffres, hiéroglyphes, annotation manuscrites, éléments graphiques sont autant d'éléments perturbateurs pour les OCRs.[18]

3.5.2 Langues et alphabets

3.5.2.1 La langue à reconnaître

Les accents

Les premiers OCRs ont été développés pour des archives du XXème siècle. Ces documents, en anglais, ne comportent quasiment aucun signe diacritique (seul le point sur le "i" est un accent. Lorsque les OCRs ont commencé à travailler sur des langues étrangères, comme le français, sont apparus les problèmes de reconnaissance des accents. Les accents ont la spécificité d'être beaucoup plus petits qu'un caractère. Or les OCRs fonctionnent tous de la même manière, ils commencent par former une image noir et blanc (binarisation) puis recherchent ensuite les éléments connexes (dites "composantes connexes") qui pourraient être susceptibles d'être des caractères, notamment par leur forme (rapport hauteur /largeur), leur taille (hauteur, largeur). Ces deux étapes sont autant d'occasion pour les algorithmes de faire disparaître les accents : la binarisation peut "effacer des éléments très petits" ; la recherche de composantes connexes de taille d'un caractère éliminera presque sûrement les accents trop petits. La fréquence d'accents dans une langue rend donc le français et l'espagnol et a fortiori les langues slaves plus difficiles à reconnaître que l'anglais.[19]

La longueur des mots

Une des étapes d'un OCR est la "mise en mot". Celle – ci consiste à séparer une ligne en mots. Elle se base sur un algorithme statistique qui cherche parmi les espaces entre caractères ceux qui sont les plus grands, et cherche à déterminer s'ils sont significativement plus grands. Ce procédé est souvent optimisé pour une longueur moyenne de mot proche de celle de l'anglais, c'est-à-dire environ 6 caractères.[19]

Le nombre de symboles

L'alphabet latin contient 26 caractères, mais à peut près 80 caractères supplémentaires sont utilisés dans des documents non techniques. Ces caractères sont des signes de ponctuation, des caractères majuscules, des caractères minuscules et des symboles spéciaux. Tous ces caractères sont incorporés dans le code ASCII (American Standard Code for Information Interchange) qui réunit 96 caractères adoptés par les sociétés industrielles Américaine. Le code Uniforme (unicode) regroupe tous les caractères et les symboles imprimables de toutes les langues. Même les symboles spéciaux qui figurent dans certaines publications spéciales comme les dictionnaires, les textes scientifiques sont inclus dans le code Uniforme.[19]

3.5.2.2 Nombre de langues et alphabets

Les OCRs travaillent mieux [18] avec une seule langue par unité documentaire. Même s'il est possible d'ajouter un dictionnaire d'une autre langue, cette opération peut causer des dégâts considérables sur le reste de la reconnaissance. Notamment le nombre d'erreurs et d'ambiguïtés du texte augmente. De même, la présence d'alphabets non latins peut être préjudiciable à la reconnaissance globale du texte, et affecter même la qualité de la segmentation., les langues anciennes sont mal connues par les OCRs qui travaillent généralement avec de nouveaux dictionnaires.

3.5.3 Références et citations

Les textes comportant des références et citations utilisent souvent les notes en bas de page. Ces notes sont des petites écritures en bas de la page, qui servent à expliquer ou référencer des termes spéciaux. La taille de ces polices est gênante pour l'OCR qui aura du mal à estimer la taille moyenne de la police sur la page. Par ailleurs, la structure du document, pour peut qu'elle soit aussi une des tâches de l'OCR, en est grandement complexifiée.[19]

3.5.4 Ponctuation

Dans les textes narratifs et descriptifs environ 60% des signets de ponctuation sont des virgules et des points. Dans les périodiques scientifiques, les points sont beaucoup plus nombreux que les virgules. En effet, les points sont utilisés dans les nombres réels, dans les abréviations et dans les fonctions scientifiques ce qui augmente leur nombre [, la fréquence des virgules a subit des diminutions au cours de ces dernières siècles. Compte tenu de leurs apparences dans le texte, les points et les virgules sont souvent similaires. En effet leurs petites tailles empêchent les méthodes de reconnaissance de forme de faire la distinction entre les deux.[19]

3.6 Notions typographiques

Le terme " typographie" regroupe l'ensemble des techniques et des procédés permettant de reproduire des textes par l'assemblage de caractères en relief (définition du dictionnaire petit Robert). Le principe de la typographie est apparu au XXème siècle . Avant 1450, les seules reproductions de livres étaient le fruit de scribes. Le premier livre imprimé fut la Bible de Gutenberg. C'est grâce à l'utilisation, par Fust et Schaeffer, des techniques

développées par Gutenberg que cet ouvrage a put être produit en gros volume. Le principe de la typographie consiste à utiliser des poinçons d'acier représentant l'image miroir du type (ou caractère) pour marquer une page de métal tendre. Cette page est ensuite recouverte d'encre et pressée sur le papier. En développant sa technique, Gutenberg fut amené à créer la première police de caractères appelée "blackletter". Cette unique police fut utilisée pendant près d'un siècle. Ce n'est qu'au xvrrme siècle que les principales polices sont apparues.

Le nom de fonte est issu de la façon de fabriquer les poinçons d'acier qu'utilisaient les typographes dans l'imprimerie. Cette notion, synonyme de police, définit l'ensemble des caractères d'un même type (fondus ensemble). Ainsi, la police est essentiellement caractérisée par la nature des empattements utilisés et le dessin utilisé pour les caractères. Chaque caractère de la police est défini par sa hauteur, son corps (qui est défini par la hauteur plus l'espace interligne), sa largeur et sa chasse définie par la largeur et l'espace inter-caractères .[20]

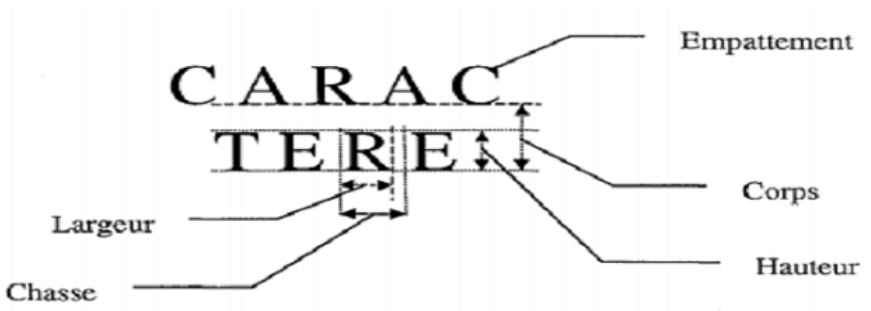


FIGURE 3.2 – Caractéristiques d'une fonte[20]

Les unités utilisées en typographie pour les mesures dimensionnelles sont le point Didot pour les français (1 point Didot =0.3759 mm) et le point PICA pour les anglais (=0.35277 mm). La typographie pouvant par extension définir la mise en forme de documents, une normalisation de la taille des caractères est apparue en fonction des différentes zones du document .

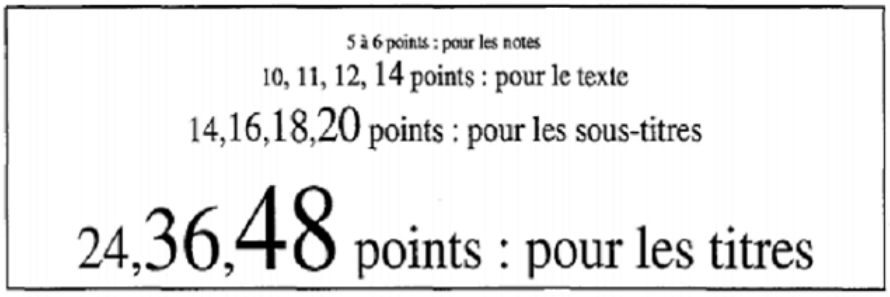


FIGURE 3.3 – Echelle des corps en point Didot[20]

Le dessin représente la forme et l'épaisseur du caractère. L'épaisseur, ou graisse, peut varier en fonction de la taille du caractère variant du maigre, au gras jusqu'au noir. La forme appelée aussi casse, décrit différents caractères de même type .

Police capitale	ABCDEF	Police capitale gras	**ABCDEF**
Police bas de casse	Abcdef	Police capitale gras italique	***ABCDEF***
Police petite capitale	ABCDEF	Police capitale noir	**ABCDEF**
Police capitale italique	*ABCDEF*	Police capitale condensée	ABCDEF

FIGURE 3.4 – Variations du dessin[20]

Il existe deux grandes classifications françaises des caractères d'imprimerie : la classification Thibaudau et la classification Vox.. Ces classifications sont basées sur la différenciation des empattements.

La classification Thibaudau réalisée en 1920 par le typographe Francis Thibaudau distingue 4 grandes familles de caractères : [21]

- Les Didots ayant des empattements fins et filiformes.
- Les Elzévirs ayant des empattements tliangulaires.
- Les Egyptiennes ayant des empattements quadrangulaires.
- Les Antiques n'ayant pas d'empattements.

Didot	ABCD	Elzevir	ABCD
Egypte	ABCD	Antique	ABCD

FIGURE 3.5 – Classification Thibaudau[20]

La classification Vox, réalisée en 1950 par le typographe Maximilien Vox, distingue 9 grandes familles de caractères didones : garaldes, humanes, incises, linéales, mamuaires, mécanes, réales et scliptes. Cette distinction est basée sur 1'évolution de la typographie.

Actuellement, ces classifications ne sont plus suffisantes. En effet, l'informatique et l'apparition des traitements de texte ont permis d'augmenter le nombre et la variété des polices de caractères. De nombreux logiciels permettent à chaque utilisateur de définir sa propre police de caractères. Les polices sont maintenant définies de manière vectolielle, représentation mathématique des contours des polices : ce sont les polices vectorielles TRUE TYPE. En outre, de plus en plus, les nouveaux types de polices développés sont des polices fantaisistes ou des polices imitant l'écliture manusclite .[20]

Airbus special	ABCDEF	Maiandra	ABCDEF
Lucidia	ABCDEF	Goudy	ABCDEF
Flexure	ABCDEF	cmb	ABCDEF
Comic	ABCDEF	Harrington	ABCDEF

FIGURE 3.6 – Exemples de polices True Type[20]

3.7 Chaine de numérisation

Un texte est une association de caractères appartenant à un alphabet, réunis dans des mots d'un vocabulaire donné. L'OCR doit retrouver ces caractères, les reconnaître d'abord individuellement, puis les valider par reconnaissance lexicale des mots qui les contiennent. Cette tâche n'est pas triviale car si l'OCR doit apprendre à distinguer la forme de chaque caractère dans un vocabulaire de taille souvent importante, il doit en plus être capable de la distinguer dans chacun des styles typographiques (polices), chaque corps et chaque langue, proposés dans le même document.

Un système de reconnaissance de texte est composé de plusieurs modules : segmentation, apprentissage, reconnaissance et vérification lexicale [23].

La structure d'un système d'OCR comporte trois parties principales : l'acquisition et traitement d'image, l'analyse du document puis l'interface de sortie vers l'environnement.

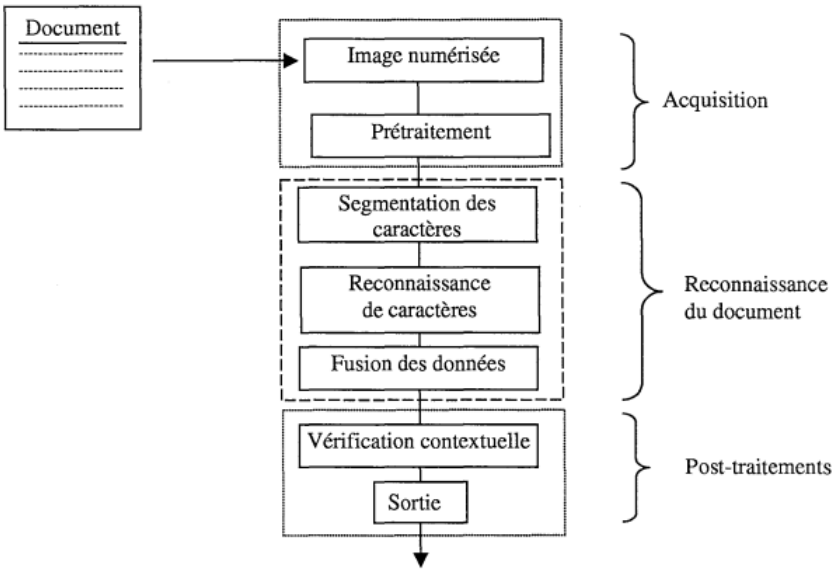


FIGURE 3.7 – Synoptique d'un système de reconnaissanc[23]

3.7.1 Acquisition

3.7.1.1 Principe

L'acquisition du document est opérée par balayage optique. Le résultat est rangé dans un fichier de points, appelés pixels, dont la taille dépend de la résolution. Les pixels peuvent avoir comme valeurs : 0 (éteint) ou 1 (actif) pour des images binaires, 0 (blanc) à 255 (noir) pour des images de niveau de gris, et trois canaux de valeurs de couleurs entre 0 et 255 pour des images en couleur. La résolution est exprimée en nombre de points par pouce (ppp). Les valeurs courantes utilisées couramment vont de 100 à 400 ppp. Par exemple, en 200 ppp, la taille d'un pixel est de 0,12 mm, ce qui représente 8 points par mm. Pour un format classique A4 et une résolution de 300 ppp, le fichier image contient 2 520 $\times$ 3 564 pixels. Il est important de noter que l'image n'a à ce niveau qu'une simple structure de lignes de pixels qu'il faudra exploiter pour retrouver l'information. [23]

3.7.1.2 Matériel

La technicité des matériels d'acquisition (scanner) a fait un bond considérable ces dernières années. On trouve aujourd'hui des scanners pour des documents de différents types (feuilles, revues, livres, photos ...etc.). Leur champ d'application va du "scan" de

textes au "scan" de photos en 16 millions de couleurs (et même plus pour certains). La résolution est classiquement de l'ordre de 300 à 1200 ppp selon les modèles.[23]

3.7.1.3 Formats

Il existe différents formats de représentation des fichiers images : TIFF, JPEG, GIF, PNG,...etc. dépendant des propriétés de l'image, comme le chromatisme et l'animation, et de la tolérance à la perte de l'information dans le processus de compression. La structuration des données est différente pour chaque format d'image. Un format d'image comprend habituellement un en-tête, contenant des informations générales sur l'ensemble du fichier (par ex. n° de version, ordre des octets ...etc.), un ou plusieurs répertoires de paramètres, caractérisant la représentation bitmap de l'image, suivis par les données de l'image qui seront lues et interprétées suivant la valeur des paramètres. Un paramètre particulier est dédié au type de compression autorisé sur le format, avec ou sans perte d'information. Il est important, dans la mesure du possible, d'écarter les formats propriétaires (comme GIF, sous Licence Unisys) et de leur préférer des formats libres de tous droits. D'autres précautions sont également à prendre en compte concernant les changements de version que peut recouvrir un format, comme c'est le cas du format TIFF dont certaines versions peuvent ne pas être reconnues par certains logiciels. En attendant la généralisation du PNG, qui est l'émanation de recommandations du consortium W3C (1996), le format TIFF est pour l'instant le format le plus répandu pour les documents textuels, avec le mode de compression sans perte CCITT groupe IV. Lancé par la société Aldus Corporation qui visait le marché de la PAO, il est devenu un standard de fait sur le marché de la microinformatique et permet d'échanger des images en mode point entre systèmes hétérogènes de PAO(publication assistée par ordinateur) (PageMaker, Xpress), PréAO (Freelance, PowerPoint), ou éditeur raster (Photoshop, Corel Draw). Il a été spécifié pour prendre en compte différents types d'images (monochromes, à niveaux de gris ou en couleurs), différentes caractéristiques qui dépendent du mode d'acquisition et divers modes de compression. Quasiment tous les produits du marché pour la manipulation des images raster offrent aujourd'hui une interface TIFF, et des librairies de manipulation d'images TIFF, libres de droits, existent [23].

3.7.1.4 Qualité des données

La qualité de la saisie est tributaire de plusieurs facteurs déterminants, comme: [23]

La résolution. Le choix de la résolution de numérisation est fonction de la qualité du contenu en termes typographiques. La difficulté souvent rencontrée est de pouvoir adapter la résolution aux différentes tailles de caractères et épaisseurs de graphiques

présents dans le document, ne nécessitant pas le même niveau de précision (échantillonnage).

Le contraste et la luminosité. La luminosité permet de jouer sur l'éclairage du document à capturer : de plus clair à plus sombre. Le contraste permet de faire varier l'accentuation ou l'atténuation des transitions Noir/Blanc. Ces deux paramètres sont souvent corrélés entre eux et jouent un grand rôle dans la qualité de reconnaissance.

La qualité du support influe sur la qualité du résultat. Le choix de la résolution peut permettre de corriger certains de ses défauts comme ceux relatifs, par exemple, au grain du papier. En effet, un bruit à 200 ppp est plus important qu'à 300 ppp. Aujourd'hui, les logiciels d'OCR peuvent être paramétrés pour prendre en compte la qualité du support.

L'inclinaison est une source d'erreur classique, relativement gênante pour les systèmes de reconnaissance qui utilisent l'horizontale comme référentiel de base pour l'extraction des lignes de texte et la modélisation de la forme des lettres. Elle est de plus en plus maitrisée grâce à l'existence de logiciels de redressement appliqués systématiquement sur les documents à leur entrée.

3.7.2 Prétraitement de l'OCR

Lors de la saisie du document, des déformations dues à la chaîne d'acquisition peuvent intervenir. De nombreux facteurs entrent en compte : la qualité des objectifs, la chaîne de transmission et la numérisation. Les premières opérations effectuées consistent à "nettoyer" l'image initiale. Il s'agit d'éliminer les déformations induites par l'acquisition, de concentrer et de localiser la représentation du caractère.

3.7.3 Reconnaissance du texte

3.7.3.1 Principe

La reconnaissance de caractères est réalisée à l'aide de systèmes dédiés appelés OCR. Son but est de convertir l'image du texte en un texte lisible par ordinateur, en faisant le moins de fautes possibles sur la conversion des caractères. L'existence aujourd'hui de plusieurs outils de ce type a conduit peu à peu à définir des critères de choix pour sélectionner l'OCR le plus efficace et surtout le mieux adapté à son application. Longtemps, le critère d'efficacité était lié à un taux de reconnaissance élevé, pensant qu'une technologie efficace est une technologie sans défaut. En effet, il faut admettre que le taux de 100% reste un objectif à atteindre. Mais réussir une opération de numérisation exploitant la

technologie d'OCR nécessite un certain nombre de règles dans la mise en œuvre de ces applications. Il est confirmé que le taux de reconnaissance ne dépend pas du seul moteur de reconnaissance mais d'un ensemble de précautions à prendre lors de la mise en œuvre de l'application. Parmi ces précautions, nous trouvons :

- L'aide dans la préparation en amont du document papier pour réussir une bonne acquisition (ex : réglage du couple contraste/luminosité) et disposer ainsi d'une image facilement traitable.
- L'aide du moteur d'OCR dans le choix de ses paramètres pour mieux s'adapter au type du contenu, en prenant en compte la qualité du document, la langue du texte, la mise en page employée...etc.
- La mise en place de plusieurs moteurs de reconnaissance de caractères, permettant en travaillant sur plusieurs seuils de confiance et sur les résultats apportés par chacun des moteurs d'OCR, de prendre une décision confortée sur le caractère analysé.[23]

3.7.3.2 Fonctionnement

Un texte est une association de caractères appartenant à un alphabet, réunis dans des mots d'un vocabulaire donné. L'OCR doit retrouver ces caractères, les reconnaître d'abord individuellement, puis les valider par reconnaissance lexicale des mots qui les contiennent. Cette tâche n'est pas triviale car si l'OCR doit apprendre à distinguer la forme de chaque caractère dans un vocabulaire de taille souvent importante, il doit en plus être capable de la distinguer dans chacun des styles typographiques (polices), chaque corps et chaque langue, proposés dans le même document. Cette généralisation omnifonte et multilingue n'est pas toujours facile à cerner par les OCRs et reste génératrice de leurs principales erreurs. Un système de reconnaissance de textes est composé de plusieurs modules : segmentation, apprentissage, reconnaissance et vérification lexicale[23].

3.7.3.3 La segmentation

La segmentation est le processus consistant à décomposer l'image d'un texte en entités (mots, caractères ou graphèmes) qui font partie d'un alphabet prédéfini selon le but visé. Par ailleurs, elle permet de réduire la complexité des modules de traitements utilisés par la suite [21]. Dans les systèmes de reconnaissance la segmentation est une opération très critique [24]. En effet, la séparation des lignes, des mots, des pseudo-mots, des caractères et des graphèmes constituent des opérations difficiles et coûteuses, tant les écritures sont variées, les lignes sont parfois enchevêtrées et les caractères généralement liés (cas de l'arabe : l'écriture est semi-cursive) les unes aux autres. De ce fait, au cours de la mise au point de la segmentation, les chercheurs ont souvent recourt à deux phénomènes qui sont naturels, la sur-segmentation ou fausse détection (un caractère est découpé en

plusieurs entités qui peuvent être problématiques ou non, selon l'utilisation qui en sera faite. Par exemple, la segmentation en petits segments) et la sous segmentation ou non détection (quelques objets de segmentation sont à cheval sur deux caractères consécutifs. Ce cas pose un problème, car on ne peut pas reconnaitre correctement certains caractères qui composent le mot). Selon la littérature, le problème le plus ardu c'est le cas de la segmentation de l'écriture cursive. Dans le but de résoudre cette problématique plusieurs algorithmes de segmentation existent mais posent souvent une polémique autour de choix de l'un par rapport à l'autre, de ce fait, l'utilisation d'un algorithme est conditionnée par son efficacité et le type de la graphie étudiée. Cette efficacité ne peut pas être déterminée que relativement au traitement qui sera fait en aval de cette étape. Les solutions proposées se basent sur deux stratégies de segmentation différentes ci-dessous :

Segmentation explicite

La segmentation explicite, s'appuie sur un découpage à priori de l'image en sous-unités qui peuvent être des lettres ou des graphèmes. Cette décomposition se base directement sur une analyse morphologique du texte ou de mot, ou sur la détection des points caractéristiques tels que les points d'intersection, les points d'inflexion, les boucles à l'intérieur du texte ou de mot pour localiser les points de segmentation potentiels. Dans le cas de segmentation en graphèmes, les textes ou les mots sont alors reconnus non comme une suite de lettres reconnues indépendamment, mais comme une suite de graphèmes globalement comparés à l'entrée. L'avantage de cette segmentation c'est que l'information est localisée explicitement, puisque la sépararation des lettres non pas d'après leur reconnaissance, mais d'après des critères topologiques ou morphologiques. il n'existe pas de méthode de segmentation fiable à 100%, toute erreur de segmentation pénalise les performances de système.

Plusieurs approches proposent la segmentation directe d'un texte ou mot en graphèmes primitifs, suivi par une étape de combinaison de ces graphèmes en caractères [25]. Il existe quatre approches pour la mise en œuvre d'une segmentation explicite :

- Les approches basées sur des analyses par morphologie mathématiques [26], permettent la sélection des points de segmentation en utilisant le principe de régularité et singularité.
- Les approches basées sur l'analyse des contours [27], déterminent les candidats de coupure en s'appuyant sur les extremas locaux du contour0.
- Les approches basées sur l'analyse du profil d'histogramme de projection verticale .
- Les approches basées sur l'analyse du squelette consistent à repérer les points de coupure sur le squelette on se basant sur des seuils ajustés.

Segmentation implicite

La segmentation est dite implicite lorsque celle-ci est basée sur un moteur de reconnaissance pour valider et classer les hypothèses de segmentation (recherche de chemin des points de segmentation possibles). Dans ce cas, la segmentation et la reconnaissance sont réalisées conjointement, d'où le nom parfois employé de "segmentation-reconnaissance intégrée". Contrairement à la segmentation explicite, on ne procède pas à une segmentation à priori en entrée, mais à une segmentation aveugle du mot dans le sens où elle ne dépend en aucun cas d'une analyse de l'image à segmenter et qui dépend d'une compétition des classes des lettres ou graphèmes en sortie du classifier. Ce dernier recherche dans la séquence des segments, des composantes ou des regroupements de graphèmes qui correspondent à ces classes de lettres. Cette recherche peut se faire au moyen de la segmentation basée sur les fenêtres glissantes [28].L'avantage de cette segmentation c'est que l'information est localisée par les modèles des lettres et la validation se fait par ses modèles. Dans les approches à segmentation implicite, la tâche de segmentation est accomplie par le système. Elle est soutenue simultanément par un processus de reconnaissance, en évitant la pré-segmentation d'un mot en lettres ou entités plus fines.

3.7.3.4 La reconnaissance de caractères

Le processus de reconnaissance de caractères est le plus complexe. Généralement, il comporte deux phases : l'extraction de caractéristiques et la classification. L'extraction de caractéristiques permet de déterminer un vecteur dont les composantes caractérisent chaque type de caractère. La classification va permettre de déterminer la classe d'appartenance du caractère à l'aide de ce vecteur de caractéristiques.

Extraction des caractéristiques

Le processus d'extraction consiste à représenter un caractère par un vecteur de caractéristiques, le codage le plus élémentaire consiste à construire un vecteur constitué d'autant de composantes qu'il y a de pixels dans l'image. Leur niveau de gris définit alors le codage de ce vecteur. Afin d'éviter l'utilisation de vecteurs de taille trop importante, les techniques d'extraction de caractéristiques cherchent à définir un codage, déduit d'un ensemble de mesures, qui distingue le mieux les différents types de caractères.

Ces mesures doivent être les plus génériques possible pour ne pas dépendre des polices utilisées mais aussi suffisamment précises pour identifier chaque caractère. Les mesures sont effectuées à partir de la matrice image du caractère et représentent généralement des propriétés locales et/ou globales. Les caractéristiques obtenues tiennent compte des particularités métriques, statistiques ou topologiques. Les méthodes les plus courantes sont [29] : la détection de traits horizontaux et verticaux, le calcul d'intersection de lignes, le

calcul de concavités, la détection et la localisation de boucles et les mesures liées à la dimension et à la surface du caractère.

Classification

La partie classification est généralement composée de deux phases : l'apprentissage des types de caractères à reconnaître et l'identification de caractères en rapport avec l'apprentissage effectué. Certaines techniques ont tenté d'utiliser un processus de reconnaissance de caractères sans apprentissage, mais aucune jusqu'à présent ne s'avère fiable .[20]

- L'apprentissage permet au système d'élaborer sa bibliothèque de caractéristiques. Avant de pouvoir effectuer la reconnaissance, le système doit apprendre les caractéristiques de chaque caractère et les garder en mémoire. La méthode d'apprentissage la plus triviale consiste à mémoriser différents vecteurs représentant chaque caractère présenté durant cette phase.
- L'apprentissage étant fait, l'identification consistera à déterminer, à partir de ces caractéristiques apprises, la classe d'appartenance du caractère présenté.

On référence trois grandes catégories d'approches utilisées pour la reconnaissance de caractères : 1'approche statistique, l'approche structurelle et l'approche stochastique. Cette classification regroupe les techniques d'extraction de caractéristiques ainsi que des techniques de classification associées.

- L'approche statistique : ou approche globale va utiliser des mesures faites sur le caractère afin de le coder sous forme de vecteur. Ce vecteur définit le caractère dans un nouvel espace de représentation. Cette approche consiste à construire un classifieur permettant de distinguer les différentes classes à l'intérieur de cet espace de représentation.
- L'approche structurelle : est plus proche du fonctionnement du raisonnement humain. Elle consiste à ordonner hiérarchiquement les caractéristiques déduites des caractéristiques locales de la forme, sous forme d'un arbre de décision. L'identification s'effectuera par étapes successives.
- L'approche stochastique est essentiellement utilisée pour la reconnaissance manusctite. Elle permet de modéliser 1'écriture en tenant compte des états rencontrés par le stylet. L'information temporelle et le sens du tracé vont permettre de coder l'écriture. Cependant, dans le cas de l'écriture imprimée, cette approche n'apporte que peu d'intérêt.

Fusion de données

Lorsqu'individuellement aucune des techniques choisies ne se révèle satisfaisante, une combinaison de ces techniques permet d'obtenir un meilleur résultat. Ce procédé constitue une des tendances actuelles. En effet, il permet d'utiliser la complémentarité des techniques associant ainsi leurs qualités. Cependant, la combinaison des techniques doit respecter

des règles qui tiennent compte des spécificités de chacune. L'élaboration de ces règles est assumée par l'étape dite de fusion. Il existe trois approches différentes [30] :

- Une approche séquentielle qui consiste à appliquer successivement chaque technique en s'appuyant sur le résultat de la précédente.
- Une approche parallèle qui consiste à appliquer l'ensemble des techniques séparément puis à combiner les résultats dans une étape finale. Chaque technique de classification est appliquée aux mêmes données
- Une approche hybride qui combine approche séquentielle et approche paralléle.

3.7.3.5 Le post-traitement

Est effectué quand le processus de reconnaissance aboutit à la génération d'une liste de lettres ou de mots possibles, éventuellement classés par ordre décroissant de vraisemblance. Le but principal est d'améliorer le taux de reconnaissance en faisant des corrections orthographiques ou morphologiques à l'aide de dictionnaires de digrammes, tri-grammes ou n-grammes. Quand il s'agit de la reconnaissance de phrases entières, on fait intervenir des contraintes de niveaux successifs : lexical, syntaxique ou sémantique.[23]

3.8 Types des erreurs

3.8.1 Erreurs de segmentation

La segmentation du document conduit à la décomposition du document en unités structurelles telles que des régions textuelles ou des graphiques. Une mauvaise application de la méthode de segmentation conduit à des erreurs [31].

Ces erreurs sont :

- Fusion horizontale de régions textuelles : cette erreur conduit à la fusion de lignes adjacentes appartenant à des colonnes différentes. Elle influe sur l'ordre de lecture comme le montre la figure 3.8 où l'ordre normal : 1, 2, 3, 4 est transformé en 1, 3, 2, 4.

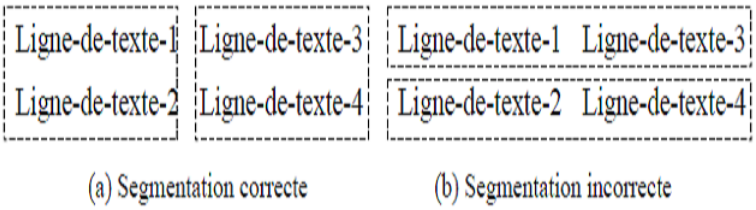


Figure 3.8: Fusion horizontale des régions.[23]

- Fusion verticale de régions textuelles : cette erreur conduit à regrouper deux paragraphes. Elle n'altère pas l'ordre de lecture mais sa correction est nécessaire pour la classification du texte,
- Scission horizontale de régions textuelles : cette erreur conduit à un faux ordre de lecture.
- Scission verticale de régions textuelles : ce cas est similaire au précédent, il ne cause pas d'erreur grave car l'ordre de lecture n'est pas changé.
- Région non détectée : cette erreur indique la non détection d'une région de texte, assimilée sans doute à un graphique ou à du bruit.
- Graphique/bruit confondu à du texte : Ceci indique que l'OCR a dû interpréter un graphique ou du bruit comme du texte. Cette erreur conduit à des séquences perturbées de caractères dans le texte. Cette erreur concerne également les formules mathématiques.
- Fusion horizontale avec graphique/bruit : ceci conduit, comme dans le cas précédent, à l'insertion de séquences erronées de caractères dans le texte.
- Fusion verticale avec graphique/bruit : Ce cas est identique au cas précédent, sauf qu'il se produit verticalement.

3.8.2 Erreurs de reconnaissance de caractères

Un OCR peut faire quatre types d'erreur sur la reconnaissance des caractères :[23]

- Une confusion, en remplaçant un caractère par un autre, si les caractères sont morphologiquement proches (par exemple. « o,0 », « c,(», « n,h », « s,5 ») .
- Une suppression, en ignorant un caractère, considéré comme un bruit de l'image,
- Un ajout, en dédoublant un caractère par deux autres dont la morphologie de leurs formes accolées peut être proche du caractère (par exemple. « m, rn», « d,cl », « w,vv »).

En plus de ces erreurs, les OCR peuvent indiquer les doutes qu'ils ont eu sur certains caractères et certains mots. Ces doutes peuvent servir lors de la phase de correction.

Cette figure donne des exemples d'erreurs de reconnaissance possibles.

Château	Chanteau	Chat eau	Chapeau	Gâteau
Original	*Ajout*	*Ajout*	*Confusion*	*Confusion & Suppression*

FIGURE 3.9 – Différents cas d'erreurs de reconnaissance possibles sur le mot "Château"[23]

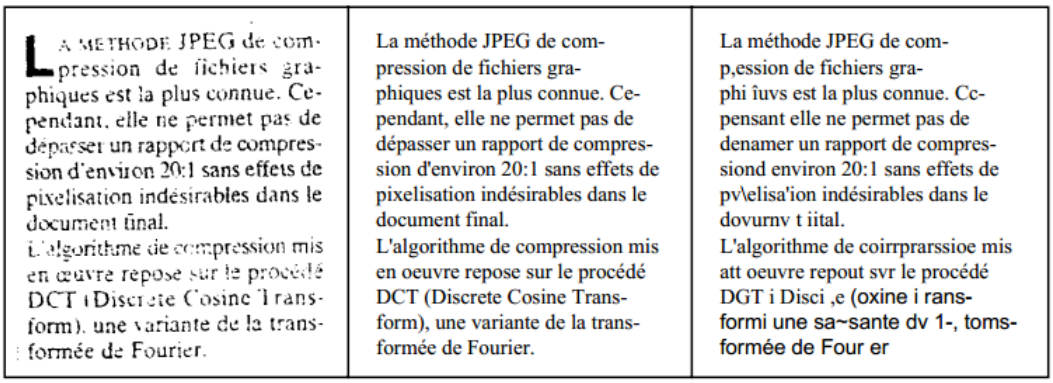


FIGURE 3.10 – Texte bruité et sa reconnaissance OCR.

3.8.3 Erreurs de reconnaissance de mots

Une cause d'erreur fréquente est la mauvaise interprétation par l'OCR de la largeur des espaces. Cela peut conduire soit à la fusion de deux mots, soit à la scission d'un mot. La cause principale de suppression correspond à une mauvaise image du mot, comme le surlignage , le fond grisé, le raturage ...etc. ne permettant pas à l'OCR de prendre de décision même partielle sur les caractères, le conduisant à rejeter le mot entier. La suppression peut même toucher des lignes entières, comme les lignes d'en-tête à cause de l'inverse vidéo, de la pliure du papier ...etc.[23]

3.9 Conclusion

Dans ce chapitre,nous avons présenté la reconnaissance de caractère d'une manière général ainsi que l'outil permettant cette reconnaissance en détaillant ses différents aspects. Finalement nous avons présenté quelques limites ou obstacles rencontre par les OCRs. Dans le chapitre suivant nous allons présenter des méthodes pour améliorer la performance d'OCR.

Chapitre 4

Prétraitement

4.1 Introduction

Le prétraitement d'une image regroupe un ensmeble d'opération ayant pour objectif d'amélioration la qualité des images dans le but de faciliter l'extraction des informations présenté dans ces derniers. Il existe différent type de prétraitement permettant d'améliorer la qualité des images numériques. Nous allons présenter quelques-uns d'entre eux dans ce qui suit.

4.2 Le prétraitement sur l'image

Le pré-traitementre groupe l'ensemble des processus visant à améliorer les caractéristiques d'une image

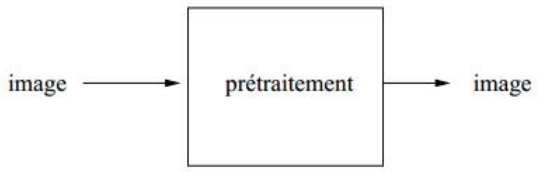


FIGURE 4.1 – schéma de processus d'analyse d'image

- Le lissage local : il s'agit de supprimer le bruit, ou les petites variations, présent dans une image. L'intensité d'un pixel est transformé en fonction des intensités sur un petit voisinage du pixel
- L'amélioration d'images : consiste à modifier les caractéristiques visuelles de l'image (contraste, ...) pour faciliter son interprétation par l'œil humain.

– La restauration d'images : a pour but de supprimer les dégradations subies par une image à l'aide de connaissance a priori sur ces dégradations.

4.3 Binarisation

Les systèmes de reconnaissance nécessitent une étape de binarisation qui vise à séparer les pixels de texte des pixels de l'arrière-plan de l'image traitée. En fait, la plupart des systèmes ne fonctionnent que sur des images binaires. La plus simple façon pour obtenir une image binaire est de choisir une valeur seuil, puis de classer tous les pixels dont les valeurs sont au-dessus de ce seuil comme étant des pixels d'arrière plan, et tous les autres pixels comme étant des pixels de texte. Soit l'image I(M×N), supposons que f(x,y) représente le niveau de gris du pixel aux coordonnées (x,y), $0 \leq x \leq M$, $0 \leq y \leq N$ et s est le seuil choisi, les pixels de l'objet sont ceux ayant le niveau de gris inférieur à s et les autres ayant le niveau de gris supérieur à s sont des pixels du fond. Alors, l'image binarisée g est déterminée par les pixels (x,y) dont la valeur est donnée par l'équation :

g(x,y)=1 si f(x,y)>s

g(x,y)=0 si f(x,y)<=s

Dans la pratique, cette situation idéale ne se rencontre que très rarement. Les niveaux de gris associés au fond et aux objets présents sur l'image sont supposés être suffisamment différents pour qu'une bonne discrimination puisse être faite. Cependant, cette dichotomie n'est évidemment pas parfaite en raison de défauts d'éclairage ou de bruits introduits par l'opération d'acquisition lui-même. Par conséquent, un mauvais choix d'un seuil de binarisation peut détruire une grande part d'information utile contenue dans l'image en dégradant notamment la qualité des caractères à reconnaître par l'OCR , ces caractères peuvent ainsi être fragmentés ou fusionnés.[32]

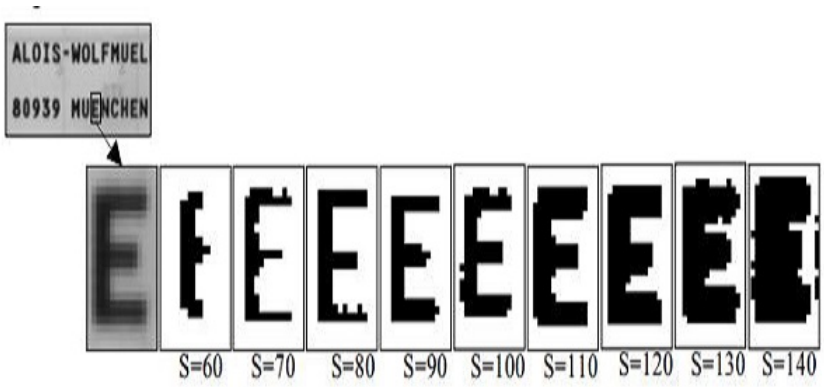


FIGURE 4.2 – Effet de seuillage sur la qualité des caractères.[32]

FIGURE 4.3 – Résultat de la binarisation de différentes images par le même seuil S=120.[32]

Ainsi, une binarisation appliquée directement sur les images de documents dégradés introduit de nombreux artéfacts qui entraînent des erreurs dans les modules suivants d'analyse. Par conséquent, il est nécessaire d'appliquer des prétraitements de rehaussement du contraste, d'égalisation de l'histogramme et de réduction du bruit par filtrage afin d'améliorer la qualité de cette binarisation.

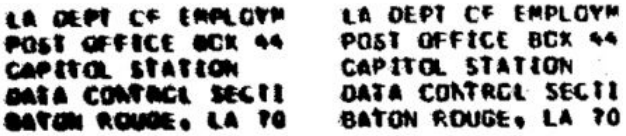


FIGURE 4.4 – Exemple d'image binaire (à gauche) sans prétraitement et à droite avec prétraitement (filtrage)[32]

Cette figure illustre le fait qu'une bonne binarisation doit être capable de conserver tous les caractères et sans récupérer trop de bruit. Pour résoudre ce problème, . On distingue essentiellement trois catégories de méthodes selon la nature de seuillage utilisé : les méthodes globales, les méthodes locales et les méthodes hybrides qui exploitent les deux approches précédentes.[32]

4.3.1 Seuillage global

La méthode de seuillage globale consiste à calculer un seuil unique à partir d'une mesure globale sur toute l'image et la recherche de seuil s'effectue par l'analyse de l'histogramme des niveaux de gris et par la détermination d'un minimum local (voir figure ci-dessous). Il nous permet de décider de l'appartenance d'un pixel à l'objet ou au fond sur toute l'image. telque les pixels ayant un niveau de gris inférieur au seuil sont mis en noir et les autres en blanc.

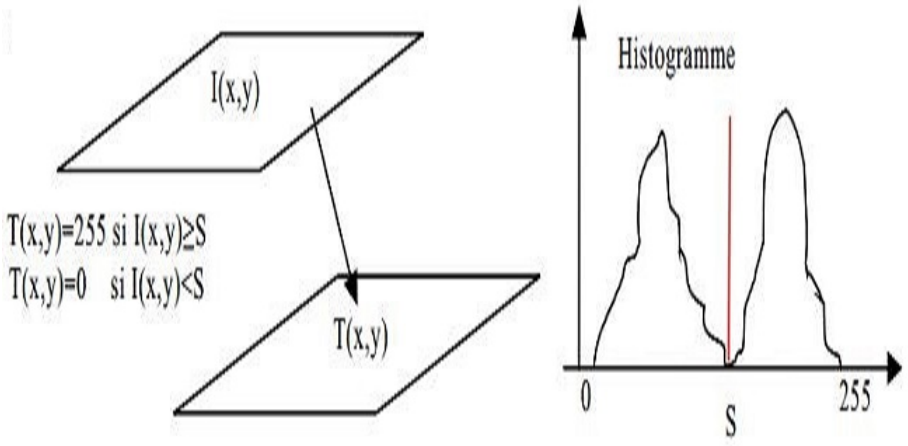


FIGURE 4.5 – Principe de la binarisation par seuillage globale[32]

4.3.1.1 Approches basées sur l'analyse discriminante(méthode Otsu)

La méthode d'OTSU est utilisée pour effectuer un seuillage automatique à partir de la forme de l'histogramme de l'image.Cette méthode nécessite donc le calcul préalable de l'histogramme de l'image. L'algorithme suppose alors que l'image à binariser ne contient que deux classes, (Les objets et l'arrière-plan). L' algorithme itératif calcule alors le seuil optimal T qui sépare ces deux classes afin que la variance intra-classe soit minimale et que la variance inter-classe soit maximale.[33]

Dans le cadre de la binarisation par la méthode d'Otsu, la séparation s'effectue à partir de la moyenne et de la variance. On calcule donc :

Variance intra-classe :

$$\delta_w^2 = w_1(T) * \delta_1^2(T) + w_2(T) * \delta_2^2(T).$$

- Oméga 1 représente la probabilité d'être dans la classe 1.
- Oméga 2 représente la probabilité d'être dans la classe 2.

- Sigma 1 représente la variance de la classe 1.
- Sigma 2 représente la variance de la classe 2

Calcul de la probabilité de la classe 1 et 2 :

Pour calculer la probabilisé d'être dans la classe 1 ou 2 en fonction du seuil T, il suffit de sommer les probabilités de chaque niveau de gris.

$$w_1(T) = \sum_{K=0}^{T-1} P(K)$$
$$w_2(T) = \sum_{K=T}^{255} P(K)$$

Calcul de la variance de chaque classe :

$$\delta_1^2(T) = \frac{\sum_{i=1}^{T-1} (N1(i) - Moy_1(T))^2 * P(i)}{w_1}.$$
$$\delta_2^2(T) = \frac{\sum_{i=T}^{255} (N2(i) - Moy_2(T))^2 * P(i)}{w_2}.$$

- N1 est un vecteur de 1 à T-1.
- N2 est un vecteur de T à 255.
- Moy1 représente la moyenne de la classe 1.
- Moy2 représente la moyenne de la classe 2.

Calcul de la moyenne de chaque classe :

La moyenne de chaque classe est calculée en sommant le vecteur N qui est multiplié par la probabilité de chaque niveau de gris. Le tout est ensuite divisé par la probabilité de la classe.

$$Moy_2(T) = \frac{\sum_{i=1}^{T} N1(i) * p(i)}{w1(T)}$$

$$Moy_2(T) = \frac{\sum_{i=T+1}^{255} N2(i) * p(i)}{w2(T)}$$

- N1 est un vecteur de 0 à T-1 .
- N2 est un vecteur de T à 255.

Discutions de la méthode Otsu :

Une variante de cette approche a été proposée par Tsai [34] : elle consiste initialement à découper l'image récursivement en quadtree, et à appliquer ensuite un seuillage de type d'Otsu dans chaque bloc. Cette méthode s'adapte bien à la forme des tracés et permet de résoudre les problèmes liés à une distribution non uniforme de l'intensité lumineuse. L'inconvénient de cette approche est son coût calculatoire et le risque de générer des blocs entièrement noirs.

4.3.1.2 Approches basées sur les réseaux de neurones

Babaguchi et al. [35] ont proposé une méthode de binarisation basée sur un modèle connexionniste (CMB). Cette technique s'articule sur deux phases (apprentissage et binarisation). Dans la phase d'apprentissage, le réseau utilise l'algorithme de rétro propagation sur toute la base d'images, chacune étant représentée par son histogramme propre et un seuil désiré. Dans la phase de binarisation, le réseau reçoit à l'entrée un histogramme d'une image inconnue et retourne en sortie le seuil optimal de binarisation.[32]

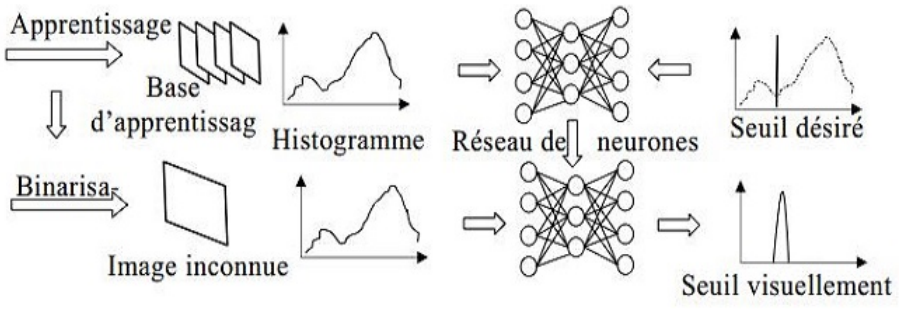


FIGURE 4.6 – Principe général de la méthode de binarisation par réseaux de Neurones.[32]

4.3.1.3 Les techniques d'entropie

Elle est proposée par Kapur et al[36], dans cette méthode, les classes de premier plan et d'arrière plan sont considérées comme deux sources différentes. Lorsque la somme des deux entropies de classe est un maximum l'image elle est dite de seuil optimal.

4.3.1.4 Discussion des méthodes de seuillage global

Ces méthodes conviennent pour des documents simples et de bonne qualité. Néanmoins, elle n'est plus applicable lorsque la qualité d'impression du texte n'est pas constante

dans toute la page et également si le fond est bruité ou non homogène, dans ce cas des taches parasites peuvent apparaitre. La Figure ci-dessous montre le résultat insuffisant retourné par un seuil global, dans le cas d'une mauvaise illumination du document, ou dans le cas où le texte passerait de noir sur fond blanc à blanc sur fond noir. Pour pallier à ces problèmes, il fallut trouver des techniques permettant d'adapter localement le niveau du seuil.

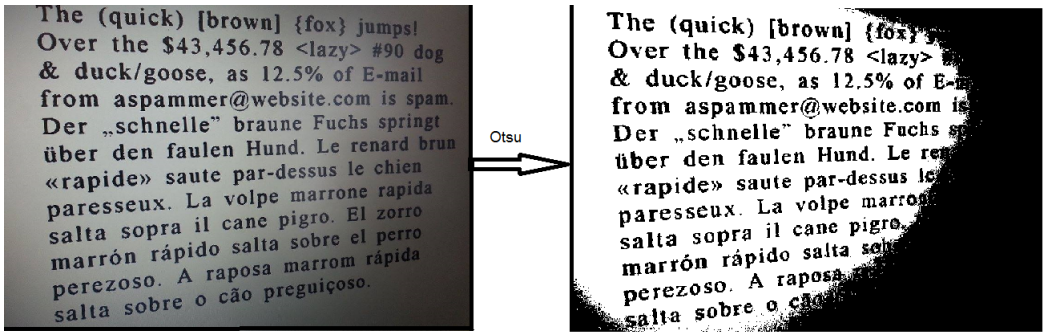


FIGURE 4.7 – Problème de seuillage global

4.3.2 Seuillage local

Le principe du seuillage local est d'adopter une étude localisée autour du pixel pour déterminer quel seuil utiliser. Pour réaliser cette étude locale, les techniques utilisent une fenêtre centrée sur le pixel à étudier (voisinage de pixel). Cette fenêtre peut avoir différentes tailles, souvent en fonction de la taille moyenne du texte dans le document.

Il existe plusieurss méthodes de seuillage local,nous allons presenter dans ce qui va suivre celles qui sont les plus utilisées.

4.3.2.1 Méthode de Bernsen 1986

C'est une méthode locale adaptative dont le seuil est calculé pour chaque pixel de l'image [37]. Ainsi pour chaque pixel de coordonnés (x, y), le seuil est donné par :

$S(x, y) = \frac{Pbas+Phaut}{2}$

Tel que Pbas et Phaut sont le niveau de gris le plus bas et le plus haut respectivement, dans une fenêtre carré $r * r$ centré sur le pixel (x, y). Cependant si la mesure de contraste C(x, y)= (Phaut - Pbas) est inférieur à un certain seuil S, alors le voisinage consiste en une seule classe : fond ou bien texte. Les tests réalisés montrent que le choix de S=15 et r=15 donne les meilleurs résultats.[38]

4.3.2.2 Méthode de Niblack 1986

L'algorithme de Niblack [39] est une amélioration de la méthode de Bernsen qui calcule un seuil local à chaque pixel en glissant une fenêtre rectangulaire sur toute l'image. Le seuil S est calculé en utilisant la moyenne m et la variance delta de tous les pixels dans la fenêtre (voisinage du pixel en question). Ainsi le seuil S est donné par :

$S = m + k * \delta$

Tel que k est un paramètre utilisé pour déterminer le nombre de pixels de contours considérés comme des pixels de l'objet, et prend des valeurs négatives (k est fixé (- 0.2) par les auteurs).

Dans [40], une fenêtre de taille 25 * 25 donne des bons résultats.

Discutions de la méthode Niblack :

Principal avantage de l'Niblack est qu'elle reconnaît correctement les zones de texte, mais crée un beaucoup de données de binarisation bruyants pour les régions non-textuelles de la Contexte.

4.3.2.3 Méthode de SAUVOLA

L'algorithme de Sauvola est une modification de celui de Niblack elle insère dans la méthode de Niblack des constantes afin d'améliorer la méthode sur les zones uniformes et pour donner plus de performance dans les documents avec un fond contient de texture claire ou bien trop de variation et illumination inégal.[41]

La méthode de SAUVOLA est une technique de seuillage local. Avec cette méthode, le seuil T pour chaque pixel de l'image est donnée par :

$$T(x, y) = mean(x, y) * [1 + K(\frac{S(x,y)}{R} - 1)].$$

- R représente la valeur maximale de l'écart-type dans un document en niveau de gris (R = 128).
- k est un paramètre qui prend une valeur positive dans l'intervalle [0.2, 0.5].
- mean(x,y) représente la matrice des moyennes locales pour chaque pixel de l'image.
- s(x,y) représente la matrice des écarts-types locaux pour chaque pixel de l'image.

2- Calcul de la moyenne locale de chaque pixel de l'image :

Consiste à faire la somme de tous les pixels sur une fenêtre carrée de taille donnée W centrée sur le pixel et ensuite de diviser par W^2.

Voici la formule suivante pour determiner la moyenne locale de chaque pixel de l'image :

$$mean(x,y) = \frac{1}{W^2} \sum_{i=x-W/2}^{x+W/2} \sum_{j=y-W/2}^{y+W/2} (Image(i,j))$$

3- Calcul de l'écart-type local d'une image :

L'écart-type local de chaque pixel étant égal à la racine carré de la variance locale de chaque pixel, nous allons donc passer par le calcul de la variance locale pour déterminer l'écart-type local.

Le calcul de la variance locale par la première méthode est la technique la plus connue. Le principe consiste à sommer tous les pixels de la fenêtre qui ont été retranchés par la moyenne et élevés à la puissance de deux. Ensuite, il faut diviser le résultat de cette somme par le nombre de pixel dans la fenêtre.

$$S^2(x,y) = \frac{\sum_{i=x-W/2}^{x+W/2} \sum_{j=y-w/2}^{y+w/2} (Image(i,j)) - mean(x,y))^2}{W^2}$$

- W représente la taille de la fenêtre.
- mean(x,y) représente la matrice des moyennes locales de chaque pixel .
- Image(i,j) représente l'image de départ .

Discutions de la méthode Sauvola

la méthode de Sauvola montre plus efficace que la méthode de Niblack dans le cas ou le niveau de gris du texte est proche de 0, et celui du fond est proche de 255. Cependant dans les images ou le niveau de gris des pixels du fond et du texte sont proches, les résultats sont peu satisfaisants.

4.3.2.4 La méthode de Nick

L'avantage majeur de cette méthode est qu'elle améliore considérablement la binarisation des images de page blanches et claires, et dans le cas où l'image présente de faible contraste, en déplaçant vers le bas, le seuil de binarisation [51]. Le calcul du seuil est réalisé comme suit :

$$S = m + k * \sqrt{\frac{(\sum pi^2 - m^2)}{NP}}$$

Tel que : k est le facteur de Niblack et varie entre -0.1 à -0.2 selon les besoins de l'application, m : le niveau de gris moyen, pi : niveau de gris du pixel i et NP est le nombre total de pixels. Khurshid et a1 assument que cette méthode marche très bien pour plusieurs

(si pas tous) types de documents anciens dégradés. Pour évaluer ses performances, ils ont appliqué la méthode globalement et localement sur les images de documents. Dans le cas du test local, la taille de la fenêtre utilisée est 19 X19.

4.3.3 Les avantages et les inconvénients de quleque techniques de binarisation

La binarisation est encore un sujet de recherche très actif. En effet, il est bien souvent nécessaire d'adapter une technique en fonction de la problématique. Le tableau suivant résume les différentes méthodes de binarisation rencontrées en mettant en avant les avantages et les inconvénients de chacune des méthodes. Par ailleurs, il est important de bien séparer les difficultés de binarisation liées à un éclairage non uniforme (ce dont la majorité des techniques traitent correctement) des difficultés liées à un fond non uniforme.

Nom	Année	Type	Principe	Inconvénients
Otsu	1979	Seuillage global	D'après l'histogramme, cherche à maximiser la variance intra-classe du «texte» et du «fond».	Problèmes pour les documents mal éclairés
Bernsen	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 lorsquela fenêtre est centrée sur du fond.
Niblack	1986	Seuillage local	Amélioration de Bernsen : prise en compte de la variance et de la moyenne.	Même problème que Bernsen apparition de bruit sur les zones uniformes.
Sauvola	2000	Seuillage local	Insère des constantes dans la méthode de Bernsen afin d'améliorer la méthode sur les zones uniformes.	Les constantes à ajuster empêchent la méthode de traiter parfaitement des documents non uniformes.
Nick		Seuillage local	améliore considérablement la binarisation des images de page blanches et claires, et dans le cas où l'image présente de faible contraste, en déplaçant vers le bas.	Même problème que sauvola, Les constantes à ajuster empêchent la méthode de traiter parfaitement des documents non uniformes.

FIGURE 4.8 – Différentes techniques de binarisation

4.3.4 Méthode de binarisation récent

Zemouri.E-T et al (2014) [52], ont proposé l'utilisation de la Transformation Contourlet pour évaluer la qualité du document historique dégradé. Pour faciliter la binarisation, ils améliorent tout d'abord la qualité de l'image du document par l'application de la « Contourlet Transform », afin de sélectionner les coefficients significatifs. Après la reconstruction, une méthode locale de binarisation est utilisé pour séparer les pixels de texte des pixels de l'arrière-plan de l'image traitée Cette méthode est évaluée sur l'ensemble de données d'analyse comparative utilisé dans l'international « Document Image Binarization Contest » (Dibco 2009/2011 et H-Dibco 2010/2012) en fonction du type de dégradation. Des résultats prometteurs sont obtenus relativement aux méthodes classiques.

Gaceb et al. (2013) [53] ont proposé un systèmes automatique pour la lecteur des documents en utilisant la technologie OCR.et utilise un seuillage local afin de le détecter les pixels de premier plan, fond ou pixel ambigu. La qualité globale de l'image est donc prévisible à partir de la densité des pixels dégradés. Si elle est considérée comme dégradée, ils appliquent une deuxième séparation sur les pixels ambigus pour les séparer.

Rabeux et al. (2013) [54] ont proposé une approche de binarisation sur une 'image en fonction de son état de dégradation. Ils ont proposée pour caractériser la dégradation d'un document l'image en utilisant des caractéristiques différentes en fonction de l'intensité, la quantité et l'emplacement de la dégradation. Ces fonctionnalités nous permettent de construire des modèles et des algorithmes de prédiction de binarisation qui sont très précis en fonction des valeurs R2 et les valeurs p. Les modèles de prévision sont utilisés pour choisir le meilleur algorithme de binarisation d'une image de document donnée. Cette stratégie améliore la binarisation de l'ensemble de données.

Wagdy et al. (2013) [55] ont proposé que la modification de Niveau de gris de l'image ou la couleur dans l'image binaire est l'étape principale de la reconnaissance optique des caractères (OCR) et des systèmes d'analyse de document. La plupart des méthodes de binarisation précédentes, le seuillage locale consomme plus de temps. Cependant Ils présentent un nettoyage efficace de l'image et le seuil global rapide et génère des résultats de haute qualité.

Smithet al. (2012) [56] ont discuté la binarisation de l'image a un grand résultat sur le reste du processus d'analyse d'images documents dans la reconnaissance des caractères. La préférence de terrain de binarisation vérité affecte la la conception d'algorithmes de binarisation, soit directement si la conception est par algorithme automatisé tente de contester la réalité de terrain fourni, ou indirectement si les concepteurs humains ajuster leurs conceptions pour exécuter mieux sur les données fournies. Trois variations de pixel précis ont été utilisés pour former un binarisation classificateur. La performance peut varier considérablement selon le choix de la réalité de terrain, qui peut manipuler le choix de conception de binarisation.

4.4 Le Filtrage

Nous allons intéressé à du filtrage spatial : c'est-à-dire un filtrage qui s'applique sur un voisinage d'un pixel dans une image. Parmi les différents types de filtrage, certains sont linéaires, s'exprimant sous forme de convolution, d'autres sont non-linéaires (filtrage conservatif, filtrage médian,...). Les filtres peuvent effectuer plusieurs types d'opérations comme du lissage ou du rehaussement de contours.

Le principe du filtrage est de modifier la valeur des pixels afin d'améliorer la qualité visuelle de l'image et d'obtenir une image proche à la réalité qui aurait pu être obtenue si le système de capture était parfait.

4.4.1 Principe général des filtres

Pour chaque pixel, le filtre utilise les valeurs des pixels voisins pour calculer la valeur finale du pixel.

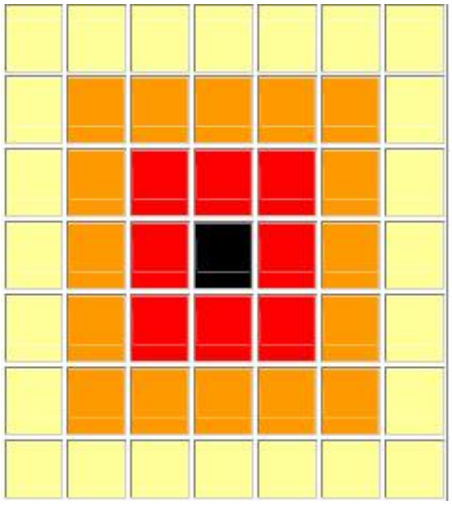


FIGURE 4.9 – Caractréstique d'un filtre

Dans l'exemple ci-dessus le voisinage du pixel central est :
- de 3x3 (rayon 1) si on considère les pixels rouges.
- de 5x5 (rayon 2) si on considère aussi les pixels oranges
- de 7x7 (rayon 3) si on considère également les pixels jaunes.

Un filtre est donc caractérisé par :

1. la forme du voisinage (généralement un carré centré sur le pixel)
2. la taille (ou rayon) du voisinage.
3. l'algorithme de calcul de la valeur finale.

4.4.2 Convolution

Pour mettre en œuvre un filtrage avec des filtres linéaire, on utilise un opérateur mathématique nommé convolution (noté $\otimes$) qu'on utilise pour multiplier des matrices déférentes entre elles. comme une image numérique est en quelque sorte une carte de pixels, et que chaque pixel peut être identifier par ces coordonnées x et y, on peut lui affecter une valeur liée a sa luminosité en utilisant une sorte de tableau de n colonnes et m ligne. Une convolution est un traitement d'une matrice d'image par une autre appelée matrice de convolution ou « noyau ».

L'expression general de la convolution d'une image I1 par un noyau N de taille k*k est :[42]

$$I2(x,y) = \sum_{i=0}^{K-1} \sum_{j=0}^{K-1} I1(x+i-\frac{K}{2}, y+j-\frac{k}{2}) * N(i,j)$$

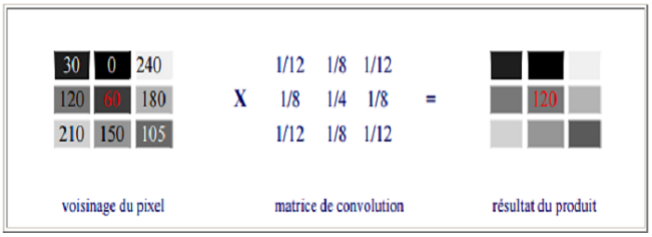


FIGURE 4.10 – Matrice de Convolution[42]

4.4.3 Filtrage Linéaire

Un filtre linéaire transforme un ensemble de données d'entrée en un ensemble de données de sortie selon l'opération « convolution ».

4.4.3.1 Filtre passe-haut

Un filtre passe-haut est un filtre qui laisse passer les hautes fréquences et qui atténue les basses fréquences, c'est-à-dire les fréquences inférieures à la fréquence de coupure. Il pourrait également être appelé filtre coupe-bas,ainsi que, il améliore le contraste. Toutefois, il produit des effets secondaires

- Augmentation du bruit : dans les images avec un rapport Signal/ Bruit faible, le filtre augmente le bruit granuleux dans l'image.
- Effet de bord : il est possible que sur les bords de l'image apparaisse un cadre.[44]

0	-1	0
-1	5	-1
0	-1	0

FIGURE 4.11 – Masque de convolution passe-haut

4.4.3.2 Filtre passe-bas (lissage)

Un filtre passe-bas est un filtre qui laisse passer les basses fréquences et qui atténue les hautes fréquences, c'est-à-dire les fréquences supérieures à la fréquence de coupure. Il

pourrait également être appelé filtre coupe-haut. Le filtre passe-bas est l'inverse du filtre passe-haut et ces deux filtres combinés forment un filtre passe-bande.

1	1	1
1	4	1
1	1	1

FIGURE 4.12 – Masque de convolution passe-bas

L'intensité d'un pixel est transformée en fonction des intensités de voisinage de pixel courant [43]. Voir la figure ci-dessous :

FIGURE 4.13 – Exemple de filtrage par filtre passe-bas.

4.4.3.3 Filtre moyenneur

Le filtre moyenneur est un filtre passe-bas permettant ainsi d'éliminer les hautes fréquences, correspondant au bruit. Son inconvénient est qu'il élimine également les hautes fréquences correspondant aux détails de l'image : il rend ainsi l'image moins bruitée mais plus floue

Principe : Le niveau de gris du pixel central est remplacé par la moyenne des niveaux de gris des pixels environnants.

Voici le masque 3×3 :

FIGURE 4.14 – Le masque d'un filtrage moyen

La figure suivante montre un exemple de calcul de la moyennes :

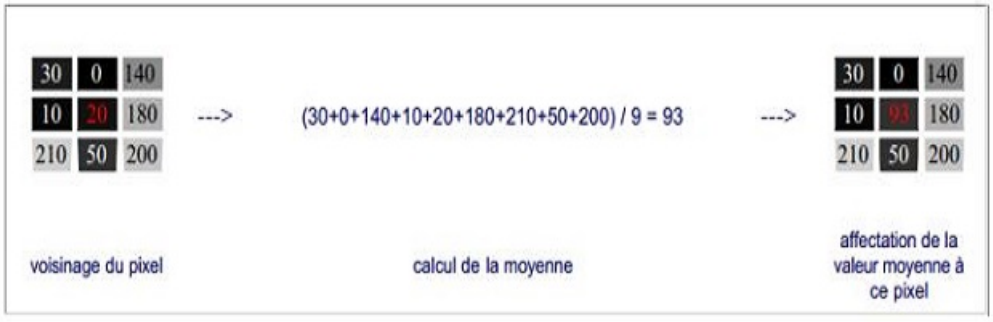


FIGURE 4.15 – Un exemple d'un filtrage moyen

Les effets du filtre moyenneur varient avec la taille du noyau, 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. il rend ainsi l'image moins bruitée mais plus floue. [45].

Voir la figure suivante :

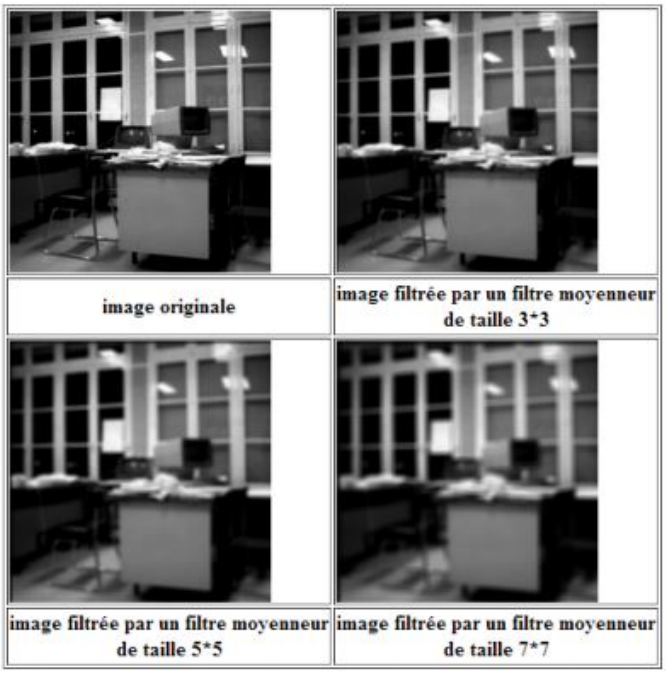


FIGURE 4.16 – Resultat d'un filtre moyen avec différent fenétre.[45]

L'effet du filtre augmente avec la taille de son masque. Les contours et les détails fins sont cependant mieux conservés qu'avec le filtre moyen.

4.4.3.4 Filtre gaussien

C'est également un filtre passe-bas. Une gaussienne à deux dimensions est donnée par l'expression suivante

$$g(x,y) = \frac{1}{2\Pi\sigma^2} exp(-\frac{x^2+y^2}{2\sigma^2})$$

Si par exemple σ =0.8 on a le filtre 3 x 3 suivant

$$\frac{1}{2\Pi(0.8)^2}\exp\left(-\frac{(-1)^2+(-1)^2}{2\,(0.8)^2}\right)=\frac{1}{16}$$

$$\begin{array}{|c|c|c|}\hline G(-1,-1) & G(0,-1) & G(1,-1)\\ \hline G(-1,0) & G(0,0) & G(1,0)\\ \hline G(-1,1) & G(0,1) & G(1,1)\\ \hline\end{array} \simeq \frac{1}{16}\cdot\begin{array}{|c|c|c|}\hline 1 & 2 & 1\\ \hline 2 & 4 & 2\\ \hline 1 & 2 & 1\\ \hline\end{array}$$

FIGURE 4.17 – Un exemple pour calculer la matrice de convolution gausienne avec σ =0.8 on a le filtre 3 x 3

Le filtre gaussien donne un meilleur lissage et une meilleure réduction du bruit que le filtre moyenne.

En général un filtre gaussien avec $\sigma < 1$ est utilisé pour réduire le bruit, et si $\sigma > 1$ c'est dans le but de fabriquer une image qu'on va utiliser pour faire un « masque flou » personnalisé. Il faut noter que plus σ est grand, plus le flou appliqué à l'image sera marqué.

4.4.3.5 les filtres adaptatifs

Le fiLtre adaptatis est utilisés pour diminuer le bruit tout en préservant les contours et en améliorant la qualité de l'image en lissant l'arrière plan mais ils ont aussi des problèmes comme la génération du ou et un lissage des contours lorsque le bruit est trop présent. Le filtre de Wiener, par exemple, a subi des améliorations dans divers travaux selon les besoins, certains ont réussit à réaliser un filtre de Wiener passe-bas adaptatif qui s'adapte au voisinage de chaque pixel à l'aide des statistiques pour filtrer des images de documents en niveau de gris. Une telle démarche parait intéressante dans la mesure où nous essayons de partir sur le même principe de s'adapter aux besoins pour améliorer des méthodes existantes sans oublier que notre traitement se focalisera sur les images de document en couleurs,

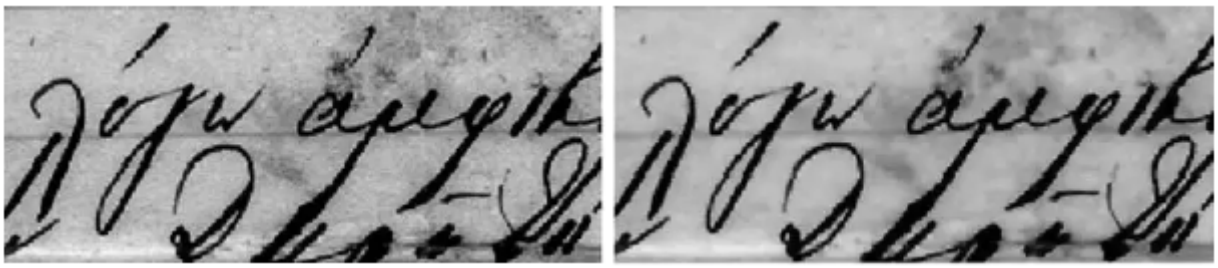
Une adaptation du filtre de Wiener a été conçue pour se comporter comme un filtre passe-bas sur les zones uniformes tout en conservant les discontinuités. La modification du filtre permet de le rendre adaptatif à l'image. Le filtre peut s'exprimer ainsi :

$I(x,y) = \mu + ((\sigma^2 - v^2) * (I_s(x,y) - \mu))/\sigma^2$

Avec :

- I : image filtrée.
- μ : moyenne des valeurs des pixels dans une fenêtre de taille $N \times N$ centrée en (x,y).
- σ : variance des valeurs des pixels dans une fenêtre de taille $N \times N$ centrée en (x,y).
- υ : moyenne de la variance des pixels dans une fenêtre de taille $N \times N$ centrée en (x,y).
- I_s : image source.

Bien que ce filtre améliore la qualité des documents, il ne se comporte pas bien lorsque l'image est fortement bruitée. Le filtre est cependant utilisé par des technique de binarisation, dont [50] car il est tout particulièrement adapté à la restauration de documents. Il permet d'augmenter le contraste entre le texte et le fond tout en lissant le fond. De plus, il se comporte aussi bien sur des documents manuscrits que sur des documents imprimés, car aucune hypothèse n'est faite sur la forme des discontinuités.



(a) (b)

FIGURE 4.18 – Exemple de filtres. (a) Image d'origine. (b) Image après filtrage de Wiener adaptatif.

4.4.4 Filtrage non-linéaire

Le filtrage local est dit non linéaire si nous ne pouvons pas exprimer un filtre par une combinaison linéaire. Ces filtres sont plus complexes à mettre en œuvre que les filtres linéaires. Cependant leurs résultats obtenus sont très souvent de meilleure qualité que ceux obtenus par les filtres linéaires. L'un des filtres non-linéaire les plus connus est le filtre médian, qui prend la valeur médiane des niveaux de gris de voisinage (3x3, 5x5 ...etc.). Ce type de filtre permet d'avoir un meilleur contraste de l'image par rapport au filtres gaussien en assurant une bonne réduction de bruit. D'autres filtres existent aussi comme les filtres morphologiques, filtres FAS, filtres de Nasgao ...etc [46].

4.4.4.1 Le filtrage médian

Le filtre médian consiste à remplacer un pixel par la médiane de ses voisins. Ainsi, même si plusieurs pixels voisins sont bruités, on peut corriger le pixel courant. Ce filtre induit cependant un lissage puisque même des pixels corrects peuvent être modifiés. De plus, ce filtrage est plus coûteux car nécessite d'effectuer un tri des voisins pour chaque pixel. Plus le voisinage considéré est plus, plus l'algorithme sera coûteux. On pensera donc, lors de l'implémentation, à utiliser un algorithme de tri rapide tel que le quick sort. Le filtre médian permet d'obtenir de bons résultats sur du bruit poivre et sel[46].

La figure suivante montre un exemple de calcul de la valeur médiane d'un voisinage d'un pixel :

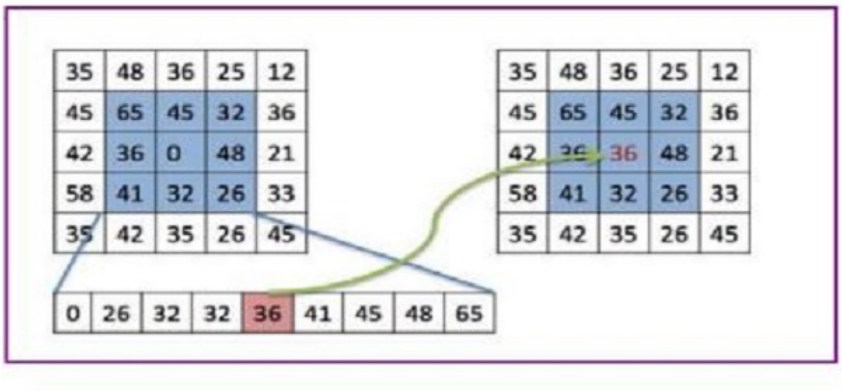


FIGURE 4.19 – Étapes de filtre médian.

4.4.5 Filtres sélectifs

principe de ce filltre est de combiner les avantages des filtres médian et convolutif tout en se débarrassant de leurs défauts. L'idée simple consiste à faire une moyenne gaussienne ou rectangulaire mais sur les pixels de couleurs proches dans le voisinage. Il y a donc deux paramètres : la taille de la fenêtre et la distance de niveau de gris au delà de laquelle on ne compte pas le pixel dans le voisinage. Cet algorithme s'avère être très efficace : robuste et rapide et il préserve les contours.

4.5 Bilan du chapitre

Pour conclure, cette étude bibliographique nous a permis de retracer les différents travaux de prétraitement des images textuelles et de souligner la grande diversité des problématiques qui en découlent. Nous avons constaté que ces dernières présentent différents niveaux de dificultés que les auteurs ont essayé de résoudre. Quelques soit le niveau de difficulté qui s'impose, l'objectif s'agit de mettre en évidence l'information utile contenue dans l'image et de réduire, voire supprimer, toute sorte d'information inutile en fonction des dégradations existantes (bruit, flou, ombre, taches, papier, etc.). Nous avons vu que certaines méthodes de prétraitement reposent sur l'espace de représentation de l'image . Les résultats insusants de certaines approches de prétraitement ne dépendent pas seulement de la technique ou la démarche suivie, mais elle dépend aussi fortement de bon choix de mécanisme (les paramètres) de prétraitement et des espaces de représentation de l'image.

4.6 Conclusion

Dans ce chapitre nous avons présenté ce qu'est un prétraitement aussi ce qu'est un ensemble de prétraitement pouvant être utilisé pour améliorer la qualité des images de documents .Nous avons sélectionné quelque uns de ces prétraitement pour les utiliser dans notre application, dont nous allons détailler la conception dans le chapitre suivant.

Chapitre 5

Conception

5.1 Introduction

Le but de notre projet est la réalisation d'une application utilisant Tesseract pour la reconnaissance de caractère, nous allons intégrer des prétraitements pour améliorer les résultats. Dans ce qui suit nous présentons les différentes étapes suivies par la conception de notre solution.

5.2 Logique de notre approche

Lors de l'utilisation de notre system, au lieu d'avoir recoure directement a tesseract l'utilisation pourra utiliser des prétraitements pour améliorer la qualité d'image avant de passer à la reconnaissance de caractère. L'exécution de notre système suivra donc l'enchaînement suivant :

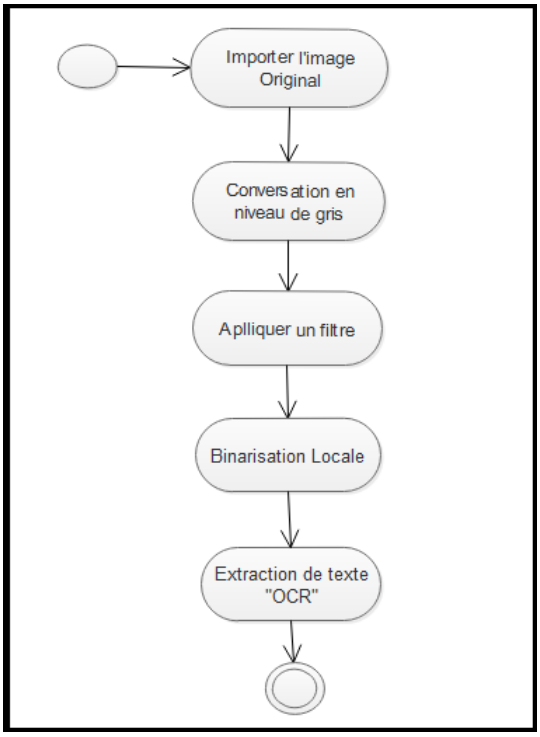


FIGURE 5.1 – Schéma qui résume l'approche de prétraitement proposée.

5.3 Les étapes de l'approche proposé

5.3.1 Conversion de l'image en couleur vers une image en niveaux de gris

Dans une image numérique codée en RGB, la représentation en niveau de gris correspond à l'égalité des intensités des trois composantes de l'intensité du pixel. Pour convertir une image couleur en niveau de gris il faut transformer, pour chaque pixel, les trois valeurs représentant les niveaux de rouge, de vert et de bleu, en une seule valeur représentant l'intensité lumineuse. Pour afficher une image en niveaux de gris sur un support qui prend un format d'image en couleurs, les trois valeurs (r, v, b) sont égales. Dans la recommandation 601 de L'UIT (Union internationale des télécommunications), la transformation est donnée par la formule suivante :

$Niveaudegris(Luminence) = 0,299 \times r + 0,587 \times v + 0,114 \times b.$

La somme des 3 coefficients vaut 1. Dans sa recommandation 709, qui concerne les couleurs vraies ou naturelles :

$Niveaudegris(Luminence) = 0,2126 \times r + 0,7152 \times v + 0,0722 \times b.$

Ces formules rendent compte de la manière dont l'œil humain perçoit les trois composantes (r,v,b) pour la synthèse des couleurs.Dans notre application on utilise la recom-

mandation 601.

5.3.2 Appliquer des filtres

Dans notre projet, nous avons choisi d'utiliser comme première étape dans l'amélioration de la qualité de l'image le filtrage de cette dernière. Les filtres implémentés dans notre système sont les suivants :

Le filtre Gaussien : Le filtre gaussien est un bon exemple pour les performances qu'on peut obtenir avec un filtre linéaire à réponse impulsionnelle finie. Le gros avantage de ce filtre, c'est leur facilité de conception et d'implémentation.

Le filtre médian : est un filtre efficace contre du bruit poivre et sel dans des images à niveaux de gris, et l'une de ses propriétés fondamentales, est qu'il ne crée pas de nouvelles valeurs de niveaux de gris dans l'image.

5.3.3 Binarisation (seuillage) de l'image

La binarisation est la dernière étape avant la reconnaissance de l'écriture à l'aide de l'OCR. elle fournit à ce dernier une image binaire ou le texte est isolé de son arrière-plan. pour la binarisation nous avons utilisé la technique locale de Sauvola qui offre le meilleur résultat sur les images de documents particulièrement avec les images qui ont une mauvaise qualité (sombre, lumineuse).

5.3.4 Extraction du texte dans l'image à l'aide de Tesseract-ocr

C'est la dernière étape de notre système elle applique après le prétraitement sur l'image basé sur l'extraction du texte dans l'image à l'aide d'un outil s'appelle Tesseract-ocr.

5.4 Présentation UML

Nous essayerons dans ce qui suit de décrire les différentes fonctionnalités qu'offrira notre système à l'aide des différents types de diagrammes UML

5.4.1 Diagramme de cas d'utilisation

Un cas d'utilisation est un résumé des scénarios pour un but ou une tâche unique. Un acteur est la personne ou l'objet qui engage les événements impliqués dans cette tâche,

Notre diagramme de cas d'utilisation sera présenté comme suit :

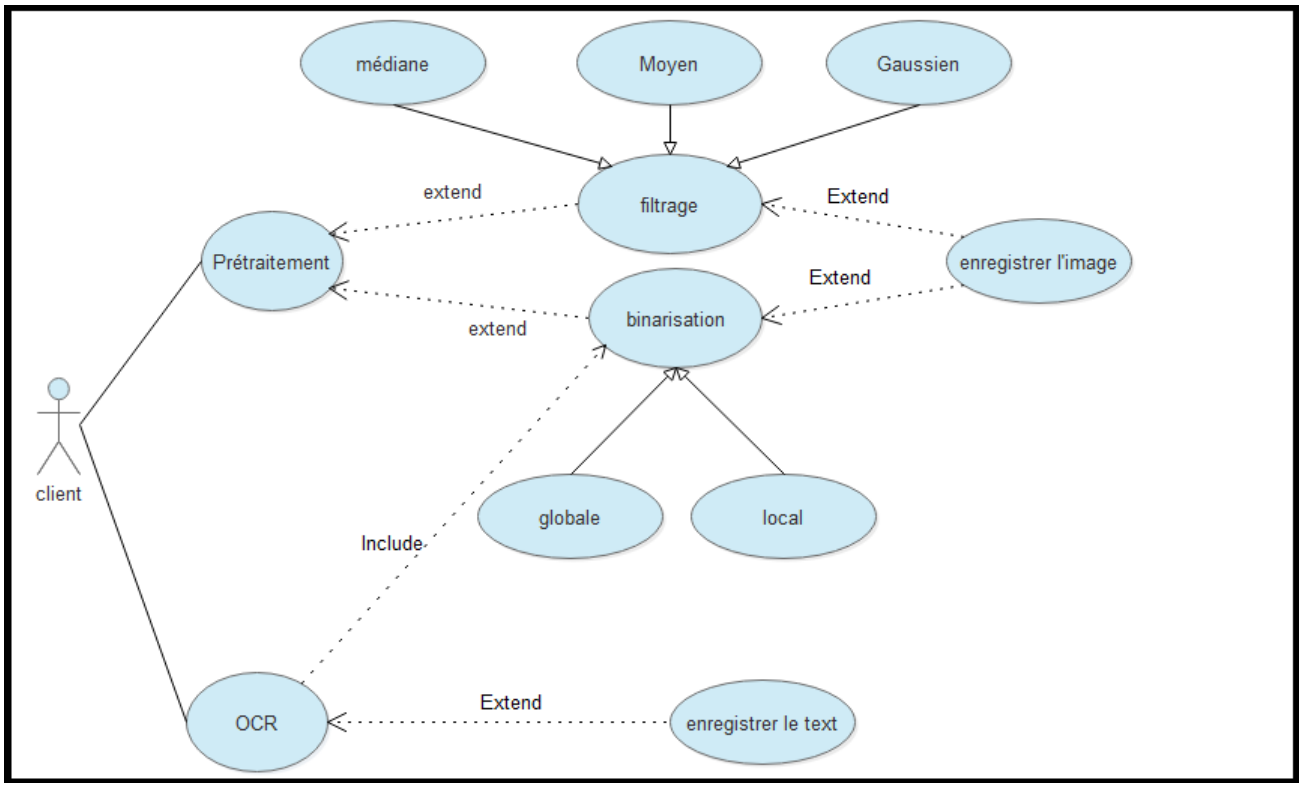


Figure 5.2: Diagramme de cas d'utilisation

5.4.2 Diagramme de classe

Le diagramme de classes permet de définir la structure de toutes les classes qui constituent un système. Une classe est définie en plus de son nom par des attributs et des méthodes.

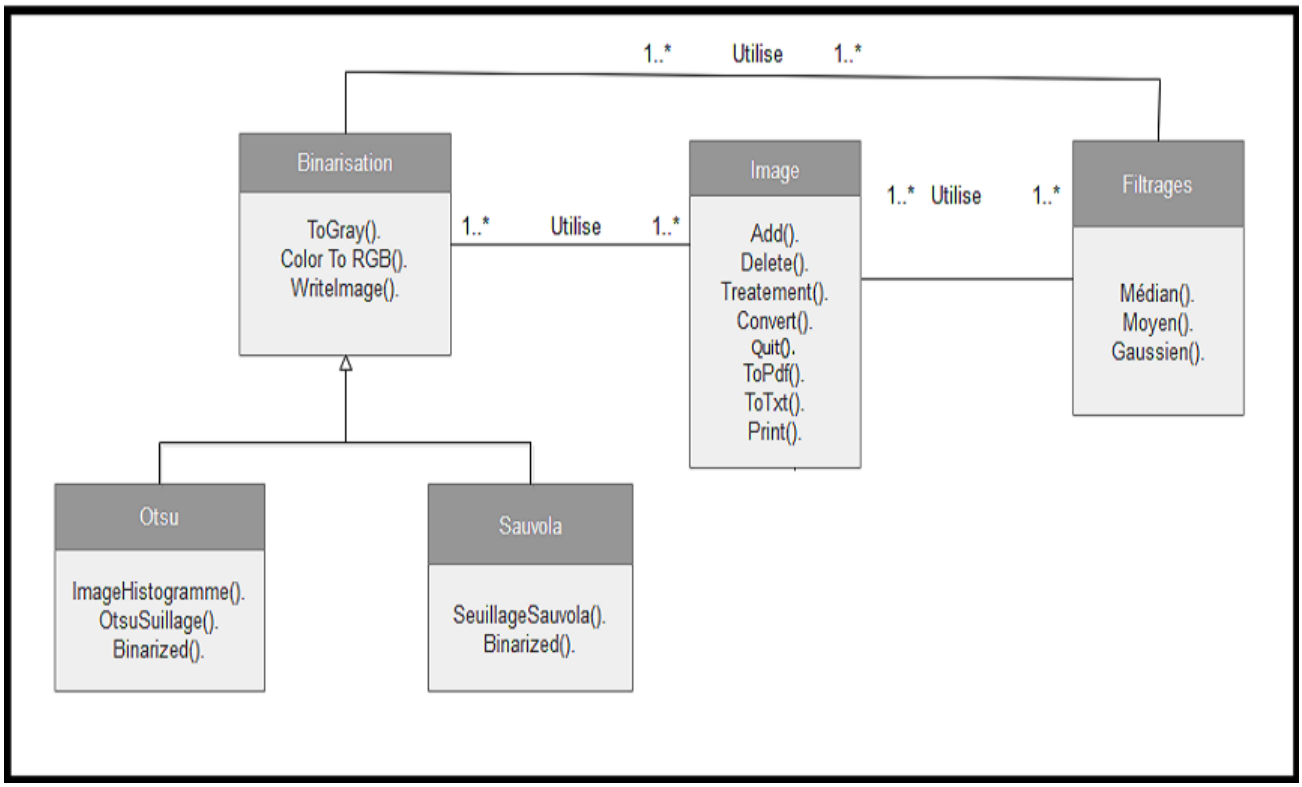


FIGURE 5.3 – Diagramme de Classe

5.5 Conclusion

Dans ce chapitre nous avons illustré la logique de notre approche ainsi que les étapes de cette dernière après nous avons construit le diagramme de cas d'utilisation et le diagramme de classe tout en s'appuyant sur le langage UML.

Le chapitre suivant sera consacré à la partie de la réalisation de l'application java ainsi que les différentes fonctionnalités dont elle dispose avec une etude comparative entre les resultat proposé.

Chapitre 6

Implemantation Et Expérimentation

6.1 Introduction

Ce chapitre est le dernier de notre travail, il sera consacré à la mise en œuvre de notre application en indiquant les outils de développement, les langages de programmation utilisés pour son implémentation. Nous présentons ensuite les résultats de nos tests expérimentales obtenue grâce à notre application.

6.2 Outils et Langages de developpement

6.2.1 language java

Nous avons utilisé le langage JAVA, ce choix est justifié par ses nombreux avantages dont voici quelques uns :

- Il est orienté objet : permet l'encapsulation et l'héritage, qui vont nous aider à bien organiser et structurer l'application.

- La portabilité : car le compilateur java produit un code intermédiaire qui sera interprété par une JVM (Java Virtual Machine).

– Il dispose d'une bibliothèque extensible avec des classes très riches.

FIGURE 6.1 – Logo de java

6.2.2 Standard Widget Toolkit (SWT)

est une bibliothèque graphique libre pour Java, initiée par IBM. SWT n'est pas un standard Java reconnu par le JCP. Cette bibliothèque se compose d'une bibliothèque de composants graphiques (texte, label, bouton, panel), des utilitaires nécessaires pour développer une interface graphique en Java, et d'une implémentation native spécifique à chaque système d'exploitation qui sera utilisée à l'exécution du programme.

La deuxième partie de SWT n'est en fait qu'une ré-encapsulation des composants natifs de système (Win32 pour Windows, GTK ou Motif pour Linux). Plusieurs projets travaillent aujourd'hui sur une implémentation utilisant les composants de Swing.

L'environnement de développement libre Eclipse, commandité lui aussi par IBM, repose sur cette architecture[48].

6.2.3 Environnement de développement JAVA(Eclipse)

Eclipse est un environnement de développement intégré libre extensible, universel et polyvalent, permettant de créer des projets de développement mettant en œuvre n'importe quel langage de programmation. Eclipse IDE est principalement écrit en Java (à l'aide de la bibliothèque graphique SWT, d'IBM), et ce langage, grâce à des bibliothèques spécifiques, est également utilisé pour écrire des extensions[49].

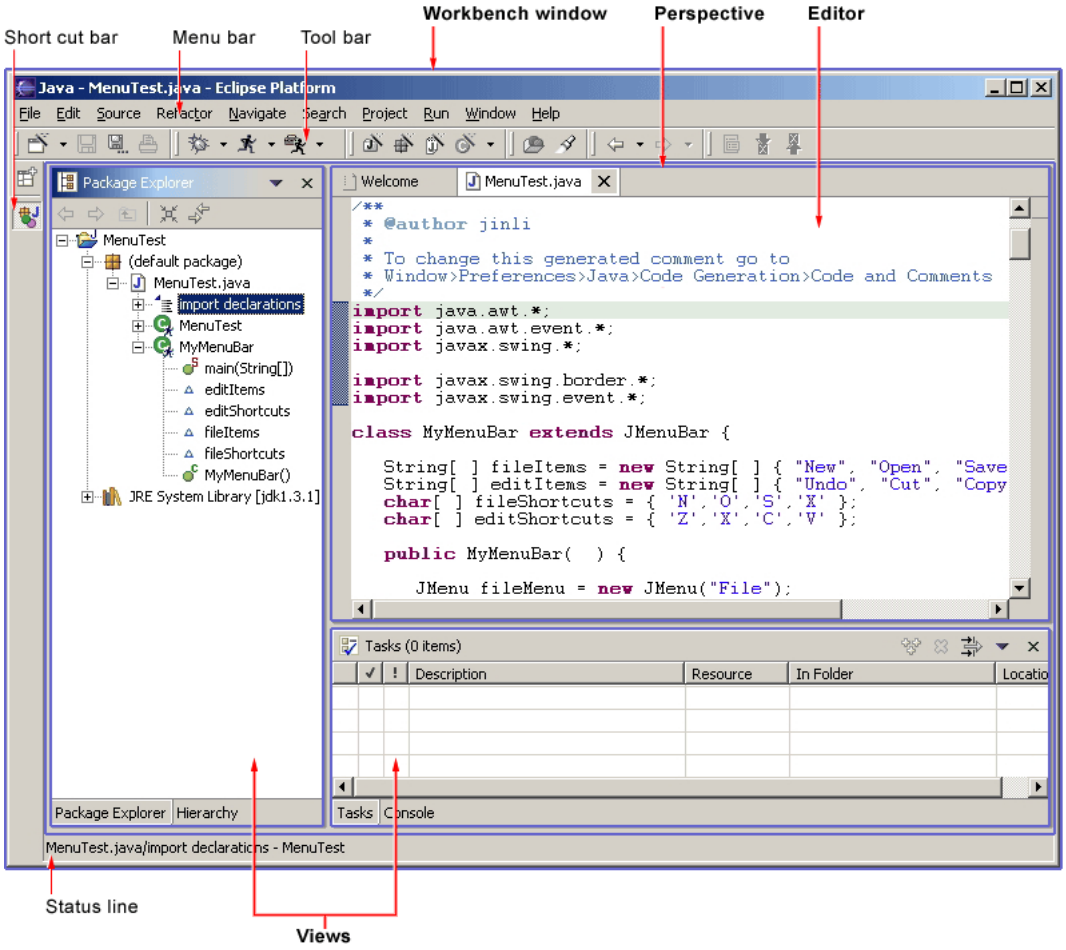


FIGURE 6.2 – Interface Eclipse

6.2.4 Moteur de reconnaissance tesseract-ocr

6.2.4.1 Déffinition

Tesseract est un logiciel de reconnaissance optique de caractères sous licence Apache. Conçu par les ingénieurs de Hewlett Packard de 1985 à 1995, son développement est abandonné pendant les dix années suivantes , en 2005, les sources du logiciel sont libérées sous licence Apache et le logiciel est actuellement développé par Google. Initialement limité aux caractères ASCII, il supporte parfaitement les caractères UTF-8 et reconnait maintenant 40 langues.[47]

6.2.4.2 Le fonctionnement de tesseract-ocr

La analyse de mise en page pour Tesseract a été conçue. dès le début pour être indépendant du langage, mais le reste du moteur a été développé pour l'anglais, sans beaucoup de réflexion sur la façon dont il pourrait fonctionner pour les autres langues. Après avoir noté

que les moteurs commerciaux à l'époque étaient strictement pour texte noir sur blanc, parmi les objectifs de la conception originale du Tesseract est qu'il devrait reconnaître un texte blanc sur noir) aussi facilement que le noir sur blanc.[47]

6.3 Présentation les interfaces de l'application

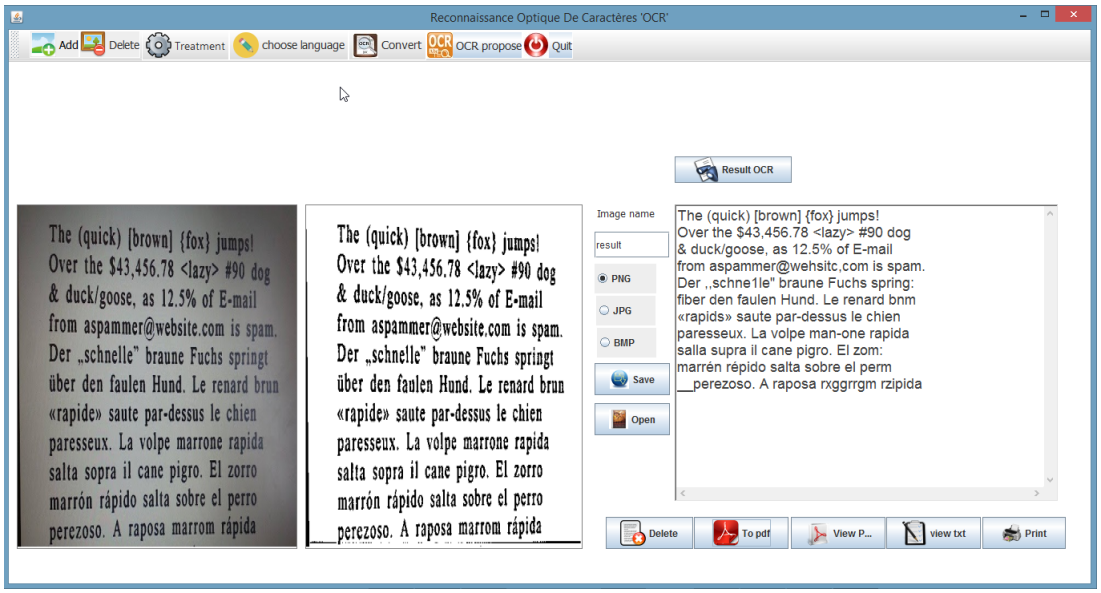


FIGURE 6.3 – interface de notre application

Traitement :

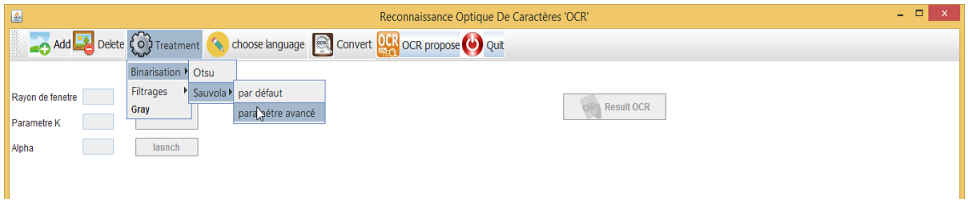


FIGURE 6.4 – Pour les binarisations

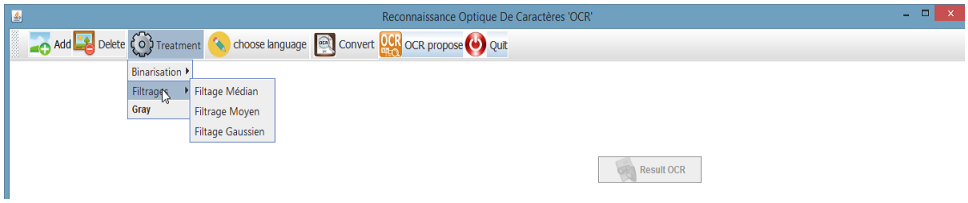


FIGURE 6.5 – Pour les filtrages

Le choix de la Langue

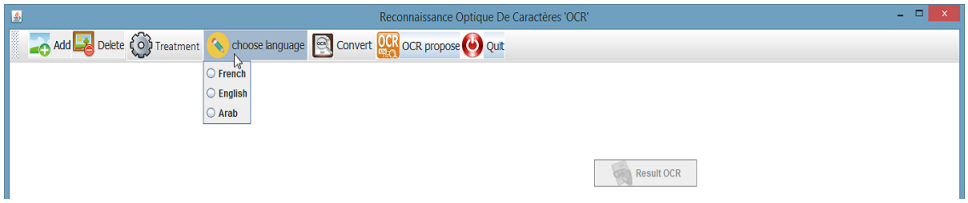


FIGURE 6.6 – Pour choisir la langue de l'ocr

6.3.1 Résultats visuels des approches de prétraitement :

Nous allons vous présenter dans cette partie les résultats des prétraitements développées avec des techniques :

6.3.1.1 Pour les filtres

Filtre Gaussien

FIGURE 6.7 – Résultat du filtre gaussien

Filtre Moyenneur

FIGURE 6.8 – Résultat du filtre moyenneur

Filtre Median

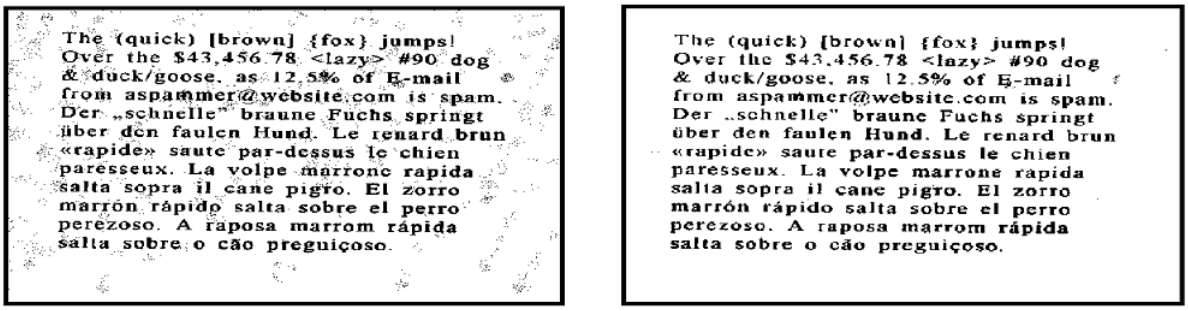


FIGURE 6.9 – Résultat du filtre Médian

6.3.1.2 pour les binarisation

la binarisation Otsu

Nous allons presenté le resultat de la binarisation globale Otsu.

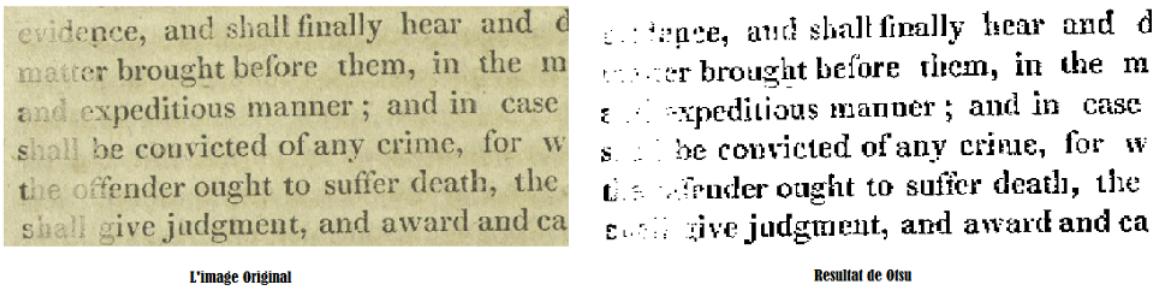


FIGURE 6.10 – Résultat de la méthode d'Otsu

la binarisation Sauvola

Nous allons presenter le resultat de la binarisation locale Sauvola avec quelque exemple de k et la taille de rayon

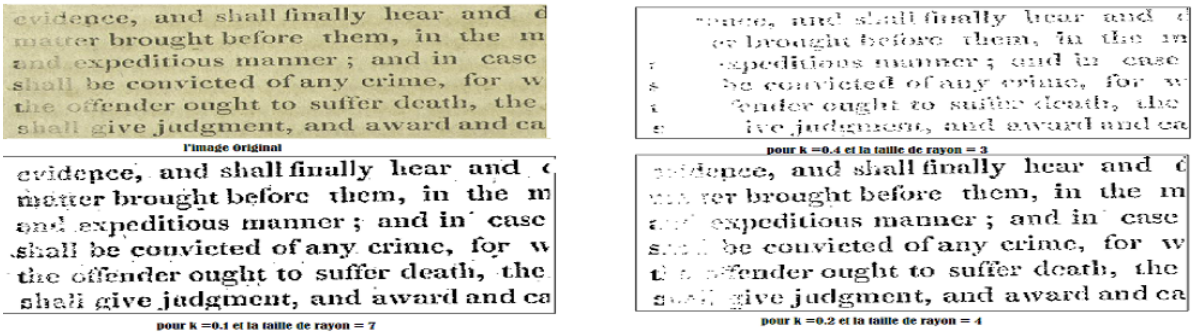


FIGURE 6.11 – Résultat de la méthode Sauvola

Remarque :

Nous pouvons remarquer que le résultat de Sauvola est meilleur par rapport a Otsu si nous choisissons la taille de rayon et le parametre K correctement.

6.4 Etude Comaprative

Cette étude désigne la comparaison des résultats obtenu à partir de tresseract sans prétraitements sur l'image et avec la méthode principale de notre étude Sauvola plus les différents prétraitements ainsi que les différents états et le niveau de qualité des images.

Nous allons présenter quelque image et leurs résultat de tesseract avec différent état :

Exemple N°1

Image Net N°1

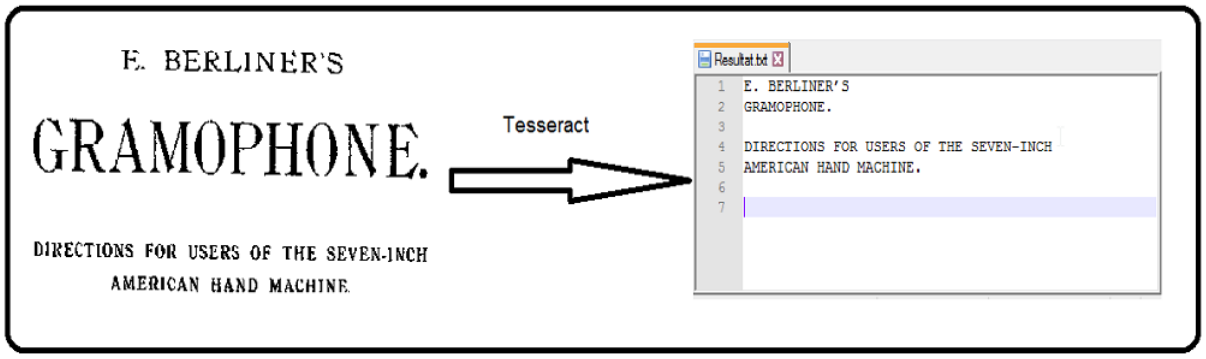


FIGURE 6.12 – Resultat de tesseract avec image net

Image Fortement Bruité N°1

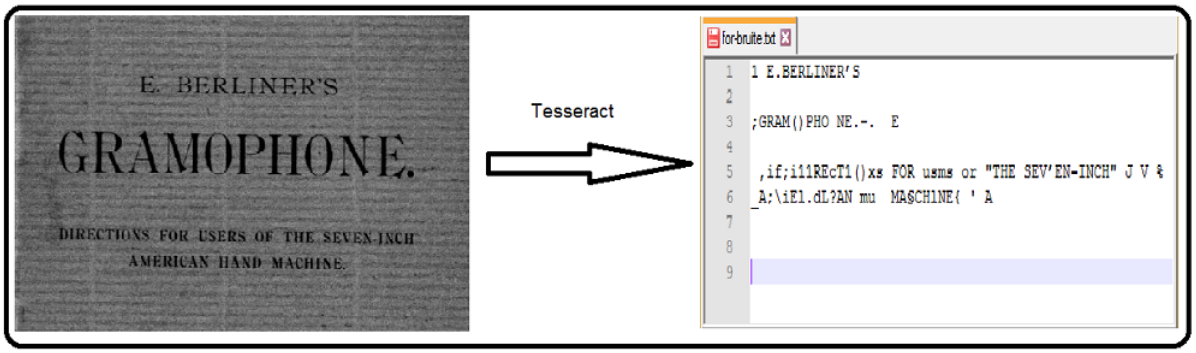


FIGURE 6.13 – Resultat de tesseract avec image fortement bruité

Image Sombre N°1

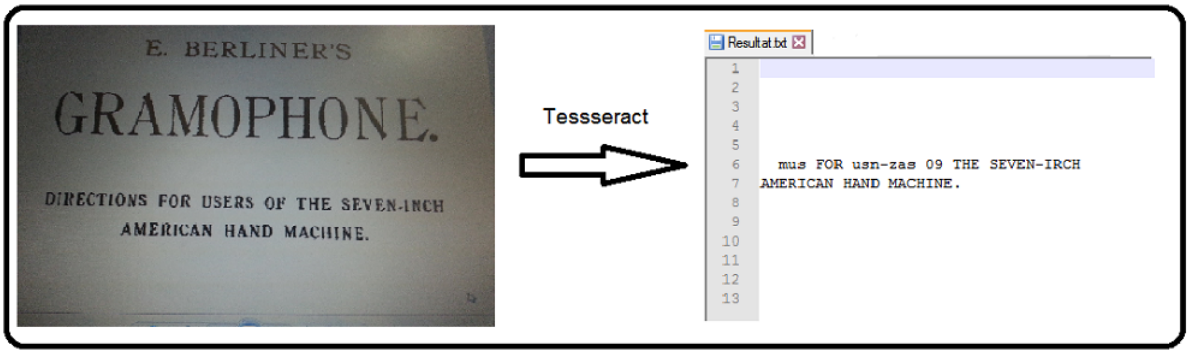


FIGURE 6.14 – Resultat de tesseract avec image sombre

6.4.1 Tableau de comparaison N°1

Nous allons présenter cette étude dans le tableau suivant :

Ce tebleau presente le pourcentage des caractères reconnus par tesseract dans les différents cas possible :

l'image originale contient 73 caractères et 12 mots

Image / Prétraitement	Image Net		Image Fortement Bruité		Image Sombre	
	% Caractère	% Mot	% Caractère	% Mot	% Caractère	% Mot
OCR seul	100%	100%	38.35%	25%	50%	45.20
OCR+ Sauvola	100%	100%	41.09%	50 %	93,15%	66,66%
OCR+ Otsu	100%	100%	52.05%	25%	50,68%	75%
OCR+ Moyen +sauvola	100%	100%	75.34%	58.33%	34,24%	41,66%
OCR + Médian +sauvola	100%	100%	71.32%	58.33%	75,34%	91,66%
OCR+ Gaussien +sauvola	100%	100%	79%	66.66%	83.56%	83.33%

FIGURE 6.15 – Tableau de Comparaison

Exemple N°2

Image Net N°2

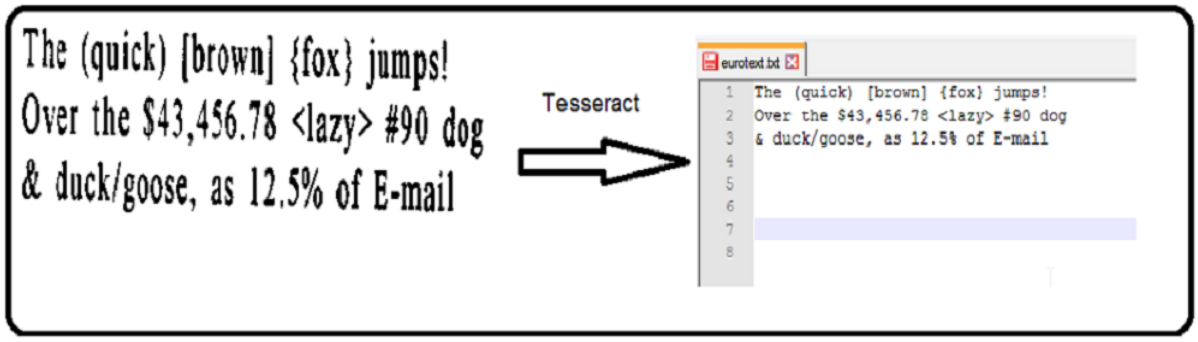


FIGURE 6.16 – Resultat de tesseract avec image net

Image Fortement Bruité N°2

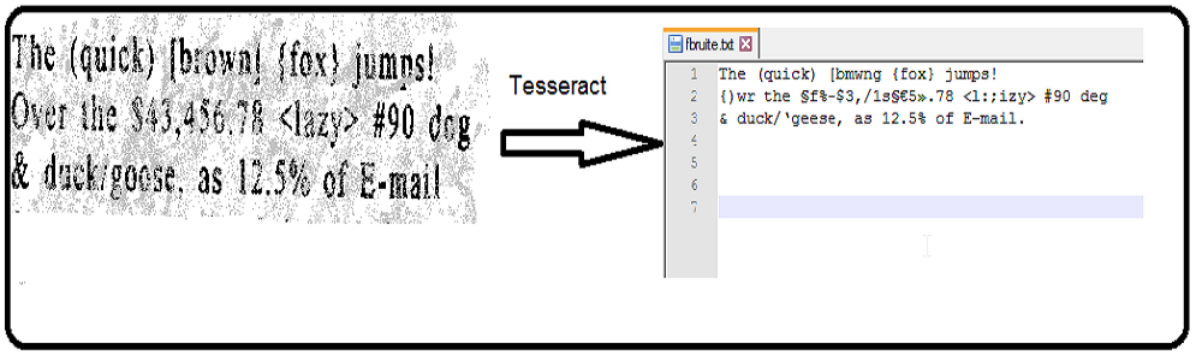


FIGURE 6.17 – Resultat de tesseract avec image fortement bruité

Image Sombre N°2

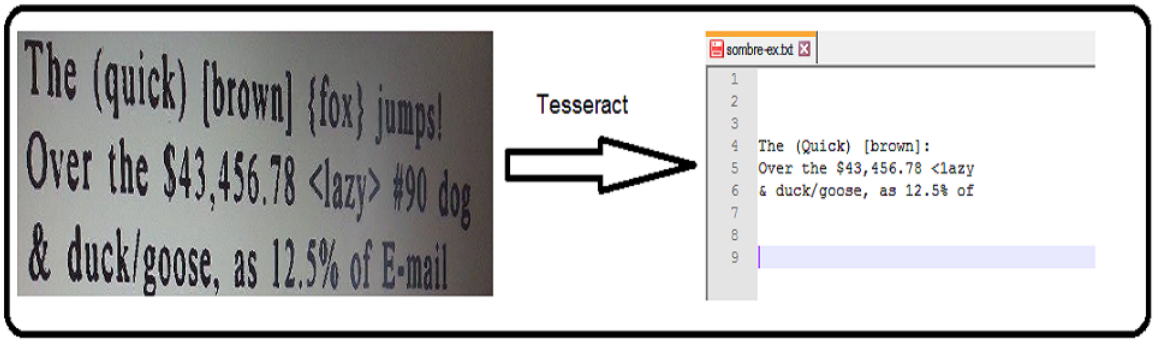


FIGURE 6.18 – Resultat de tesseract avec image sombre

6.4.2 Tableau de comparaison N°2

Nous allons présenter cette étude dans le tableau suivant :

Ce tebleau presente le pourcentage des caractères reconnus par tesseract dans les différents cas possible :

l'image originale contient 83 caractères et 20 mots

Image / Prétraitement	Image Net		Image Fortement Bruité		Image Sombre	
	% Caractère	% Mot	% Caractère	% Mot	% Caractère	% Mot
OCR seul	100%	100%	49%	25%	57.83%	55%
OCR+ Sauvola	100%	100%	68.67%	50%	85%	90.36%
OCR+ Otsu	100%	100%	65%	55%	81.92%	60 %
OCR+ Moyen +sauvola	100%	100%	100%	100%	93.79%	90 %
OCR + Médian +sauvola	100%	100%	89.15%	90%	78.31%	75%
OCR+ Gaussien +sauvola	100%	100%	97.59%	96%	69.38%	96%

FIGURE 6.19 – Tableau de Comparaison

6.5 Conclusion

Dans ce chapitre nous avons présenté les démarches que nous avons suivre pour le développement de la plateforme de reconnaissance de caractére , les différentes étapes de la réalisation de notre application, et à la fin nous avons présenteté les différents résultats expérimentaux de méthode développée.

Conclusion générale

Le traitement des images dans les documents est une étape importante dans leur analyse. Ce traitement s'opère sur plusieurs niveaux : prétraitement, segmentation, extraction, et la reconnaissance. Toutes ces manipulations permettent notamment d'améliorer la qualité de l'image (nettoyage, restauration), reconnaître les différents éléments graphiques (traits, lignes, symboles,...) et de délimiter les zones de textes (caractères, mots, lignes, paragraphes) et des images. L'objectif donc de notre travail était la conception et le développement d'un système pour la reconnaissance de caractère imprimé.

Afin de bien mener notre projet, nous avons mené une étude bibliogaphique sur le système de reconnaissance de caractère OCR ainsi que une étude sur les prétraitements de l'image afin d'améliorer sa qualité. Cette étude nous a permis à choisir les méthodes les plus utiles pour améliorer la performance de l'OCR pour atteindre à des résultats satisfaisants.

Enfin nous n'avons pas pu aborder quelques concepts dont, nous avons décidés de les mettre comme perspectives de notre travail :

- Améliorer la base de connaissance de tesseract avec de nouveaux fonds et de nouveaux caractères.
- Réduire le temps de traitement et trouver une technique adaptative de prétraitemnt-d'image.
- Etablir une technique de binarisation basée sur la binarisation locale et globale en même temps (binarisation hybride).

Bibliographie

[1]

[1] J. Fruit et, "Outils et méthodes pour le traitement des images par ordinateur " , Université de Villeneuve-la-Garenne,article, France 2000.

[2] A.Christophe, Transmettre et stocker de l'information '' Caractéristiques d'une image numérique : pixellisation, codage RVB et niveaux de gris", octobre 2011.

[4] K. Chakra, "La Compression des Images Fixes par les Approximations Fractales Basée sur la Triangulation de Delaunay et la quantification Vectorielle ", mémoire de fin d'étude, 1999.

[5] R Isdant ,'Traitement numérique de l'image",article,2009.

[6] M chakib, '' Généralités sur le traitement d'images"article,1999.

[7] H. Schahrazed '' SEGMENTATION DE TEXTES EN CARACTERES POUR LA RECONNAISSANCE OPTIQUE DE L'ECRITURE ARABE ",mémoire fin d'étude, 08, Juillet, 2007.

[8]].M.Ogier, Contribution à l'analyse automatique de documents cartographiques : Interprétation de données cadastrales, Thèse de doctorat, Rouen, 1994.

[9] P.T.Wright, "On-line recognition of handwriting", GEC Journal of research, vol 8 n°l, pp 42-48, 1990.

[10] A.Belaid , Y.Belaid , caractéristique d'image : méthodes et applications, Interéditions,1992.

[11] E.Lecolinet, O. Barett : « Cursive word recognition : Methods and strategies ». In NATO/ASI Fundamentals in handwriting recognition, Bonas, France June 21-july 3, 1993.

[12] P.M. Lallican, C. Viarp-Gaudin, S. Knerr : « From off-line to on-line handwriting recognition ». Proc. 7th workshop on frontiers in handwriting recognition, pp. 303-312, Amsterdam 2000.

[13] I.R. Tsang : «Pattern recognition and complex systems». Thèse de doctorat, université d'Anterwerpen, 2000.

[14]] J. Trenkle, A. Gillies, S.Schlosser : « An off-line Arabic recognition system for machine printed documents ». Proc. Of the symposium on document image understanding technology (SDIUT'97), pp. 155-161 1997.

[15] N. Benamara « Utilisation des modèles de Markov cachés planaires en reconnaissance de l'écriture arabe imprimée ». Thèse de doctorat, spécialité Génie Electrique, Université des sciences, des Techniques et de médecine de Tunis II, 1999.

[16]] B. Al-Badr , S.A. Mahmoud : « Survey and bibliography of Arabic optical text recognition ». Signal processing , vol. 41, pp. 49-77, 1995.

[17] A. Amin, H.B. Al-Sadoun , S. Fisher : « Handprinted A rabic character recognition system using an artificial network ». Pattern recognition, vol. 29, No 4, pp. 663-675, 1997.

[18] Shalev Vayness, Catherine Lerouge, « Charte de Traitement "OCR brut et HQ, ALTO" » documentation interne de la Bibliothèque nationale de France, 25/02/2008.

[19] G. Cron – A.Salah – N.Ragot – K.Mohand – T.Paquet ,'' Etat de l'art sur la caractérisation d'un document à OCRiser '' Projet ANR DigiDoc, septembre 2012.

[20] S. Lecoeuche,"Reconnaissance de caractères industriels par application d'un système de réseaux de neurones à boucle de rétroaction " pour obtenir le grade de DOCTEUR DE L'UNIVERSITE. 20 Novembre 1998.

[21] A.&Y.Belaid, Reconnaissance des formes InterEditions, Paris, 1992.

[22]] M.M.M. Fahmy, S.Al Ali : « Automatic recognition of handwritten Arabic characters using their geometrical features ». Studies informatics and control journal (SIC journal), vol. 10, N°2, 2001

[23] A. Belaïd et H. Cecotti ,« La numérisation de documents : Principe et évaluation des performances », Université Nancy 2 – LORIA.

[24] V.RICE S, V.NAGY , T. A.NARTKER , « Optical Charcater Recognition : An illustrated guide to the frontier », Kluwer Academic Publishers, 1999.

[25]] E. Lecolinet. « Segmentation d'images de mots manuscrits : Application à la lecture de chaîne de caractères majuscules alphanumériques et à la lecture de l'écriture manuscrites ». Thèse de doctorat, Université Paris 6, Mars 1990.

[26]] D.Motawa,A. Amin , R.Sabourin., "Segmentation of Arabic cursive script", Proceedings of the ICDAR'97, the 4th Conference on Document Analysis and Recognition, Vol. 2, pp. 625-628, Ulm, Germany, August 1997.

[27] S. Madhvanath, V. Krpasundar, and Venu Govindaraju. Syntactic methodology of pruning large lexicons in cursive script recognition. Pattern Recognition, Vol.34, N°.1, pp.37–46, 2001.

[28] F.Biadsy, SEl-Sana , N.Habash, "Online Arabic handwriting recognition using Hidden Markov Models", Proceeding of IWFHR'06, 10th International Workshop on Frontiers in Handwriting Recognition, pp. 85-90, La Baule, France 2006.

[29] G.Gaillat, Méthodes syntaxiques de la reconnaissance de formes, ENSTA, Paris, 1983.

[30]] J.Kittler, M.Hatef R.Duin & J.Matas, "On combining classifiers", IEEE on pattern analysis and machine intelligence, vol20, n°3, pp 226-239, 1998.

[31] S.AGNE, and M. ROGGER, « Benchmarking of Document Page Segmentation », Part of the IS&T/SPIE Conference on Document Recognition and Retrieval VII, San Jose, California, p. 165-171, January 2000.

[32] D.GACEB « Contributions Au TriAautomatique De Documents Et De Courrier D'entreprises »,INSA De Lyon,Thèse, 2009.

[34]] Yao-Hong Tsai, A New Approach for Image Thresholding under Uneven Lighting Conditions, Computer and Information Science, ICIS 2007. 6th IEEE/ACIS International Conference on, 11-13, July 2007. pp.123-127.

[35] N. Babaguchi and al, Connectionist model binarisation, Pattern Recognition, ICPR 1990. Proceedings, 10th International Conference, 16-21 June 1990, vol.2, pp. 51-56.

[36] A New method for gray-level picture threshold using the entropy of the histogram, Computer Vision, Graphics, and Image Processing, 29, p. 273-285, 1985.

[37] J. BERNSEN. Dynamic thresholding of grey-level images. in Proc. 8th International Conference on Pattern Recognition, p.

[38] A.J.O. TRIER. Goal-directed evaluation of binarization methods, IEEE Transactions on Pattern Analysis and Machine Intelligence, 17(12), p. 1191-1201, 1995.

[39] W. NIBLACK. « An Introduction to Digital Image Processing, Ed. Prentice Hall, Englewood Clis »pp 115-116, 1986.

[40] J. HE, Q. D. M. DO, A. C. DOWNTON, J. H. KIM. A comparison of binarization methods for historical archive documents. International Conference on Document Analysis and Recognition (ICDAR), p. 538-542, 2005.

[43] Y.Sculo « Introduction au traitement d'images Détection de contours et segmentation »,2009.

[44] M. Bergounioux « Quelques méthodes de filtrage en Traitement d'Image »,article, Jan 2011.

[47] R.Smith « An Overview of the Tesseract OCR Engine »,article,2006.

[50] B.Gatos, I.Pratikakis,and S.J.Perantons.« Adaptive degraded document image binarisation Pattern Recogn ». pp317-327,2006

[51] K. KHURSHID, I. SIDDIQI, C. FAURE, N. VINCENT.« Comparison of Niblack inspired Binarization methods for ancient documents. 16th International conference on Document Recognition and Retrieval, USA », 2009.

[52] E. Zemouri, Y. Chibani, and Y. Brik "Restoration based Contourlet Transform for historical document image binarization." ,Multimedia Computing and Systems (ICMCS), 2014 International Conference on IEEE, 2014.

[53] D.Gaceb , F.Lebourgeois, and J. Duong. "Adaptative Smart-Binarization Method : For Images of Business Documents." Document Analysis and Recognition (ICDAR), 2013 12th International Conference on. IEEE, 2013.

[54] V.Rabeux, N.Journet, A.Vialard et J-P.Domenger. "Quality evaluation of ancient digitized documents for binarization prediction." Document Analysis and Recognition (ICDAR), 2013 12th International Conference on. IEEE, 2013.

[55] M.Wagdy, I.Faye, and D.Rohaya. "Fast and efficient document image clean up and binarization based on retinex theory." Signal Processing and its Applications (CSPA), 2013 IEEE 9th International Colloquium on. IEEE, 2013.

[56] B.Smith,H.Elisa , and C.An. "Effect of" ground truth" on image binarization." Document Analysis Systems (DAS), 2012 10th IAPR International Workshop on. IEEE, 2012.

Webographie

[3] http ://tecfaetu.unige.ch/staf/staf-k/benetos/staf13/per1/tache5/resolution.html.consulté le 25/04/2016

[33] https ://sites.google.com/site/lizantchristopher/services/binarisation-1.consulté le 14/05/2016

[41] https ://sites.google.com/site/lizantchristopher/services/otsu.consulté le 14/05/2016

[42] https ://docs.gimp.org/fr/plug-in-convmatrix.html. consulté le 1/05/2016

[45]http ://www.tsi.telecomparistech.fr/pages/enseignement/ressources/beti/filtres_lin_nlin/filt consulté le 23/05/2016

[46] http ://www.unit.eu/cours/videocommunication/filtrage_non-linéaire.consulté le 23/05/2016

[48] https ://fr.wikipedia.org/wiki/Standard_Widget_Toolkit.consulté le 01/06/2016

[49]https ://fr.wikipedia.org/wiki/Eclipse_(projet).consulté le 1/06/2016