République Algérienne Démocratique et Populaire

Ministère de l'enseignement supérieur et de la recherche scientifique

Université de Guelma

Faculté des Mathématiques, d'Informatique et des Sciences de la matière

Mémoire de fin d'étude de Master

Département d'Informatique

Spécialité : ingénierie des médias

Thème :

Recherche de mots dans les images de documents arabes

Présenté par : Hammor Med Jalil

Babou Soumeya

Sous la direction de :

Mme. Mehnaoui Zahra

Juin 2011

Remerciements

Nous remercions en premier lieu Mme Mehnaoui. Nous lui sont reconnaissants pour avoir accepté de rapporter nos travaux et de nous avoir apportés beaucoup de soutien à travers les remarques et critiques qui ont permis de nous guider dans nos travaux, sa disponibilité et ses qualités scientifiques. Nous avons beaucoup appris à ses côtés et nous lui adressons toutes notre gratitude.

Nous remercions aussi Mr Sari Maître de Conférences à l'Université de Annaba pour sa disponibilité, son aide et ses conseils précieux ainsi que Mme Suissi Maître de Conférences à l'Université de Annaba et Dr. Mr Benouareth.

Nos dernières pensées s'adressent à nos familles, à nos parents et nos amis pour leur soutien continu.

Dédicace

A tous nos amis, et à tous ceux qui nous sont chers.

Résumé :

Dans ce travail, nous nous intéressons plus particulièrement à la problématique de navigation et d'accès à des collections de documents manuscrits qui sont des enveloppes postales algériennes. L'accès à ces collections nécessite des stratégies d'indexation et de recherche efficaces. Dans la plupart du temps, les indexes sont crées manuellement. Si cette approche est possible pour un petit nombre de documents, le cout devient très élevé pour des larges collections. L'OCR peut être une alternative pour les documents imprimés ou les documents manuscrits avec un lexique limité. Dés que les documents soient dégradés et avec un lexique plus large, l'OCR devient inefficace, surtout dans le cas des documents arabes qui présentent d'autres difficultés relatifs aux traitements de l'écriture arabe.

Afin d'atteindre les objectives visées, le système proposé regroupe plusieurs traitements issus principalement du domaine de l'analyse de documents : binarisation, segmentation, extraction de contours, codage de Freeman et du domaine de la recherche d'information, notamment la technique de la recherche approximative.

Mots clés : analyse de documents, recherche d'information, reconnaissance de l'écriture arabe, Word-spotting.

Table des matières

Table des figures

Introduction Générale

L'écriture manuscrite est toujours omniprésente dans notre vie quotidienne, et constitue un lien étroit et privilégié entre les hommes car elle leur permet d'échanger de façon naturelle des idées, des informations, des sentiments, etc. Son importance peut être quantifiée de façon permanente, par le volume du courrier manuscrit acheminé chaque jour par la poste, par le nombre de chèques bancaires et postaux traités quotidiennement par les services financiers, de formulaires administratifs remplis à la main, d'ordonnances médicales, etc. Aujourd'hui, malgré l'avènement des nouvelles technologies (ordinateurs, réseaux de communication, assistants personnels (PDA), téléphones mobiles, etc.), l'écriture manuscrite reste un moyen de communication incontournable.

L'écriture est restée, jusqu'au siècle dernier, le seul moyen matériel de transmission des connaissances. De ce fait, l'homme a toujours cherché à développer des techniques visant sa pérennité et sa diffusion le plus largement possible à bas prix. Actuellement, la prolifération des ordinateurs dans notre société conduit à une dématérialisation de l'écrit, au profit de sa forme numérique qui offre des possibilités de diffusion, de traitement, de stockage, d'indexation et d'accès à l'information beaucoup plus importantes que celles classiquement offertes par le support papier. Ainsi sont apparus les concepts de société sans papier et de bibliothèque virtuelle. Paradoxalement, notre société de consommation produit un nombre considérable de documents écrits : lettres, enveloppes, chèques, formulaires, mémentos, livres, journaux… Il n'est alors pas étonnant de constater le développement de nombreuses techniques visant à la conversion de ces documents vers la forme numérique de l'écrit en vue de leur traitement automatique par l'intermédiaire d'ordinateurs.

L'automatisation de la lecture des documents écrits qu'il s'agisse de documents imprimés ou de documents manuscrits consiste à doter les machines de capacité de lecture similaire à celle des hommes. La lecture constitue un des apprentissages de base de l'être humain, et lui semble une tâche suffisamment maîtrisée, toutefois son automatisation est délicate, à cause de la difficulté pour une machine de prendre en compte tous les aspects liés à la richesse et la variabilité de l'écriture, et ceci est d'autant plus vrai pour l'écriture manuscrite puisque chaque individu possède un style d'écriture unique. Ces aspects se manifestent au niveau de la grande diversité des styles d'écriture et des formes (caractères, mots, etc.), l'imprécision et

l'ambiguïté du processus d'écriture, l'inclinaison et la direction de l'écriture, l'existence de recouvrements et de liaisons entre les caractères, la localisation et l'extraction de l'information écrite, la taille du lexique, etc.

Il est nécessaire de distinguer la reconnaissance en ligne (on-line) de l'écriture manuscrite, qui relève plutôt de l'interfaçage entre l'homme et l'ordinateur (un stylo spécial est connecté à la machine et ne fonctionne que sur une tablette sensible), de la reconnaissance hors ligne (off-line) qui consiste à interpréter de l'écriture contenue dans une image typiquement saisie à l'aide d'un scanner. Dans le cadre de ce travail nous nous intéressons uniquement à la l'écriture hors-ligne.

La reconnaissance de documents s'applique à plusieurs langues écrites. La langue latine a reçu la plus grande attention de la part de chercheurs. En revanche, malgré le nombre de personnes qui parlent la langue arabe, peu de travaux de recherche sur la reconnaissance de documents ont été consacrés à cette langue.

Notre travail s'intègre dans le cadre du traitement automatique du document. Nous nous intéressons à la recherche des noms de villes dans les courriers arabes sans recourir à une reconnaissance du contenu à fin d'éviter le coût élevé et l'effort ardu de la reconnaissance.

Ce travail décrit les techniques dont on a développées pour construire un système de recherche d'images de documents arabes. Ce système contient tout les niveaux de traitements à partir d'une collection non structurée d'images de documents arabes numérisées jusqu'à une interface utilisateur permettant de retrouver l'information formulée par l'utilisateur à travers des mots clés en repérant des occurrences de ces mots dans les images de documents, passant par différents traitements tirés principalement du domaine de l'analyse de documents : prétraitements, segmentation, extraction et codage, et du domaine de la recherche d'informations textuelles : indexation et recherche.

Dans le premier chapitre nous abordons le problème général du traitement, d'analyse et de la reconnaissance d'images de documents sans être limités par un type particulier de documents. Nous commençons par donner quelques définitions relatives à la notion de documents. Ensuite, nous aborderons les étapes nécessaires à l'analyse et à la reconnaissance de documents en général. On essaye à travers ce chapitre de donner un survol des techniques les plus fréquemment utilisées dans l'analyse et la reconnaissance de documents

Le deuxième chapitre est consacré à la représentation des différentes applications sur les images de documents. On commence par un contexte général sur la recherche d'information, puis nous donnons un aperçu sur la recherche d'image et la recherche de mots en image, ensuite on présente l'extraction d'information, et enfin on donne une brève présentation de la catégorisation de documents.

Dans le troisième chapitre, nous détaillons la méthodologie adoptée pour la conception d'un système de recherche de mots dans les images de documents arabes par le contenu morphologique des caractères. On commence par l'exposition des caractéristiques de l'écriture arabes afin de montrer la difficulté de traitement de ce genre de documents. Par la suite, nous décrivons, dans le reste du chapitre en détails, les différentes phases intervenant dans le système, ainsi que, les algorithmes choisis.

Dans le quatrième chapitre nous présentons notre application et la base de données utilisée pour évaluer toutes les étapes de notre système donnant ainsi les résultats de tests et la discussion sur les résultats obtenus.

Enfin nous terminons notre mémoire par une conclusion générale et les perspectives auxquelles notre travail peut s'ouvrir.

Chapitre 1

Analyse et reconnaissance de documents

Un très grand nombre de documents de différentes catégories : journaux, formulaires, cartes géographiques, dessins techniques, partitions musicales, documents historiques existent aujourd'hui dans les musés, les archives nationales, les bibliothèques …etc. La numérisation de ces documents laisse apparaître souvent de nombreux défauts. Certains dépendent de l'état de conservation du document (la qualité du papier, l'acidité de l'encre, l'humidité du lieu de stockage, vieillissement du support...) ou des modifications apportées par l'Homme (annotations, soulignements, mauvaise restauration physique...). D'autres proviennent du processus même de numérisation (résolution insuffisante, courbure apparente, restitution non fidèle des couleurs...). Mais seule, la numérisation n'est pas suffisante, elle doit être accompagnée d'outils informatiques permettant un accès rapide et pertinent à l'information y contenue.

Dans ce chapitre, nous aborderons les étapes nécessaire à l'analyse et à la reconnaissance de documents en général sans se limiter par un type particulier de documents, ni d'une application particulière. On essaye à travers ce chapitre de donner un survol des techniques les plus fréquemment utilisées dans l'analyse et la reconnaissance de documents mais avant on débute par une présentation de la notion de document.

1.1. La notion de documents

1.1.1. Définition :

Plusieurs articles scientifiques abordent des problématiques liées aux documents, mais peu d'entre eux tentent de donner des définitions génériques pour ce terme. Par contre plein de définitions existent dans les dictionnaires, les encyclopédies et les répertoires. Le grand nombre de ces définitions rend difficile, la tentative de s'accorder sur une définition universelle de la notion de « document ».

Nous commençons par une définition étymologique du mot « document ».

D’après le Petit Robert édition 2002 [ROB 02], le terme « document » provient du mot latin « documentum », qui veut dire “ce qui sert à instruire”. On remontant dans l’origine, nous trouvons aussi le mot indo-européen « docte » qui veut dire “ acquérir ou faire acquérir une connaissance”.

Certaines définitions du mot « document » se limitent par le document écrit, ou bien le document papier, parmi ces définitions nous citons :

- Dans le Petit Larousse [LAR 86] : « Un document est un renseignement écrit ou objet servant de preuve ou d’information : document historique, photographique; (droit) titre qui permet d’identifier des marchandises pendant leur transport »
- En 1989, le Larousse de poche [LAR 89] définit le document comme « Ecrit servant de preuve ou de titre: objet quelconque servant de preuve ».
- Le Petit Robert édition 1993, définit le document comme étant « un écrit servant de preuve ou de renseignement. »

Le mot document correspond plus précisément à la « réunion » d'un support physique et d'une information. Un document est constitué d'information portée par un support. L'information y est délimitée et structurée, de façon tangible ou logique selon le support qui la porte, et elle est intelligible sous forme de mots, de sons ou d'images.

Une définition plus générale donnée par Karim HADJAR [HAD 06] dans sa thèse de doctorat indique qu’un document peut avoir plusieurs types (textuel, sonores, vidéo, graphique…etc.) selon le support choisi. Pour lui « Un document est le support physique pour conserver et transmettre de l’information ».

Bachimont [BAC 98] considère que le document est indissociable d’un support matériel. En effet, «un document est un objet matériel exprimant un contenu ». L’objet matériel est le support d’inscription où un contenu est exprimé. Le contenu est l’ensemble d’informations, de savoir à exprimer.

Le document joue un rôle important dans le développement des civilisations et de leurs cultures ; il permet de conserver et de transmettre les connaissances d’une génération à l’autre [DRI 07].

Après l'introduction de l'informatique et l'émergence des réseaux, et de l'internet, la définition du terme « document » devient plus générale. Cela est montré par les définitions suivantes :

- Un document est un ensemble de Données consignées sur support papier, électronique ou autre, pouvant être utilisées pour consultation, étude ou preuve. [OFF 06]
- Un document est un Œuvre fixée à un support matériel au moyen du langage ou d'autres symboles [OFF 06].
- L'ISO définit le document comme «l'ensemble d'un support d'information et des données enregistrées sur celui-ci sous une forme, en général, permanente et lisible par l'homme ou par une machine»

1.1.2. Le document numérique :

Le document numérique ou électronique est un objet informatique immatériel et manipulable avec un ordinateur. Il peut être une image, un fichier son, un ensemble de données organisées en fichier. Rappelons qu'un écrit électronique est tout équivalent d'un écrit papier dont la création est réalisée sur ordinateur. Il apparaît donc qu'un document électronique permet de séparer les caractéristiques d'un document classique, à savoir sa présentation (métadonnées), son contenu (informations), son architecture, offrant alors la possibilité d'une exploitation séparée.

À l'inverse des documents papiers, le document électronique peut permettre de séparer l'aspect présentation (mise en forme, mise en page...) et l'aspect information (contenu, données...), offrant alors la possibilité d'une exploitation séparée.

1.1.3. Cycle de vie d'un document :

Avec la numérisation est apparu un véritable cycle de vie des documents. Les pages manuscrites ou imprimées peuvent être capturées et leurs images analysées afin d'en extraire le contenu informatif. Ce dernier pourra être diffusé, modifié sur le fond (corrections, annotations, *etc.*) ou la forme (modifications de la mise en page, des signes typographiques, *etc.*) et finalement réédité pour donner lieu à de nouveaux documents physiques. Ainsi, un document n'est plus nécessairement prisonnier d'un support donné (et donc périssable) puisqu'il peut être recréé. Dans l'optique traditionnelle, l'´evolution d'un document est linéaire et peut se résumer en deux phases que sont la production et la consommation. La numérisation introduit l'opportunité d'un recyclage, ajoutant ainsi une boucle dans le cycle de vie de l'´ecrit, **(*Figure 1*)** [DUO 05].

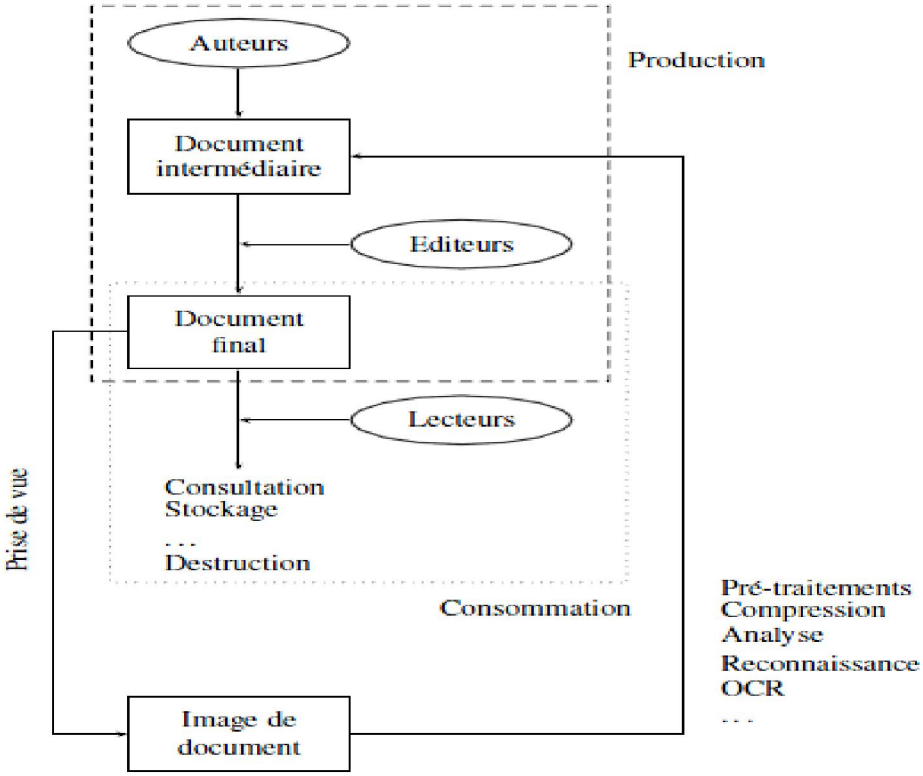


Figure 1.1 : *cycle de vie des documents [DUO 05].*

1.1.4. Diversité des documents :

Les documents manipulés représentent une grande variabilité à différentes mesures (catégorie, type, contenu, dégradation, ...) et il serait difficile de les classer de façon définitive. Dans [NAG 00] l'auteur a proposé une classification en documents structurés et en documents graphiques suivant la prédominance des zones textuelles ou graphiques et dans (Rolf, 2002) l'auteur caractérise le document papier suivant trois vues : le contenu, la complexité et la qualité. Un document peut être imprimé ou manuscrit. Il peut être contemporain ou ancien. Il peut contenir des zones textuelles et/ou graphiques. Un document peut être à structure simple ou à structure complexe suivant l'organisation spatiale des différentes zones. Un texte peut être écrit en plusieurs langues, avec plusieurs fontes et en différentes tailles. Un document est plus au moins dégradé (noir/blanc, niveau de gris ou en couleur). Tous ces points doivent être pris en compte par le système de reconnaissance, ce qui nécessite des traitements divers et très complexes.

a- Caractéristique des documents imprimés contemporains

Les documents imprimés modernes représentent une typographie (police, taille,…) bien connue et compréhensible par les OCR de nos jours, ce qui présente un avantage

lors de la reconnaissance des zones textuelles. En contre partie ces documents représentent une structure très variable, on peut trouver : des documents à structure simple, des documents à structure complexe et stable (article scientifique, formulaire, …) et des documents à structure complexe et variable (journaux, magazines,…) **(*figure 1.2*).**

b- Caractéristique des documents anciens

Les documents anciens se caractérisent par des présentations et des écritures très variées différentes de celles appliquées sur les documents contemporains. Ils se caractérisent par des variabilités de styles d'impression non utilisés à nos jours (fontes, polices, taille, lettrine, …). Dans [BEL 04] les auteurs incluent dans cette catégorie de documents les manuscrits anciens, les calligraphiés et les imprimés reconnaissables par OCR avec un entraînement spécial sur certains caractères.

L'usure du temps a de plus produit des altérations au document original et l'image numérisée qui en découle contient alors des imperfections (tâches, écritures fragmentées) qui n'existent pas dans les documents plus modernes, [LIK 03]. **(*Figure 1.3*)**

c- Caractéristique des manuscrits

Les manuscrits (*courriers manuscrits, les brouillons, les agendas, enveloppe, etc*) sont caractérisés par des lignes de longueur différente, plus ou moins fluctuantes. Les difficultés majeures sont l'imbrication des lignes, le chevauchement de composantes (composantes appartenant à plusieurs lignes de texte du fait de la présence de hampes et de jambages) et la fragmentation des caractères (due à la binarisation ou à la non homogénéité de l'encre), [LIK 03]. **(*Figure 1.4*)**

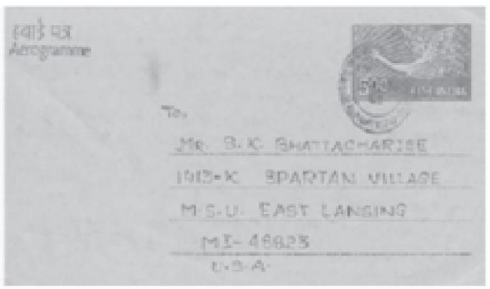
Journal [HAD 06] formulaire médicale [LEM 07] article scientifique [LEM 07]

Figure 1.2 : *exemple de document imprimé contemporains*

Figure1.3 : *exemple de document ancien [DRI 07]*

Brouillon de « Charmes »Paul Valéry [LEM 07]

Exemple de courrier manuscrit [LEM 07]



Enveloppe [LEM 07]

Figure 1.4*. Exemples de documents manuscrits*

1.1.5. Structure de documents :

Un document possède une structure qui apporte une information supplémentaire sans laquelle il ne serait pas possible de le lire ou de l'interpréter correctement. La structure du document n'est pas seulement une aide à la lecture, elle porte aussi une information, aussi importante que le contenu du document, qui traduit à la fois la fonction du texte et l'intention de son auteur. En modifiant la structure d'un texte, on peut changer radicalement la signification du

message souhaité par l'auteur et l'interprétation du document par le lecteur. « Tout document textuel est construit selon une structure, qui est reconnue par les lecteurs humains grâce à des marques typographiques, des conventions de mise en page, des connaissances 'pragmatiques', culturelles, relatives aux informations génériques qu'est susceptible de contenir ou bien que doit contenir tout document particulier, appartenant à une certaine catégorie de documents ». on peut cependant distinguer deux types de structure dans un document : la structure physique et la structure logique.

La structure logique : décrit l'organisation hiérarchique du texte contenu dans un document au moyen d'entités logiques telles que les chapitres, les sections, les titres, les paragraphes, les notes, les citations, les formules, les tableaux, les cellules ou les graphiques. Les entités logiques sont des concepts servant à structurer le message de l'auteur ; en retour, elles servent de repères au lecteur. Cette abstraction offre l'avantage de rendre la description du texte contenu dans un document indépendant de tout support physique. [AZO 95]

La structure physique : La structure physique est définie à la fois par la typographie et l'organisation du document. La typographie définit le style des éléments graphiques (polices de caractères, couleurs, traits, cadres…), et la forme de mise en page (colonage, interlignage, justification…). L'organisation du document décrit l'agencement de tous les objets visuels qui composent le document (caractères, mots, lignes, blocs, paragraphes, colonnes, images) et les relations spatiales entre ces objets (hiérarchie, inclusion, voisinage, position). Cette structure physique est la structure perceptible telle qu'elle apparaît visuellement au lecteur [EMP 03].

1.1.6. L'intérêt de la numérisation des documents :

Le récent développement des technologies numériques de l'information et de la communication a révolutionné la transmission des connaissances et remis en cause la suprématie du papier comme support unique de diffusion des savoirs. Mais les données numériques ne remplaceront jamais les supports physiques traditionnels, elles ne sont qu'une représentation plus ou moins fidèle des documents originaux. Par conséquent, la «copie numérique» ne constitue pas une alternative pérenne à la conservation du patrimoine culturel. Cependant, la numérisation présente de nombreuses propriétés très intéressantes [DRI 07] :

- Facilité de reproduction : Les données numériques peuvent être dupliquées très rapidement sans aucune perte d'information. En outre, ces données peuvent être échangées, stockées, manipulées, interrogées et recopiées avec des coûts très réduits.

- Indépendance du support physique : Les données numériques peuvent être stockées sur des supports variés de façon permanente ou bien itinérante sur des disques d'ordinateurs ou des supports amovibles. Ces données peuvent être reproduites sur de nombreux supports (papiers, écrans, papiers électroniques...) et même être consultables n'importe où et par n'importe quel moyen informatique fixe ou mobile (PC, PDA, ebook, téléphone portable...). La numérisation va donc révolutionner l'accès à l'information et à la connaissance. De nombreux chercheurs peuvent travailler simultanément à distance sur les mêmes documents ou bien consulter des documents rares autrefois peu accessibles.
- Facilité de consultation : La consultation de documents numériques offre un confort visuel sans précédent, une navigation plus performante et une recherche d'information plus efficace en comparaison avec les formes traditionnelles de consultation des livres ou des microfilms. En effet, les moteurs de recherche permettent de retrouver efficacement et de comparer rapidement les informations, dans plusieurs bibliothèques numériques. Néanmoins, ces moteurs de recherche fonctionnent uniquement sur des textes électroniques et ne peuvent pas indexer l'information directement à partir des images. Une solution à ce problème consiste soit à décrire manuellement leurs contenus, soit à développer des systèmes de reconnaissance capables d'extraire automatiquement des informations dans les images.
- Préservation et conservation des documents originaux : La numérisation permet de garder une copie des documents originaux risquant de partir en poussière. Nous rappelons, à titre d'exemple, l'incendie qui a eu lieu le 12 juin 1999 à la Bibliothèque Universitaire des Lettres et Sciences Humaines et Sociales de Lyon. Cet incendie a permis de faire prendre conscience du caractère éphémère des ouvrages anciens et de l'intérêt d'une conservation alternative. Les copies numériques réduisent les consultations excessives des lecteurs qui contribuent à l'usure des documents originaux fragiles. C'est la raison qui pousse les gestionnaires des fonds patrimoniaux à financer de coûteux projets de numérisation. L'UNESCO, responsable de la protection du patrimoine culturel mondial, a lancé en 1992 le programme «Mémoire du monde», pour sauvegarder et promouvoir ce patrimoine par la numérisation [STE 95].

Mais la numérisation seule ne suffit plus [AND 99]. Il faut absolument qu'elle cohabite avec le développement d'outils informatiques destinés à améliorer les conditions d'accès et de

recherche. L'analyse d'images de documents est un domaine de recherche actif à la frontière de la reconnaissance des formes et de l'analyse d'images.

1.2. Analyse et reconnaissance de documents :

L'analyse de documents, ou plus précisément l'analyse d'images de documents est une discipline qui étudie les possibilités algorithmiques de reconstituer une information structurée à partir de la forme visuelle brute (à partir de son image). [BAP 98].

L'analyse de documents consiste à chercher dans le traitement d'images, des solutions génériques à des problèmes de type document avec comme but la reconstitution du contenu du document selon une forme définie par l'application en question. Notons que l'analyse de documents fait l'objet d'un grand nombre de travaux de recherche, et les bons résultats obtenus avec certains types de documents fait sortir le domaine de l'analyse de documents des laboratoires vers des applications réelles dans l'industrie comme le cas d'analyse de chèques, traitement de formulaires et le tri du courrier postal.

Très tôt dans l'histoire de l'informatique dés qu'elle quitta le domaine strict du calcul scientifique et des applications militaires, au début des années 1950, une des premières applications explorées fut la reconnaissance optique de caractères, ou OCR (Optical Character Recognition). À l'époque, on pensait aboutir rapidement à une machine qui saurait lire automatiquement n'importe quel document. Mais malgré des premiers résultats spectaculaires et encourageants, il s'avéra rapidement qu'un taux de reconnaissance supérieur à 90% de caractères reconnus ne suffit pas à fournir un service satisfaisant pour l'usager : un seul chiffre d'un code postal mal identifié et non classé comme « inconnu » suffit à envoyer par erreur une lettre à l'autre bout du pays, dix ou vingt caractères erronés ou non reconnus par page de texte induisent un coût de reprise manuelle non négligeable, etc. [HIL 04]

Au fil des années la discipline s'est continuellement enrichie, profitant des progrès réalisés dans des domaines proches, comme le traitement du signal, l'analyse d'images, la reconnaissance des formes, ou l'intelligence artificielle. Les chercheurs se sont attaqués à une multitude de sous- problèmes, explorant autant les fondements théoriques que les approches empiriques. Chaque tentative d'application a mis en évidence de nouvelles difficultés pratiques, et des moyens pour les contourner [BAP 98].

Aujourd'hui, l'analyse de documents possède tous les avantages d'un domaine scientifique intensif et bien organisé. Le savoir-faire accumulé est riche et varié. La technologie se

concrétise par des prototypes de recherche opérationnels et des logiciels commerciaux [BAP 98].

L'analyse et la reconnaissance d'images de documents regroupe un ensemble de techniques informatiques dont le but est de reconstituer le contenu d'un document à partir de son image. Alors qu'elle est longtemps restée cantonnée dans la problématique de la reconnaissance de caractères, elle vise aujourd'hui des objectifs beaucoup plus larges, allant de la simple classification de documents à l'interprétation complète du contenu en passant par l'indexation ou la réédition. Ainsi, le but ultime de la reconnaissance d'images de documents est de générer une représentation de haut niveau sous la forme de documents structurés, selon une forme adéquate pour l'application visée.

La reconnaissance de l'écriture est l'une des applications les plus populaires de l'analyse d'image de documents. Elle a connu ces dernières années de grands progrès, et les succès des travaux de recherches ont donné lieu à de nombreuses applications industrielles, dans plusieurs domaines, citons par exemple la lecture automatique de formulaires, de chèques ou d'adresses postales. La reconnaissance de l'écriture suppose une localisation préalable des entités textuelles qui peuvent êtres mots ou bien caractères. Deux modes d'écritures existent : l'écriture imprimée, et l'écriture manuscrite.

Pour la reconnaissance de l'imprimé ou du manuscrit l'approche n'est pas la même selon qu'il s'agisse de reconnaître un imprimé ou un manuscrit. Dans le cas de l'imprimé, les caractères sont bien alignés et souvent bien séparés verticalement, ce qui simplifie la phase de lecture, bien que certaines fontes présentent parfois des accolements qu'il faut défaire. De plus, le graphisme des caractères est conforme à 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. Cela nécessite l'emploi de techniques de délimitation très spécifiques et souvent des connaissances contextuelles pour guider la lecture [DER 09].

Dans le cas de l'imprimé, la reconnaissance peut être monofonte, multifonte ou omnifonte. Un système est dit *monofonte* s'il ne traite qu'une fonte à la fois, c'est-à-dire que le système ne connaît l'alphabet que dans une seule fonte. L'apprentissage y est simple puisque l'alphabet représenté est réduit. Un système est dit *multifonte* s'il est capable de reconnaître un mélange de quelques fontes parmi un ensemble de fontes préalablement apprises. Dans ce cas, le prétraitement doit réduire les écarts entre les caractères (taille, épaisseur et inclinaison),

l'apprentissage doit gérer les ambiguïtés dues aux éventuelles ressemblances de caractères des différentes fontes, et la reconnaissance doit identifier les subtiles différences entre ces caractères. Enfin, un système est dit omnifonte s'il est capable de reconnaître toute fonte sans l'avoir absolument apprise, ce qui relève actuellement du domaine de la recherche. [DER 09]

Dans le cas du manuscrit, la reconnaissance peut être monoscripteur, multiscripteur ou omniscripteur. L'écriture manuscrite hors-ligne peut être classée en deux catégories d'écritures : écriture cursive et écriture semi-cursive. [BOU 06]

-Un système est dit Mono-scripteur (propres au scripteur) : c'est le fait que le système ne peut reconnaître qu'une seul écriture. Tous ces éléments influent sur la forme des lettres (écriture penchée, bouclée, arrondie, linéaire, etc.) et bien sûr sur la forme des ligatures, compromettant parfois le repérage des limites entre lettres.

- Un système est dit Multi-scripteur (propres à l'écriture manuscrite): c'est que le système peut identifier et reconnaître l'écriture pour un certain nombre de scripteurs.

-Et un système est dit Omni-scripteur (propres à n'importe quelle écriture manuscrite): c'est le fait de réduire l'information contenue dans l'image au minimum nécessaire pour modéliser précisément la structure des caractères.

a-Approches de reconnaissance :

Il existe deux approches possibles pour la reconnaissance des mots: [BEL 01] [CHA 06].

L'approche globale **:** Elle possède une vision générale du mot. Cette approche est basée sur une description unique de l'image du mot dans son ensemble, sans chercher à identifier chacune des lettres qui le compose. L'inconvénient de cette approche réside dans sa sensibilité à la variabilité des mots. Cependant, l'aspect généraliste de cette approche la limite à des vocabulaires distincts et réduits (cas des montants numériques de chèques).

L'approche analytique **:** vise à reconnaître les mots en identifiant les lettres qui le composent. Une étape de segmentation est donc nécessaire afin de déterminer les limites entre les lettres. Cette approche est la seule applicable dans le cas de grands vocabulaires. Son inconvénient principal demeure la nécessité de l'étape de segmentation en caractères.

b. Reconnaissance de caractères :

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 OCR et reste génératrice de leurs principales erreurs [BEL 01]. Ils sont aussi incapables de traiter des documents scientifiques ou techniques qui contiennent des formules mathématiques, des symboles ou autres schémas normalisés. Ils ne permettent pas encore de lire certains alphabets (arabe, grec…) ni les documents anciens qui utilisent des polices de caractères aujourd'hui disparues.

Les logiciels de reconnaissance de caractères sont sensibles à la résolution et à la qualité des formes de caractères. Par exemple, les documents imprimés de mauvaise qualité comme les télécopies sont encore aujourd'hui difficilement traitables. De plus il ne faut pas oublier qu'un taux de reconnaissance de 99% se traduit par plus d'une erreur toutes les deux lignes. Les OCR ne permettent pas de reconnaître le contenu des documents manuscrits dont la nature cursive demande un traitement totalement différent. [EMP 03]

1.3. Etapes d'analyse et de reconnaissance de documents :

De nos jours les organisations utilisent encore un grand nombre de documents papiers imprimés ou manuscrits qui nécessitent d'être représentés sous forme numérique et exploiter de la manière la plus efficace, sans recourir à une saisie manuelle. De même un grand nombre d'ouvrages et de documents anciens du monde entier sont conservés dans les archives et qui sont menacés de disparaître à cause de l'humidité, l'acidité du papier, etc. Il se trouve alors important de préserver ce patrimoine, de le rendre accessible à tout le monde et de l'interpréter facilement.

La numérisation est la solution adoptée, mais elle ne fournit que des images de documents, ce qui n'est pas toujours suffisant. En effet, il est souvent nécessaire d'accéder aux contenus des documents numérisés et de les modifier éventuellement. C'est l'objet de l'Analyse et la Reconnaissance des Documents.

1.3.1. L'acquisition :

Avant l'avènement de l'informatique, une grande partie des documentations était éditée sur support papier. Afin de faciliter le stockage et l'indexation de ces documentations, des projets

d'analyse de documents ont commencé à émerger. Ces analyses de documents se décomposent en plusieurs étapes, dont la première est la numérisation du document [DOS 00].

La numérisation ou l'acquisition d'images est le résultat de la conversion du document papier en une image numérisée. Ce processus est effectué soit par le biais d'un scanner soit d'une caméra. Le résultat de cette numérisation est une image. La qualité de l'image numérique obtenue dépend de plusieurs facteurs : la qualité du papier, la qualité du scanner ou de la caméra et le format d'image numérisée (compressé ou pas). En effet, s'il s'agit d'un document très ancien, le papier a de fortes chances d'avoir une couleur d'aspect jaunâtre. Ceci se reflète sur le résultat de la numérisation. Les trois points suivants diminuent la qualité : un scanner contenant de la poussière sur sa vitre, un scanner possédant une basse résolution et une caméra dont la mise au point est mal effectuée. Dans la chaîne de la qualité nous notons aussi le format de l'image numérisée, un format d'image compressé avec perte à l'instar de JPEG dégrade l'image numérisée obtenue [HAD 06].

L'acquisition est alors importante car elle définit la qualité des images, c'est à partir de cette image numérisée que l'analyse du document commence et le succès des étapes suivantes en dépend. C'est une étape coûteuse qui dépend du choix du matériel d'acquisition, de ses performances, de son réglage et de son utilisation. Une mauvaise numérisation avec des paramètres inadaptés et des choix techniques insuffisants implique une perte d'information que le traitement et l'analyse d'images ne peuvent pas toujours retrouver.

1.3.2. Prétraitement :

La première étape du traitement consiste à améliorer la qualité de l'image en éliminant les défauts dus à l'éclairage et au processus d'acquisition. Le vieillissement (trous, taches d'humidité,...) du document produit lui aussi des imperfections lors de la numérisation du document. Plusieurs techniques de traitement d'image sont mises en œuvre, ces techniques permettent de rehausser la qualité de l'image du document et aussi de préparer le terrain pour les processus suivants mais l'application de ces techniques de traitements se diffère d'un système à un autre selon le type du document traité. Les premières techniques appliquer à l'image numérisée sont appelées prétraitement et consistent en un ensemble d'algorithmes (filtrage, redressement, lissage, squelettisation, binarisation…) **(*Figure 1.5*)** dont l'objectif est de préparer le terrain à la reconnaissance. Le résultat de ce prétraitement est une image épurée dépourvue de bruit. Nous allons présenter quelques traitements.

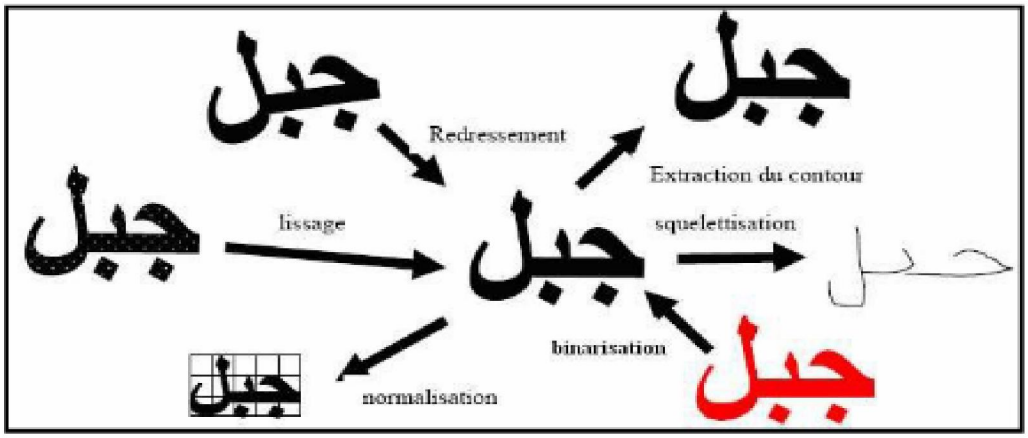


Figure 1.5 : *effets de certaines opérations de prétraitement. [BOU 06]*

1.3.2.1. Suppression du bruit :

Le bruit peut être dû aux conditions d'acquisition ou encore à la qualité du document d'origine comme on la vu précédemment. Le bruit s'insère dans le signal des images de différentes manières. Il peut être aléatoire et cohérent avec ce signal et dans ce cas, il ne peut être supprimé qu'en faisant appel à des connaissances a priori sur l'image. Le bruit peut être aussi périodique et donc se trouve en dehors des informations utiles. Dans ce cas, sa suppression est souvent aisée par des techniques classiques de filtrage et ne cause aucune destruction des informations utiles [BEL 92].

Le principe du filtrage est de modifier la valeur des pixels d'une image, généralement dans le but d'améliorer son aspect. En pratique, il s'agit de créer une nouvelle image en se servant des valeurs des pixels de l'image d'origine. Différentes méthodes de filtrage ont été développées suivant le type et l'intensité du bruit, ou les applications auxquelles on destine l'image. Dans un premier temps, les filtres linéaires, simples d'implantation, et se prêtant bien à l'étude analytique complète, ont été largement utilisé comme outils de traitement d'images. Cependant, ces filtres ont été progressivement remplacés par des techniques de filtrage non linéaire caractérisées par de meilleures performances tant en réduction de bruit qu'en préservation de contours

A cause de la grande quantité de données à traiter dans une image, la mise en œuvre se fait souvent par un traitement local qui consiste à balayer l'image par une fenêtre de filtrage de taille N (impaire la plupart du temps). Ce filtrage local assure la rapidité de traitement. Les performances obtenues, en terme de réduction de bruit ou de préservation de transitions,

dépendent à la fois de la taille de la fenêtre d'analyse et du traitement effectué à l'intérieur de la fenêtre. [TAB 98]

a. Filtres linéaires :

Ce sont les filtres les plus simples et les plus faciles à implanter. Grâce à cette simplicité, de nombreux filtres, typiquement de type passe-bas (pour le lissage), ont été proposés dans la littérature et appliqués au filtrage d'image. Le filtrage est effectué en utilisant un masque de convolutions, le niveau de gris du pixel central est remplacé par la convolution du voisinage avec le masque.

Les filtres linéaires sont les plus performants en termes de réduction de bruit dans le cas de bruit à distribution gaussienne. Ceci constitue une première limitation des filtres linéaires car le bruit dans une image naturelle n'est pas toujours gaussien. Dans de nombreuses situations, il est plutôt impulsionnel ou au contraire, très concentré. De plus, les filtres linéaires présentent un autre inconvénient quand il s'agit de traiter une discontinuité (par exemple une frontière entre deux régions). Ils ont tendance à lisser les transitions donnant une impression de flou sur les bords des objets et à rendre délicat l'extraction et la localisation de contours des objets [TAB 98]. Les filtres linéaires utilisés classiquement en traitement d'images peuvent être : Filtres passe bas et Filtres passe haut

Un filtre passe haut favorise les hautes fréquences spatiales, comme les détails, et de ce fait, il améliore le contraste et met en évidence les contours. Un filtre passe haut est caractérisé par un noyau comportant des valeurs négatives autour du pixel central. Les techniques permettant de détecter un contour sont basées sur l'utilisation: de filtres gradients et de filtres Laplaciens

Les filtres passe bas agissent en sens inverse des filtres passe haut et le résultat est un adoucissement des détails, ainsi qu'une réduction du bruit granuleux. Ils agissent par moyenne sur un voisinage et suppriment donc les détails. On trouve le : Filtre Moyenneur, Filtre Gaussien...

Le gros avantage de ces filtres est leur facilité de conception et d'implémentation, mais ils ne peuvent être utilisés pour des travaux trop fins (la détérioration des contours qu'ils induisent par exemple, empêchera une segmentation fine des images).
Ces limitations ont donc conduit à la conception de filtres non-linéaires

b. Filtres non-linéaires :

Ces opérateurs ont été développés pour pallier aux insuffisances des filtres linéaires principalement la mauvaise conservation des contours. Ils ont le défaut d'infliger des déformations irréversibles à l'image. Parmi les filtres non-linéaires, on peut citer Filtre Min-Max et le Filtre médian.

1.3.2.2. Restauration :

Ceci signifie que l'image observée est altérée par les instruments de mesure ou du milieu de propagation de l'information (canal). En effet, ces appareils présentent toujours des imperfections qui se répercutent sur la qualité de l'image observée. Autrement dit, une partie de l'information a été perdue. La restauration de l'image signifie retrouver l'image idéale a partir de son observation dégradée.

Les systèmes qui traitent des documents dégradés disposent souvent d'un module de restauration qui permet de reproduire une représentation, la plus proche possible, de la qualité de l'image originale avant sa dégradation par le processus de numérisation **(*figure 1.6*).**

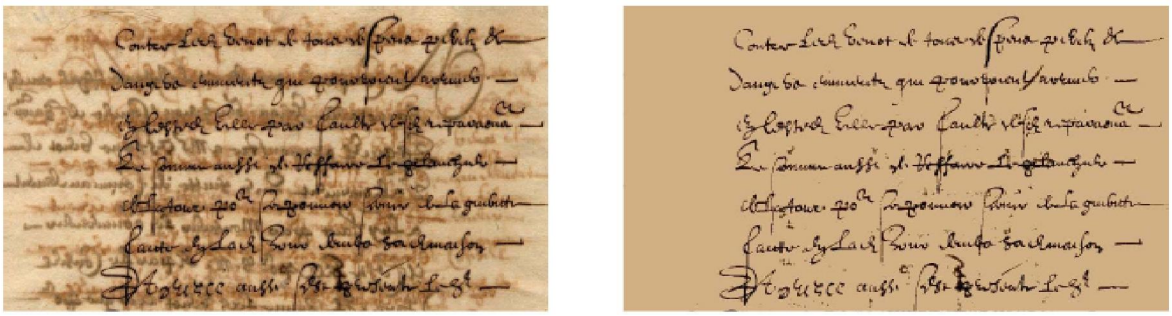


Figure 1.6 : *l'image d'origine et sa version restaurée [DRI 07].*

Les défauts attachés aux documents sont de différents types, et peuvent être classés en deux groupes : les dégradations du fond, elles se rapportent aux défauts du support papier comme les taches dues à l'humidité, au passage en transparence du verso sur le recto, aux annotations et aux ajouts indésirables, et les dégradations de formes, elles modifient la continuité des traits et la topologie des objets, différentes dégradations peuvent apparaître sur les caractères : rupture des traits, caractères tronqués, vide dans les traits, fusion de caractères, effet d'escalier [DRI 07].

1.3.2.3. Binarisation (seuillage)

La binarisation permet de passer d'une image de niveaux de gris ou couleur à une image noir et blanc, en fonction d'un seuil à définir. Elle consiste à détacher le texte du fond, ce qui permet de diminuer la quantité de données à traiter et de réduire l'espace mémoire et le temps

de calcul. Le seuil de binarisation correspond à la limite entre les contrastes forts et faibles de l'image, fixer ce seuil est très difficile quand le contraste varie dans l'image. Nous nous intéressons dans notre travail que par la binarisation des documents en niveaux de gris, car la plupart des documents en couleurs, peuvent être convertis fidèlement en niveaux de gris.

Le choix du seuil de binarisation est important. D'un côté, si le seuil est trop haut, les traits fins peuvent être coupés (comme dans le mot Tridon de la ***figure 1.17***), et d'un autre côté, si le seuil est trop bas, le trait sera plus gros et des informations peuvent être perdues (comme dans le mot Derrien de la figure 1.10 où le r est peu précis) [ROU 07].

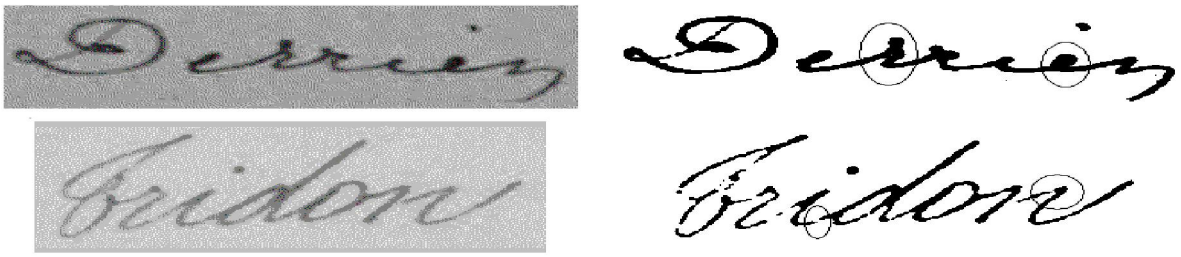


Figure 1.7 : *Binarisation : perte de précision et perte de connexité.*

Selon plusieurs travaux de recherche [ARI 01], les techniques de binarisation d'images en niveaux de gris peuvent être classées en deux catégories : **seuillage globale**, où un seul seuil est utilisé dans toute l'image pour la deviser en deux classes (texte et fond), et **seuillage local** où les valeurs des seuils sont déterminées localement, pixel par pixel ou bien région par région. D'autres [SAU 00], ajoutent un troisième groupe de méthodes hybrides, ces méthodes combinent des informations globales et locales pour attribuer les pixels à l'une des deux classes. [KEF 09]

1.3.2.4. La normalisation :

L'image du caractère ou du mot manuscrit peut être de taille quelconque, il est parfois utile de ramener sa taille à des dimensions normalisées: c'est l'opération de normalisation de l'image ou de mise à l'échelle. Cette opération permet de diminuer la complexité des algorithmes d'extraction des caractéristiques en temps de calcul et permet aussi d'obtenir des caractéristiques invariantes à la taille du caractère ou du mot [BEN 08].

La normalisation peut être effectuée soit au niveau du mot pour uniformiser la hauteur des caractères [ALM 02a, ALM 04, ALM 06], la largeur des caractères [ABD 06], ou la hauteur et la largeur des caractères [ALM 02b, PEC 03]; soit au niveau du caractère ou graphème pour uniformiser sa hauteur et sa largeur [MEH 05, SYI 06].

1.3.2.5. Correction de l'inclinaison (redressement)

L'inclinaison des images est provoquée essentiellement soit par un mauvais positionnement des pages lors de la saisie optique, soit par une mise en page fantaisiste et irrégulière de l'auteur. L'estimation de cet inclinaison est nécessaire pour certaines techniques de segmentation qui n'obtiennent de bons résultats que :

-si les images sont parfaitement redressées.
-ou connaissant l'angle d'inclinaison. [AZO 95]

L'angle d'inclinaison est considéré comme l'angle produit entre les lignes de texte de l'image et la direction horizontale. La correction de l'inclinaison ou de courbure consiste à calculer tout d'abord l'angle d'inclinaison, puis de ré-échantillonner l'image en appliquant une rotation de l'image d'angle φ.

Parmi les techniques de détection de l'angle d'inclinaison les plus utilisées : la méthode Trincklin [BEL 92, ING 99], la méthode de projection [BAG 97, BEL 92, BUN 97, DAN 99, KAV 02], la transformée de Hough [AMI 96, BER98, HUL 98, JAI 96, LE 94, PAR 96, YIN 01] et la méthode des k-plus proches voisins [ANT 97, SAF 00, SEH 00].

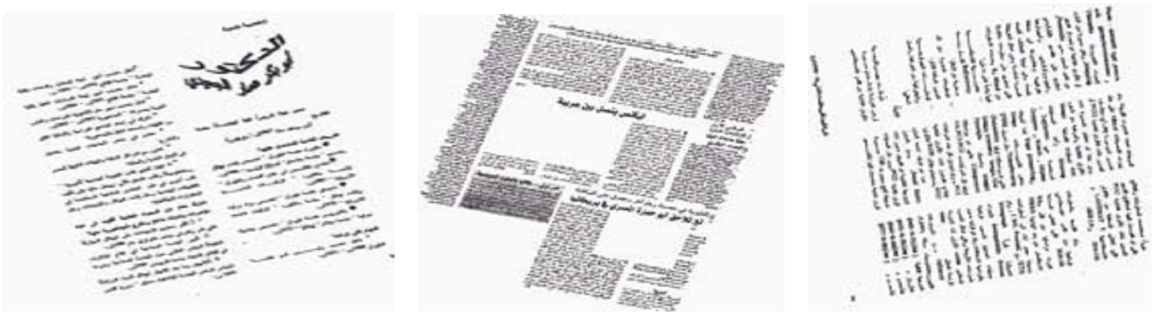


Figure 1.8 : *Exemple de quelques documents inclinés [SEH 04].*

1.3.3. Segmentation

Segmenter consiste à partitionner le document en régions homogènes. Dans le cas des documents textuels cela revient à identifier et localiser les blocs de textes, les paragraphes, les lignes de texte, les mots... etc. Cette structuration est hiérarchique et peut être représentée par un arbre. Deux approches sont possibles pour déterminer cette structure, une approche ascendante ou une approche descendante. [NIC 06]

-**Approches ascendantes**: Les méthodes ascendantes commencent par le niveau le plus bas et remontent d'un niveau à un autre jusqu'à compléter la page. En effet, elles se basent sur

l'analyse des composantes connexes. Ces dernières sont obtenues en scannant une image pixel par pixel et en regroupant les pixels en des composants en se basant sur la connexité des pixels qui peut être en 4 voisins ou en 8 voisins.

Le principe des méthodes ascendantes est le suivant : elles commencent par fusionner du plus bas niveau, en formant les mots à partir des composantes connexes, et puis remontent à un niveau supérieur en fusionnant les mots en lignes, les lignes en blocs, etc… jusqu'à ce que la page soit complètement reconstituée. [HAD 06]

-**Approches descendantes** : Les méthodes descendantes commencent par le niveau le plus élevé à savoir la page et descend d'un niveau à un autre jusqu'à arriver au niveau des composantes connexes ou au niveau pixel [HAD 06]. Elles requièrent généralement des connaissances *a priori* plus ou moins précises sur la structure des documents à traiter.

-**Approches mixtes** : cette classe regroupe souvent des méthodes qui embarquent des approches descendantes et également des approches ascendantes.

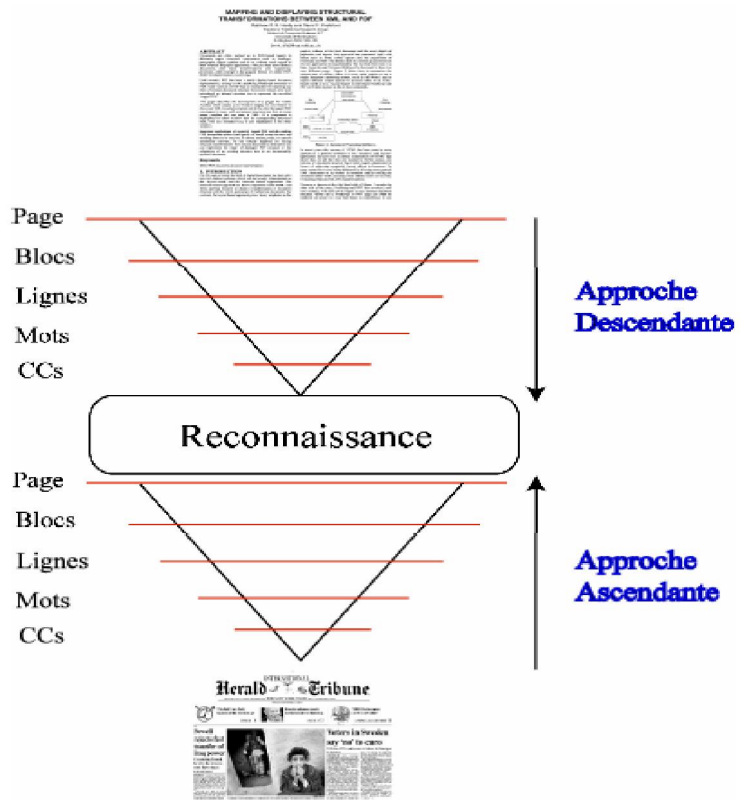


Figure 1.9 : *Approche descendante et ascendante [HAD 06]*

1.3.3.1. Segmentation texte/ graphique

Les traitements des phases ultérieures, opèrent sur une image dont on a séparé le texte des graphiques. La distinction entre les éléments textuels (caractères, symboles) et les éléments graphiques est alors nécessaire.

Plusieurs approches ont été proposées pour la séparation entre le texte et le graphique, qui se basent sur l'utilisation d'outils comme les filtres de Gabor ou les ondelettes et en s'appuyant sur l'étude des fréquences et des orientations des différentes parties de l'image. En effet, il y a une différence importante entre les fréquences caractéristiques d'une zone de texte et de graphique ce qui conduit à une bonne séparation texte/dessin. La principale difficulté de ces approches comme il est signalé dans [LEE 01] porte sur leur complexité algorithmique due à une paramétrisation des filtres utilisés pour rechercher les fréquences et les orientations.
Dans [BOU 06], les auteurs ont proposé une méthode pour la segmentation fond/ texte/ graphique d'images de documents manuscrits arabes anciens à structure complexe qui opère en deux phases : utilisation de l'analyse multi-échelle pour la segmentation avant/arrière-plan, puis segmentation texte/graphique par l'algorithme des C-moyen flous.
Dans les images binaires, la taille des composantes connexes ou leur aspect constituent un critère pour la séparation entre les composantes textuelles et les composantes graphiques. Khurshid et a1 [KHU 09b] par exemple utilise la taille des composantes comme critère pour la segmentation texte/ graphique. Ainsi les composantes graphiques vérifient:

Aire composante > (aire moyenne × A) ET Taille composante > (taille moyenne × B) ;

Tel que : aire moyenne est la moyenne des aires de toutes les composantes connexes dans cette image. A et B sont des constantes choisies par expérience égales à 5 et 4 respectivement.
La taille des composantes connexes est utilisée aussi par Waked et al [WAK 98] comme un critère de séparation. Dans cette méthode les tailles des composantes connexes sont comparées avec un seuil calculé auparavant en partant de la connaissance à priori que les composantes graphiques ont généralement une taille plus grande que les composantes textuelles. Dans [GUS 99], l'élimination des éléments graphiques notamment les lettrines dans les manuscrits médiévaux est basé sur des connaissances a priori sur la position de ces éléments. Dans le projet DEBORA[1] [BOS 08], une simple formule permet de déduire si la composante connexe CC_i est un texte ou un graphique. Elle associe à chaque CC_i une probabilité $P(x)$, si cette dernière est inférieure à un certain seuil, alors CC_i est un graphique

[1] DEBORA est l'acronyme de Digital AccEss to BOoks of the RenAissance (Accès numérique à des livres de la Renaissance).

sinon il s'agit d'un texte. HADJAR [HAD 06] s'intéresse à la séparation texte/graphique appliquée sur des images de journaux arabes. A partir de ces derniers il extrait les plus grandes composantes connexes qui seront filtrées par la suite. Après, on calcule les densités de pixels noirs et blancs pour chaque composante connexe obtenue. Si l'une de ces densités est supérieure à un seuil donné alors la composante connexe est un graphique.

1.3.3.2. Segmentation du texte en lignes

Une fois la séparation texte/graphique effectuée, on procède généralement à l'extraction du texte en lignes. L'extraction des lignes de texte est un préalable à tous les processus d'analyse et reconnaissance de mots ou de caractères. Elle a comme objectif d'assigner chaque composant du texte à une ligne appropriée ; ce qui permet de préparer les données pour les traitements ultérieurs tel que la segmentation en mots et l'extraction des caractéristiques.

Cette opération est relativement facile quand le texte est régulier, non incliné, ne comportant pas de chevauchement. Ces conditions sont sans doute réunies pour des textes imprimés mais pas souvent pour des textes manuscrits. Ces derniers sont caractérisés par une variation de la distance interligne, la présence de plusieurs lignes de base et les caractères de deux lignes de texte peuvent se toucher ou se chevaucher. Ce qui complique considérablement la segmentation en ligne. Dans le cas de l'écriture arabe, ces situations existent fréquemment à cause des caractères avec jambages et/ou hampes.

Les techniques d'extraction de lignes de texte pour les documents imprimés se divisent comme toutes les méthodes de segmentation en deux catégories : descendante et ascendante. Les méthodes descendantes sont intéressantes lorsque des connaissances sur la structure du document sont disponibles [DEF 95]. Ce sont les méthodes basées sur les profils de projections comme celles de [BAI 91] [ELL 90].

Notons que la technique de projection horizontale totale présente trois inconvénients majeurs:

- Elles sont sensibles à l'inclinaison,
- Sensible au chevauchement,
- La présence des points diacritiques produisent des faux minima.

Pour surmonter le problème d'inclinaison, Benasri et al [BEN 99] proposent une méthode d'extraction des lignes de texte manuscrit arabe, basée sur la projection partielle et le suivi de contour partiel. Tout d'abord on segmente l'image en colonnes de largeur fixe. La projection horizontale sur la première colonne nous permet d'extraire les minima locaux qui seront filtrés par la suite pour éliminer les faux minima. Un suivi de contour partiel est appliqué par la suite sur les minima locaux obtenus, en deux étapes. Il ne reste qu'annexer les points

diacritiques à l'une des deux lignes adjacentes. L'inconvénient de cette méthode est qu'elle impose que les lignes adjacentes ne doivent pas être collées. De plus la présence de chevauchement dans une colonne peut faire disparaitre le minima correspondant. Zahour et al [ZAH 04] ont modifié la méthode précédente afin de pallier au problème de collage des caractères de lignes voisines. La méthode commence par l'extraction des blocs de texte à partir des histogrammes de projection partielle. Ensuite, l'algorithme k-means est utilisé pour classer les blocs en trois classes : les symboles diacritiques, les tracés principaux des mots et les grands blocs résultant du chevauchement et collage des caractères des lignes voisines. L'étape suivante s'intéresse à la segmentation des grands blocs.

Les méthodes ascendantes partent des pixels de l'image et les fusionnent en composantes connexes puis en mots pour former des lignes. Nous trouvons sous cette catégorie la technique du *Run length smoothing algorithm* (RLSA) due à Wong [WON 82] son principe est de noircir toute séquence de pixels blancs comprise entre deux pixels noirs, de longueur inférieure à un seuil donné, la segmentation est alors obtenue en appliquant l'opérateur logique « et » sur les images résultant respectivement d'un lissage horizontal et d'un lissage vertical avec des seuils éventuellement différents pour l'horizontale et le verticale (***figure1.10***), la transformation de Hough [LIK 95], et le regroupement de composantes connexes par proximité [FEL 01] [LIK 94].

Dans [LIK 94], Likforman-Sulem et al, proposent une méthode itérative d'extraction des lignes adaptée aux documents manuscrits non-contraints. La méthode commence par l'étiquetage des composantes connexes. A partir des points d'ancrage[2], les composantes connexes qui satisfont certains critères de proximités, de similarité et continuité de direction sont regroupées dans un alignement. La résolution des conflits (qui peuvent apparaitre) est effectuée soit localement soit globalement. Likforman-Sulem et al proposent une autre méthode [LIK 95] basée sur la transformée de Hough. La transformé de Hough est utilisée ici pour la détection des lignes de texte. Un critère perceptif doit être utilisé par la suite pour valider les alignements détectés. Dans [DEF 95], une méthode de segmentation en lignes est proposée permettant de localiser puis d'extraire précisément les lignes de texte contenues dans l'image d'un document en trois étapes. La première étape se base sur une approche multi-résolution pour localiser les différentes zones de texte à partir de l'image en niveaux de gris. La binarisation est ensuite effectuée localement sur chaque zone d'intérêt. Enfin, une étape de post-segmentation est nécessaire.

[2] Ce sont des composantes connexes ayant une direction privilégiée

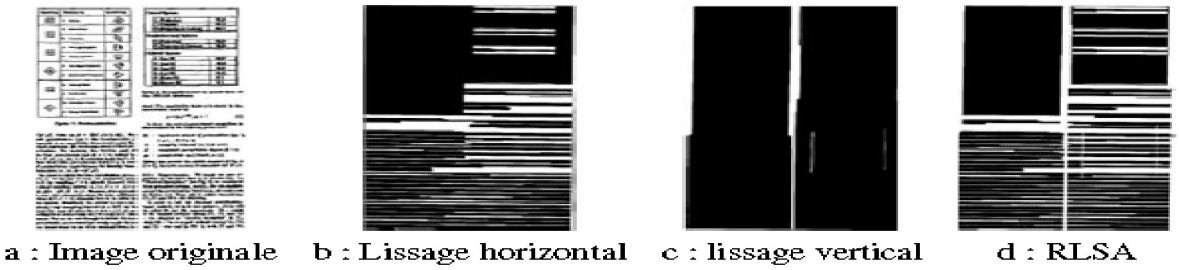


Figure1.10 : *Segmentation RLSA.*

1.3.3.3. Segmentation en pseudo mots

Un pseudo mot est une composante connexe avec ces points diacritiques, il peut être composé d'un ou plusieurs caractères. L'extraction des composantes connexes consiste à regrouper les pixels noirs voisins (dans une image binarisée) en un seul bloc délimité par un rectangle englobant. L'attribution des points diacritiques à une composante connexe est effectuée par l'analyse de proximité pour former un pseudo-mot. Un point diacritique est attribué à une composante connexe si la distance à ce dernier est minimal (***figure 1.11***).

Figure 1.11: *Attribution des points diacritiques à l'une des composantes le plus proche*

1.3.3.4. Segmentation en mots

La segmentation en mots se fait généralement après une phase de segmentation de lignes. La plupart de ces approches s'appuient sur l'analyse des espaces intra et inter mots, l'idée étant que les espaces entre deux mots sont plus grands que les espaces entre les lettres d'un mot. Ces espaces sont mesurés par une distance entre composantes connexes pour laquelle de nombreuses métriques existent [LEM 07].

Dans le cas des documents dont l'écriture est cursive, les mots sont extraits facilement, ils correspondent aux composantes connexes. Dans la (***figure 1.12)***, les mots apparaissent encadrés par des rectangles englobants.

L'écriture latine imprimée est caractérisé par un espace régulier entre les caractères et les mots, les composantes connexes correspondent aux caractères. Ces caractéristiques nous permettent de reconstruire les mots par agrégation des caractères sur des règles de proximité

[SOU 02]. Deux caractères voisins C_L et C_R (respectivement à droite et à gauche) seront agrégés dans un même mot si l'espace qui les sépare n'excède pas un seuil calculé automatiquement.

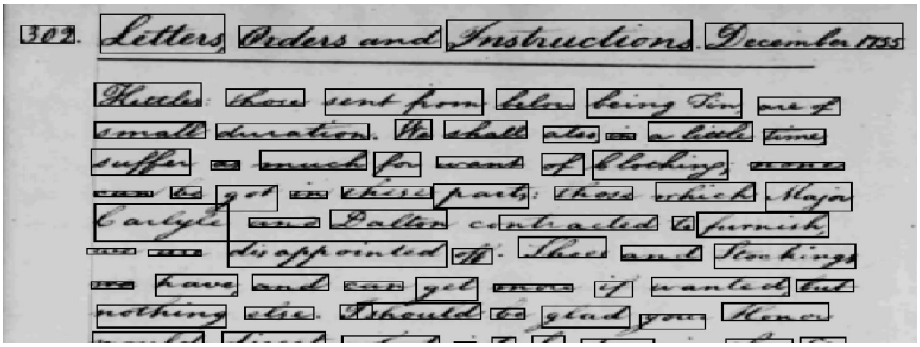


Figure 1.12: *extraction des mots dans un texte cursive*

Pour l'écriture arabe imprimée, les composantes connexes peuvent contenir un ou plusieurs caractères (***figure 1.13***), mais l'espace inter-mot et inter-caractère reste régulier, ce qui permet, également la séparation des mots par des mesures de proximité.

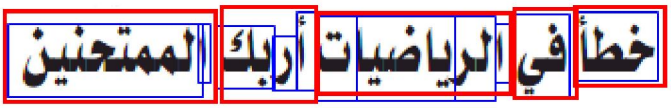


Figure 1.13: *extraction des mots d'un texte arabe imprimé, les composantes connexes encadrées en bleu, et les mots en rouge*

Le regroupement des composantes connexes en mots dans le cas de l'écriture arabe manuscrite qui est semi-cursive, est plus difficile. En effet, l'espace inter-mots et inter-caractères irrégulier rend difficile l'estimation d'un seuil pour la séparation des mots.

A. Belaid et al [BEL 06], insistent sur la difficulté de la segmentation en mots en arabe, et déplorent la trop faible quantité de travaux de recherche qui vont dans cette direction. Selon Menasri [MEN 08] les seuls travaux qui traitent cette tache (segmentation en mots dans une page de texte manuscrit) sont ceux de Srihari et al [SRI 05, SRI 06]. S. Srihari et al mettent au point un système qui utilise un réseau de neurones pour déterminer les points de coupures d'une segmentation en mots.

1.3.3.5. Segmentation du mot ou pseudo-mot en caractères

La segmentation de mot ou pseudo-mot est l'opération qui permet de décomposer l'image ou le signal d'un mot ou d'un pseudo-mot selon les caractères qui le composent. C'est une étape

critique et décisive dans quasiment tous les systèmes de reconnaissance qui sont basés sur une approche analytique. Sa délicatesse vient du fait que les limites d'un caractère sont parfois absentes ou floues et ne peut se faire sans avoir identifié le caractère. Deux stratégies de solutions sont souvent adoptées pour surmonter cette difficulté. La première néglige le problème de segmentation et procède à la reconnaissance en s'appuyant sur l'allure globale du mot ou pseudo-mot. La deuxième fait coopérer l'étape de reconnaissance avec l'étape de segmentation. Les méthodes de segmentation varient suivant le type d'écriture traitée et de leur interaction avec l'étape de reconnaissance, quoiqu'elles visent le même objectif qui est la génération d'un découpage le plus proche possible du découpage en lettres souhaité [BEN 08].

Il existe deux techniques permettant la mise en œuvre de la segmentation. La première, connue sous le nom de segmentation implicite, consiste à effectuer un découpage a priori de l'image en intervalles de grandeur régulière. Par opposition, la segmentation explicite consiste à utiliser des points caractéristiques dans le mot, tels que les minima locaux du contour supérieur, les espaces ou encore les points d'intersection. Le résultat de cette étape est la segmentation du mot en entités de base appelées graphèmes [SOU 06].

Segmentation implicite :

Les approches à segmentation implicite s'inspirent des méthodes utilisées dans le domaine de la reconnaissance de la parole. Contrairement à la segmentation explicite, il n'y a pas de présegmentation ou dissection du mot, alors la segmentation du mot est une segmentation aveugle dans le sens où elle ne dépend en aucun cas d'une analyse de l'image à segmenter. Le système recherche dans l'image, des composantes ou des groupements de graphèmes qui correspondent à ses classes de lettres. Classiquement, il peut le faire de deux manières, soit par une fenêtre glissante, soit par recherche de primitives.

Segmentation explicite :

Les méthodes de segmentation explicites s'appuient sur une analyse morphologique du mot manuscrit, ou sur la détection des points caractéristiques tels que les points d'intersection, les points d'inflexion, les boucles à l'intérieur du mot pour localiser les points de segmentation potentiels. La sélection des points de segmentation (i.e. chemin de segmentation) peut être effectuée soit sans le contrôle du moteur de reconnaissance de caractères, soit en alternant les phases de segmentation et reconnaissance de manière à valider les hypothèses de

segmentation par la reconnaissance. La segmentation explicite n'est pas parfaite pour un système de reconnaissance des mots manuscrite [BEL 02].

1.3.4. Extraction de caractéristiques (primitives)

Après avoir séparé les parties graphiques des parties textuelles, et segmenté le texte en lignes, mots puis en caractères, vient le dernier processus de réduction des informations avant l'étape de classification et reconnaissance qui est l'*extraction des caractéristiques ou des primitives*. Cette phase doit assurer un maximum de fiabilité, car l'image initiale est ignorée pour ne considérer que ses paramètres, et les performances du système de classification dépendent directement de leur choix. En effet, l'utilisation d'un classifieur très performant ne peut compenser une représentation mal adaptée ou peu discriminante. La difficulté de cette étape provient du fait que la qualité d'un ensemble de caractéristiques ne peut se juger que sur un problème particulier (reconnaissance de mots, caractères, chiffres, symboles, etc.), et qu'il n'existe pas des règles universellement établies pour l'extraction d'un ensemble de caractéristiques pertinentes de la forme à reconnaître. Par conséquent, la conception d'un nouveau système de reconnaissance d'écriture est toujours confrontée à un choix délicat lors de la définition d'un ensemble de caractéristiques car cette conception doit intégrer un ensemble de critères comme par exemple [PAQ 00]: la variabilité dans la classe à identifier ; la complexité de la procédure de détection des caractéristiques ; la complémentarité des caractéristiques sélectionnées; et le choix de la méthode de classification appropriée aux caractéristiques sélectionnées.

Devijver et al [DEV 82] ont défini l'extraction de caractéristiques comme le problème *d'extraction à partir de l'image de l'information la plus pertinente pour un problème de classification donné, c'est à dire celle qui minimise la variabilité intra-classe et qui maximise la variabilité inter-classe*.

L'objectif commun de toutes les primitives est de caractériser au mieux la forme des caractères, afin de pouvoir distinguer si deux images appartiennent à deux classes différentes ou à la même classe, c'est–à–dire qu'elles doivent diminuer la variabilité intra-classe et augmenter la variabilité inter-classe. Suivant les applications et les techniques utilisées pour le système de reconnaissance, les primitives extraites peuvent être très différentes [CHE 06].

Dans la littérature, les primitives sont classifiées de plusieurs façons différentes [CHE 06] :

Une première distinction peut être effectuée entre les primitives globales et les primitives locales. Les primitives globales cherchent à représenter au mieux la forme générale d'un caractère et sont donc calculées sur des images relativement grandes (ex : transformée de Fourier et transformée de Hough). Les primitives locales sont calculées lors d'un parcours des pixels de l'image avec un pas d'analyse qui dépend de la modélisation, du type de primitive et de la taille de l'image.

Une seconde distinction est effectuée entre les primitives topologiques, structurelles ou statistiques [TRI 95], [DAR 94] :

– Les primitives topologiques ou métriques : elles consistent à compter dans une forme le nombre de trous, évaluer les concavités, mesurer des pentes et autres paramètres de courbures et évaluer des orientations, mesurer la longueur et l'épaisseur des traits, détecter les croisements et les jonctions des traits, mesurer les surfaces et les périmètres, ...

– Les primitives structurelles : elles ressemblent beaucoup aux primitives topologiques. La différence est qu'elles sont généralement extraites non pas de l'image brute, mais à partir du squelette ou du contour de la forme. Ainsi, on ne parle plus de trous, mais de boucles ou de cycles dans une représentation filiforme du caractère. Parmi ces caractéristiques on peut citer :

- ✓ Les traits et les anses dans les différentes directions ainsi que leurs tailles.
- ✓ Les points terminaux.
- ✓ Les points d'intersections.
- ✓ Les boucles.
- ✓ Le nombre de points diacritiques et leur position par rapport à la ligne de base.
- ✓ Les voyellations et les zigzags (hamza).
- ✓ La hauteur et la largeur du caractère.
- ✓ La catégorie de la forme (partie primaire ou point diacritique, etc).
- ✓ Plusieurs autres caractéristiques peuvent être tires, suivant qu'ils soient extraits d'une courbe, un trait ou un segment de contour.

– Les primitives statistiques : l'histogramme, qui représente le nombre de pixels sur chaque ligne ou colonne de l'image, en est un exemple classique et simple à calculer. On peut citer également l'approche basée sur un moyennage des pixels situés à l'intérieur d'un masque rectangulaire : on construit une matrice de masques recouvrant la totalité de la forme qui permet une représentation statistique des valeurs correspondant à chaque masque.

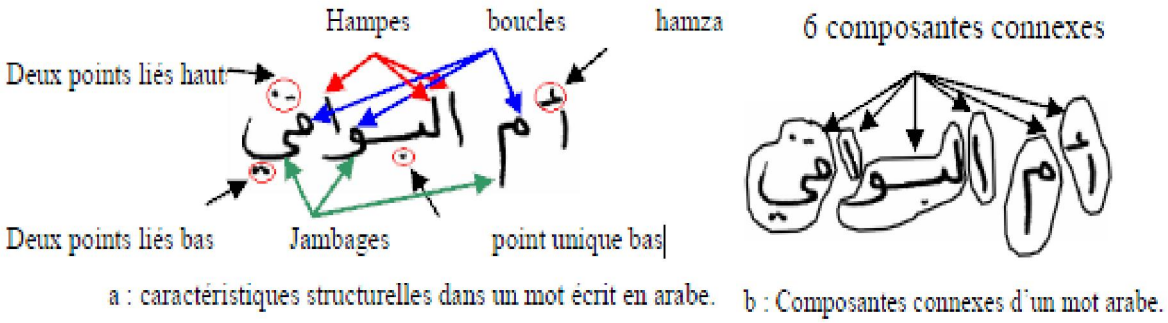


Figure1.14 : *Topologie de l'écriture arabe illustrée dans le mot « Oum-el-bouaghi »* [NEM 09]

1.3.5. Classification et reconnaissance :

La classification est l'étape noyau de reconnaissance. A partir de la description en paramètres de la forme, le moteur de reconnaissance cherche parmi les modèles de référence en présence, ceux qui lui sont les plus proches. Elle consiste à élaborer une règle de décision qui transforme les attributs caractérisant les formes en appartenance à une classe (passage de l'espace de codage vers l'espace de décision).

La décision nécessite de définir clairement la connaissance que nous avons sur les formes à traiter. Cette définition repose sur l'apprentissage qui se charge d'acquérir la connaissance et de l'organiser en classes ou modèles de référence [BEL 92].

L'apprentissage consiste en deux concepts différents : l'entraînement et l'adaptation. L'entraînement consiste à enseigner au système la description des caractères tandis que l'adaptation sert à améliorer les performances du système en profitant des expériences précédentes. Certains systèmes permettent à l'utilisateur d'identifier un caractère lorsqu'ils échouent à le reconnaître et ils utilisent l'entrée de l'utilisateur à chaque fois que le caractère est rencontré [Al- 95].

Les procédés d'apprentissage sont différents selon qu'il s'agisse de reconnaissance de caractères imprimés ou manuscrits ou de reconnaître des textes monofonte ou multifonte. D'une manière générale, on distingue deux types de techniques d'apprentissage : supervisé et non supervisé.

Dans le cas de **l'apprentissage supervisé**, un échantillon représentatif de l'ensemble des formes à reconnaître est fourni au module d'apprentissage. Chaque forme est étiquetée par un opérateur appelé professeur, cette étiquette permet d'indiquer au module d'apprentissage la

classe dans laquelle le professeur souhaite que la forme soit rangée. Cette phase d'apprentissage consiste à analyser les ressemblances entre les éléments d'une même classe et les dissemblances entre les éléments de classes différentes pour en déduire la meilleure partition de l'espace des représentations. Les paramètres décrivant cette partition sont stockés dans une table d'apprentissage à laquelle le module de décision se référera ensuite pour classer les formes qui lui sont présentées [BEN 02].

Dans le cas de **l'apprentissage non supervisé** (classification automatique), on fournit au programme un grand nombre de formes non étiquetées, le programme se chargeant d'identifier les formes appartenant à une même classe. En contre partie, le concepteur doit fournir une métrique dans l'espace de représentations afin de permettre au programme d'apprentissage de détecter des groupes de points voisins qu'on espère coïncidant au mieux avec les classes que l'on aurait à fixé préalablement. [BEN 08].

La reconnaissance peut conduire à un succès si la réponse est unique et juste (un seul modèle répond à la description de la forme). Elle peut conduire à une confusion si la réponse est erronée ou multiple (plusieurs modèles correspondent à la description). Enfin, elle peut conduire à un rejet de la forme si aucun modèle ne correspond à sa description. Dans les deux premier cas, la décision peut être accompagnée d'une *mesure de vraisemblance*, appelée aussi score ou *taux de reconnaissance* [BEL 92].

Le rôle du classifieur est de se prononcer sur l'appartenance d'une forme à chacune des classes de caractère à partir du vecteur de caractéristiques. Il existe de nombreux classifieurs possédant des caractéristiques de performances et de vitesse différentes. Les entrées du classifieur représentent les formes à reconnaître (caractères, mots, etc.) et les sorties les classes ou catégories auxquelles elles appartiennent.

Selon [Jain 00], il existe quatre grandes familles de classifieurs : le pattern matching (ou ((appariement de formes)) par une mesure de distance ou de corrélation), l'appariement structurel ou syntaxique, la classification statistique, et les réseaux de neurones. A cette classification, nous pouvons ajouter les approches stochastiques. Ces deux dernières classes peuvent être considérées comme des sous-familles de l'approche statistique.

1.4.6. Le post-traitement

Il 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. [BEL 01]

1.4. Conclusion

Les documents existants présentent de grande variabilité à différentes mesures : catégorie, type, structure, langue, etc. Ce qui nécessitent des approches et des techniques de reconnaissance diverses. Convertir ces documents papier en documents électroniques est l'objectif des systèmes de reconnaissances de documents. Cette tâche s'avère une nécessité pour offrir une manipulation plus aisée des données : archivage, indexation, recherche, etc.

Dans ce chapitre, nous avons abordé le problème d'analyse et de reconnaissance de documents sans se limiter par un type particulier de documents. Nous avons introduit la notion de document, défini le domaine d'analyse et de reconnaissance de documents. Nous avons présenté les différentes étapes qui peuvent être incluses dans le processus d'analyse et de reconnaissance documents.

Chapitre 2

Applications sur les images de documents

Nous allons dans ce chapitre voir quelques applications sur les images de documents qui sont la recherche d'information, l'extraction d'information et la catégorisation de documents.
L'application la plus connue de la recherche d'information est le moteur de recherche qui permet, à partir d'une liste de mots clés ou d'une requête, de fournir à l'utilisateur une liste de documents en correspondance avec la requête. L'intérêt d'une telle approche est de permettre une sélection automatique des documents pertinents parmi une base de documents, ou un corpus pouvant comporter plusieurs milliers de documents archivés. Le principal intérêt est donc de décharger l'utilisateur de la consultation de la base, charge à l'utilisateur d'analyser précisément le contenu des documents sélectionnés par le système pour y trouver précisément l'information qu'il recherche.
Pour les images numérisées de documents manuscrits on peut globalement distinguer deux types d'utilisation auxquelles on peut associer des requêtes de nature très différente :

• Les archives de documents manuscrits peuvent être vues sous l'angle de leurs contenus textuels. Dans ce cas l'interrogation des bases documentaires nécessite de recourir à une phase préalable de transcription des textes manuscrits permettant ensuite une analyse textuelle.

• Les archives de documents manuscrits peuvent également être vues sous l'angle de leurs contenus graphiques. Dans ce cas l'interrogation des bases documentaires est effectuée à partir de requêtes graphiques. On cherche par exemple à retrouver les documents de la base présentant certaines calligraphies correspondant à certains scripteurs. D'autres cas d'utilisation peuvent concerner la détection des différentes mains présentes, ou bien la datation des documents par rapport à la chronologie de l'œuvre de l'auteur.

L'extraction d'information, quant à elle, procède à une analyse spécifique du contenu des documents pour y détecter certains faits précis. Il existe aujourd'hui des systèmes de lecture permettant d'extraire une information manuscrite ciblée dans des documents fortement contraints. On peut notamment citer la lecture de chèques bancaires, d'enveloppes postales ou encore la lecture de formulaires. Enfin, un système de catégorisation de documents est capable d'affecter le document à un thème spécifique.

Nous commençons par un contexte général sur la recherche d'information, puis nous donnons un aperçu sur la recherche d'image et la recherche de mots en image, ensuite on présente l'extraction d'information, et enfin on donne une brève présentation de la catégorisation de documents

2.1. Contexte général

La recherche d'information est le processus de trouver un objet (en général un document) de nature non structurée (souvent) qui satisfait un besoin d'information au sein de collections de données de tailles très importantes stockées sur des supports de stockage.

La recherche d'information n'a pas commencé avec le Web mais plutôt avant. En réponse à différents problèmes de fourniture d'accès à l'information, le domaine de la recherche d'information a évolué pour donner de nouveaux principes et de nouvelles approches pour la recherche de nature variée (l'information recherchée peut être un document texte, une image, une vidéo etc.).

Chaque jours des millions d'octets sont générés et stockés sur des supports de stockage. La recherche de l'information voulue devient de plus en plus difficile et présente des difficultés majeures.

La recherche d'information ou plus précisément la recherche d'images par le contenu constitue actuellement un axe de recherche très actif. Plus de 300 papiers dans le domaine sont publiés chaque année. C'est un domaine passionnant et représente le futur des moteurs de recherche.

La recherche d'images a évolué dans le temps selon les besoins des utilisateurs. Les progrès technologiques ont beaucoup influencé cette évolution. De ce fait Del Bimbo [BIM 99] divise cette évolution en trois générations différentes:

- **La première génération :**

Dans la première génération des systèmes de recherche d'images, les images sont stockées dans une base de données, et peuvent être trouvées par le biais de chaînes de caractères qui lui sont liées. Ces chaînes de caractères peuvent être liées à un élément de l'image, ou à l'image elle-même. Ces chaînes sont stockées et peuvent être recherchées de manière structurée, comme dans les bases de données SQL. La technique utilisée dans ces systèmes souffre de limitations importantes :

Les descripteurs textuels dépendent de ce que l'utilisateur peut saisir lors de la création de la base d'images. Des utilisateurs différents peuvent décrire la même image d'une manière différente, et le même utilisateur peut décrire la même image d'une manière différente en analysant l'image une deuxième fois.

Plusieurs attributs d'image, par exemple, la texture et la distribution de la couleur, sont difficiles à décrire sans ambiguïté en utilisant les descripteurs textuels.

La saisie de chaînes de caractères dans une base de données exige beaucoup d'effort, et nécessite trop de temps. La recherche dans ce cas se fait selon l'approche textuelle traditionnelle des bases de données relationnelles.

- **La deuxième génération :**

La deuxième génération de systèmes de recherche d'images offre différentes manières d'interroger la base d'images, permettant des recherches selon les attributs visuels des images comme la texture, la forme et la couleur. Cette approche (basée sur les attributs visuels) peut être combinée avec la recherche basée sur le texte. La recherche dans les systèmes de cette génération se base sur une mesure de similarité qui classe les images selon leur degré de similitude à l'image requête.

- **La troisième génération :**

Ces systèmes sont encore en cours de réalisation. Ils sont supposés être capables de travailler d'une manière intelligente, semblable au fonctionnement du système visuel humain. Ces systèmes devaient apprendre à partir de quelques exemples et de tirer des conclusions fondées sur l'expérience. Ces systèmes restent encore hypothétiques car la connaissance du système visuel humain est limitée.

2.2. Recherche d'image et recherche de mot en image

Une immense quantité d'images de documents se fait archivés par les bibliothèques numériques, il existe un besoin pour une recherche efficace des stratégies afin de les rendre disponibles selon le besoin des utilisateurs.

Dans cette partie en va essayer de donner la définition le domaine de la recherche d'image qui fait l'un des domaines de recherche dont l'intérêt s'est augmenté en raison de la croissance rapide de la World Wide Web et la nécessité de trouver une image de notre choix à partir d'une énorme collection d'images . Ainsi qu'au repérage de mot en image qui est le domaine

d'analyse de document qui repose principalement sur la localisation et l'extraction des mots dans les images de documents manuscrits.

2..2.1. Recherche d'image

La recherche d'images est la tâche de récupérer les images numériques à partir d'une base de données. Les systèmes de récupération d'images diffèrent dans la manière dont l'interrogation et l'extraction se fait.

Les types possibles de requêtes sont par :

- **description textuelle** (meta-données) : recherche à partir des données textuelles avec des requêtes et retourne des images.
- **Information visuelle** : recherche fondé sur le contenu d'image avec les données visuelles et retourne des images.
- **La combinaison des deux**

L'approche la plus ancienne (antérieure à l'apparition des images numériques) et encore majoritairement employée aujourd'hui est l'approche par annotation de meta-données telles que le créateur de l'image, format d'image, date de création, et les descriptions des objets simples de titres ou de mot-clé. Les conditions de prise de vue et des informations variées dépendant du domaine considéré. Dans la limite de l'information que peut porter l'annotation, ce type d'indexation est conçu pour répondre à des types de requêtes spécifiques et prédéfinis. Les inconvénients majeurs de l'annotation sont : la nécessité de l'intervention d'un humain (pénible sur de grandes bases), leur rigidité (l'ajout ou la suppression des champs de meta-données sur une base entière représente un travail colossal), leur subjectivité (deux personnes annoteront-elles une image donnée avec les même mots-clés ?), les contraintes linguistiques (passage d'une langue à une autre, ambiguïté sémantique). De plus, notons que l'annotation ne pourra jamais d´écrire le contenu d'une image de façon exhaustive.

Bien que cette approche réduise considérablement le travail nécessaire à une assignation manuelle des mots-clés, il faut se rappeler que beaucoup d'images ne sont pas accompagné de texte. En outre, les besoins des utilisateurs d'images peuvent se produire à un niveau primitif qui s'exploite directement dans les attributs visuels d'une image. Ces attributs peuvent être mieux représentés par des modèles d'image et récupérées par les systèmes d'interprétation motif correspond à la couleur, texture, forme, et d'autres caractéristiques visuelles.

En même temps que l'essor récent des bases d'images numériques, une alternative à l'annotation manuelle est apparue il y a une dizaine d'années : la recherche d'images par le contenu (ou CBIR en anglais pour content-based image retrieval) d'où l'interrogation se fait par information visuelle. CBIR est l'application de la vision par ordinateur pour le problème d'extraction d'images. L'information visuelle peut être une image exemple (requête par exemple), mais il peut aussi être une esquisse du résultat souhaité ou une description des propriétés de l'image comme la proportion des couleurs souhaitée (50% rouge, 30% vert, etc.)

La recherche d'images par le contenu consiste à caractériser le contenu visuel des images par des descripteurs visuels et d'effectuer des recherches par similarité visuelle à partir de ces descripteurs. Le problème est si on utilise un descripteur visuel peu informatif ou bien peu fidèle, les images retournées ne présenteront aucun intérêt pour l'utilisateur.

Les systèmes CBIR actuels souffrent de fossé sémantique. Même si une rétroaction des utilisateurs est proposée comme un remède à ce problème, il conduit souvent à la distraction dans la recherche.

Pour remédier à ces inconvénients, de nouveau système interactif d'extraction de l'image en été proposé qui intègre textes et contenu de l'image pour améliorer la précision de récupération. Beaucoup de systèmes d'extraction d'images permettre d'affiner les résultats de recherche par le retour de pertinence. Cela signifie que l'utilisateur peut juger les images résultantes comme pertinent ou non à la requête, puis répéter la recherche avec ces informations supplémentaires.

Plus généralement, la pertinence est directement liée aux techniques choisies pour les différents éléments qui composent un système de recherche par le contenu, tels que : descripteurs (couleur, texture, forme), mesure de similarité visuelle, mode de requête, mode de représentation des images (global, partiel). Bien que de nombreux travaux sur ces différents aspects ont été proposés dans la littérature, la recherche d'images par le contenu demeure encore aujourd'hui un problème ouvert et très actif [FAU 04].

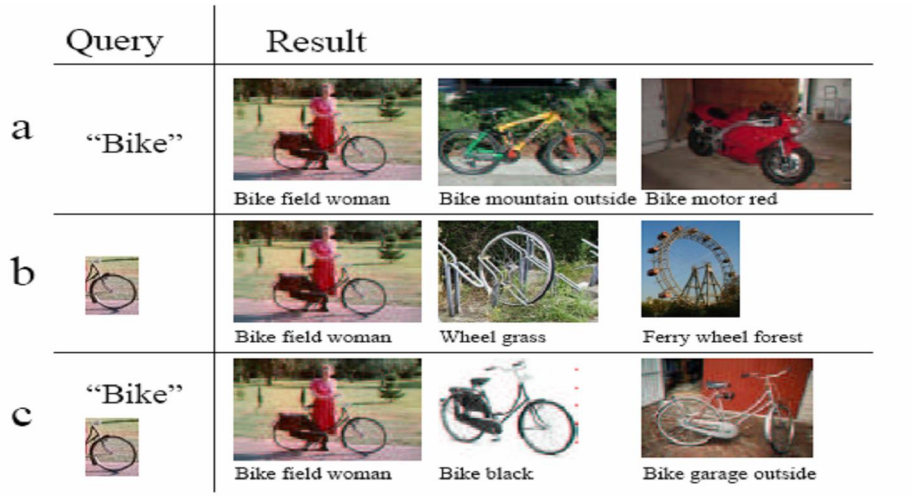


Figure 2.1 : a) requête textuel, b) exemple d'image, c) texte et image

2.2.2. Recherche de mots en image :

Repérer des mots dans les documents imprimés anciens est une tâche extrêmement difficile. Les méthodes classiques, comme la corrélation, échouent quand elles sont appliquées sur les documents anciens.

La recherche de mots par similarité de formes (le terme word-spotting est plus souvent employé) est une technique permettant de localiser des mots choisis par un utilisateur dans un texte, écrit ou parlé, sans aucune contrainte. Cette approche générique peut être appliquée à tout type de document écrit, quel que soit son langage et qu'il utilise un alphabet, un syllabaire ou des idéogrammes. Il n'est pas nécessaire de créer une base d'apprentissage adaptée à chaque document ou à chaque scripteur. Cette technique est utilisée lorsque la reconnaissance de mots est mise en échec, comme par exemple sur les documents très détériorés ou les manuscrits [LEY 06].

L'échec des méthodes de reconnaissance optique de caractères (OCR) pour transcrire les manuscrits anciens a conduit à rechercher des méthodes d'indexation de l'information n'utilisant que l'information image. Le principe de base de l'approche (word-spotting) est de choisir un prototype, ou mot-clé image (imagette), et d'en rechercher toutes les occurrences au sein du document en ne tenant compte que de l'apparence visuelle du mot [JOU 09].

Deux principaux types d'approches de word-spotting peuvent être identifiés, selon la façon dont l'entrée est spécifiée: la requête par chaîne et la requête par exemple. Dans une requête par des approches-string, des modèles de caractères ont été formés à l'avance et au moment de la requête les modèles des caractères formant la chaîne

sont concaténées dans un modèle de texte et la probabilité de chaque mot «image» est évaluée. Ainsi, ces approches sont très similaires aux systèmes de HWR. Une fois formés, ces approches permettent de rechercher n'importe quel mot-clé possible. Toutefois, ils présentent des inconvénients similaires aux systèmes de HWR.

Dans une requête par des approches-par exemple, l'entrée est une image du mot, et le résultat est un ensemble d'images mot qui sont les plus semblables en apparence à l'image requête. Cela peut être considéré comme une recherche d'images basé sur le contenu (CBIR) tâche. La recherche d'un mot-clé souhaité est sous réserve d'avoir une image exemplaire de ce mot-clé disponible. Parce que le résultat est basé sur une mesure de distance entre la requête et toutes les images mot candidat, aucune formation n'est en cause. Toutefois, la performance est limitée.

Beaucoup d'œuvres qui appartiennent à cette catégorie ont montré des performances acceptables sur des ensembles de données qui contiennent des données à partir d'un seul ou quelques écrivains. Toutefois, on ignore encore comment ces méthodes se comporteraient dans des conditions multi-écrivain. Dans certains cas, la performance peut être augmentée en interrogeant plusieurs fois avec des images différentes et en combinant les résultats. S'il est intuitif que d'un modèle statistique pourrait améliorer les performances en combinant les différentes requêtes dans un modèle unique, à la connaissance des auteurs de cette option est restée inexplorée dans la littérature de recherche de mots dans les documents manuscrits par word-spotting [ROD 09].

2.3. Extraction d'information dans les documents

L'extraction d'information dans tous les documents se déroule suivant les mêmes étapes. Une première étape de prétraitement permet de localiser au mieux l'information que l'on doit extraire. Selon les types de documents, cette localisation est plus ou moins délicate. Différentes stratégies de reconnaissances dédiées à la problématique sont ensuite mises en œuvre afin d'extraire les informations recherchées. Ces stratégies exploitent au maximum les contraintes du cadre d'application afin de fiabiliser la reconnaissance **[GUI 06].**

2.3.1. Localisation des informations

La localisation consiste à isoler toutes les composantes et seulement les composantes d'une entité que l'on cherche à identifier dans le cas d'un mot, d'une séquence de mots, d'une phrase ; ou d'un texte dans un document. La localisation se traduit donc par une étape de segmentation du document en entités distinctes.

La localisation des informations manuscrite dans les documents est un problème épineux dans la mesure où il est directement confronté au dilemme de Sayre qui stipule que dans un problème de reconnaissance de formes, la localisation et la reconnaissance des entités ne peuvent être dissociées pour être menées correctement [SAY 73]. Il existe plusieurs manières de contourner ce problème, en fonction des connaissances *a priori* que possède le système au sujet des documents traités. A partir des connaissances a priori, il est possible de constituer un modèle de document plus ou moins figé définissant l'organisation des informations à l'intérieure de ce document: structure physique, nature et position des informations, présence de repères ou symboles connus à des emplacements précis, connaissances syntaxiques réagissant tout ou partie de l'information recherchée. Toutes les méthodes de localisation reposent sur l'exploitation d'un modèle de document. On peut distinguer deux cas de figure suivant le niveau de contraintes qu'apportent les connaissances a priori [CHA 06] :

- Lorsque l'on dispose de connaissances *a priori* en quantité suffisante, le modèle de document est suffisamment contraint pour réaliser une localisation des informations en se basant sur le modèle. C'est le cas des applications de lecture automatique de chèques bancaires, de formulaires ou d'adresses postales, où les différentes entités recherchées sont facilement localisées, généralement sans faire appel à la reconnaissance.

- Lorsque les connaissances *a priori* sont trop faibles, le modèle de document n'est pas suffisamment contraint pour effectuer une localisation directe des informations. C'est le cas des textes libres qui ne possèdent pas de structure physique stable. Dans ce cas, la localisation des informations pose de nouveaux problèmes : une segmentation des entités manuscrites est nécessaire afin d'identifier les mots du texte. On connaît la difficulté d'une telle opération, et puisque le paradoxe de Sayre devient dans ce cas incontournable, la phase de reconnaissance doit être liée à la phase de localisation pour fournir des résultats fiables. A partir de ce constat, une stratégie alternative visant à extraire l'information des documents commence à émerger. Il ne s'agit plus de considérer une lecture intégrale du document mais plutôt d'effectuer une reconnaissance partielle visant à extraire l'information pertinente.

Les Techniques de binarisation, qui utilisent un seuillage global, local, ou d'adaptation, sont les méthodes les plus simples pour la localisation du texte. Ces méthodes sont largement utilisées pour la segmentation d'image du document, car ces images sont généralement des caractères noirs sur fond blanc, ce qui permet une segmentation réussie basée sur un seuillage. Cette approche a été adoptée pour de nombreuses applications spécifiques telles que

l'emplacement adresse sur le courrier postal, le montant sur les chèques, etc, en raison de sa simplicité de mise en œuvre [JUN 04].

La localisation se traduit donc par une étape de segmentation du document en entité distinctes c'est-à-dire la détermination de l'emplacement du texte dans l'image et la génération des boîtes englobantes autour du texte.

Nous présentons dans ce qui suit quelques systèmes de localisation de l'information manuscrite proposés dans la littérature dans les documents contrains (formulaires et adresses postales) et dans les documents non contrains. [CHA 06] :

2.3.2. Localisation de champs d'intérêt dans les formulaires

Les formulaires contenant des informations manuscrites possèdent généralement une structure totalement statique, autorisant une localisation immédiate des zones d'intérêt. `A partir de ces zones d'intérêt, des délimiteurs matérialisant la zone dans laquelle le scripteur doit écrire permettent d'identifier facilement les composantes appartenant au champ recherché : cases prédéfinies dans le cas de précasé, zone identifiant la région contenant l'information, ligne de base sur laquelle le scripteur doit remplir le champ. L'application d'un simple calque peut ainsi parfois suffire à extraire les informations recherchées.

2.3.3. Localisation d'entités dans les adresses postales

Dans les nombreuses applications de lecture des adresses postales développées récemment, la localisation des informations a lieu à deux niveaux. Dans un premier temps, une localisation du bloc adresse est effectuée. Dans un second temps, une interprétation du bloc adresse est réalisée afin de localiser le code postal, le bureau distributeur ou un nom de rue. La localisation du bloc adresse peut paraitre aisée sur des enveloppes ((propres)), mais il existe de nombreuses enveloppes contenant des images ou des messages publicitaires en plus du timbre et du tampon postal. Les techniques employées ne font généralement pas intervenir les connaissances a priori sur la position du bloc adresse ; elles reposent plutôt sur des approches géométriques analysant la taille et la disposition des boites englobantes des composantes connexes, ou sur des approches à base de détection de texture qui distinguent les zones ((écriture)) des zones ((fond)). Contrairement à la détection du bloc adresse dans l'enveloppe, les méthodes mises en œuvre pour la localisation des champs d'intérêt dans le bloc adresse reposent sur l'exploitation d'un certain nombre de contraintes régissant la structure des adresses postales sur les enveloppes : le bloc adresse est disposé en lignes dont le nombre peut

varier, et dans lesquelles on retrouve toujours les champs prénom, nom, numéro et nom de rue, code postal et nom de ville.

2.3.4. Localisation/reconnaissance de mots dans des textes libres

Il y a quelques années sont apparus les premiers travaux concernant la lecture de textes manuscrits dits ((libres)). Lorsque des textes libres pleine page sont traités, le modèle physique de document est peu contraint : la structure, le contenu et l'objet du document sont inconnus. La seule contrainte généralement connue est une orientation privilégiée des lignes de texte. On ne peut donc plus exploiter les connaissances a priori sur la disposition physique des entités pour localiser l'information. Contrairement aux applications de lecture automatique de chèques ou d'adresses postales où l'industrialisation a motivé les recherches, les besoins applicatifs vis-à-vis des textes libres ne sont pas encore parfaitement identifiés. Il est donc difficile de savoir ce que l'on cherche à localiser et à reconnaître. Actuellement, les travaux portent donc essentiellement sur la reconnaissance intégrale de textes dont le lexique plus ou moins grand est supposé connu. Dans tous ces travaux, on procède à une segmentation du document en lignes sans reconnaissance. Deux stratégies peuvent ensuite ˆêtre utilisées pour réaliser la segmentation des lignes en mots :

– La première stratégie consiste à effectuer la localisation sans reconnaissance. La segmentation en mots est généralement effectuée par une analyse des espaces entre composantes afin de distinguer les espaces inter-mots des espaces interlettres. Des méthodes de reconnaissance de mots isolés sont ensuite appliquées sur les hypothèses de segmentation. Une erreur lors de l'étape de segmentation ne peut donc être rattrapée.

– La deuxième stratégie que l'on peut qualifier de ((localisation/reconnaissance)) consiste à réaliser conjointement la segmentation et la reconnaissance sur l'ensemble de la ligne de texte. On peut ainsi voir cette stratégie comme une extension à la ligne de texte des méthodes de segmentation implicite ou de segmentation-reconnaissance mises en œuvre à l'échelle du mot. Plutôt que de considérer des décisions locales de segmentation ne prenant pas en compte le contexte, ce type d'approche propose des solutions de segmentation/reconnaissance sur l'ensemble de la ligne de texte. L'inconvénient de cette stratégie réside toutefois dans l'explosion combinatoire engendrée par la multiplication des hypothèses de segmentation/reconnaissance sur une ligne de texte.

Nous donnons maintenant des exemples de ces deux stratégies.

2.3.5. Localisation des mots sans reconnaissance

Dans [MAR 01a, NOS 02], une reconnaissance de textes libres où certaines contraintes sont imposées aux scripteurs est présentée. Du fait de ces contraintes, il n'existe pas d'applications industrielles mettant en œuvre de tels documents, et les travaux sont essentiellement académiques. Il s'agit en particulier de contraintes d'espacement entre les lignes et d'espacement entre les mots, afin de faciliter la segmentation en lignes et la segmentation en mots (voir ***figure 2.2***). Nous qualifions par la suite ces documents de textes faiblement contraints.

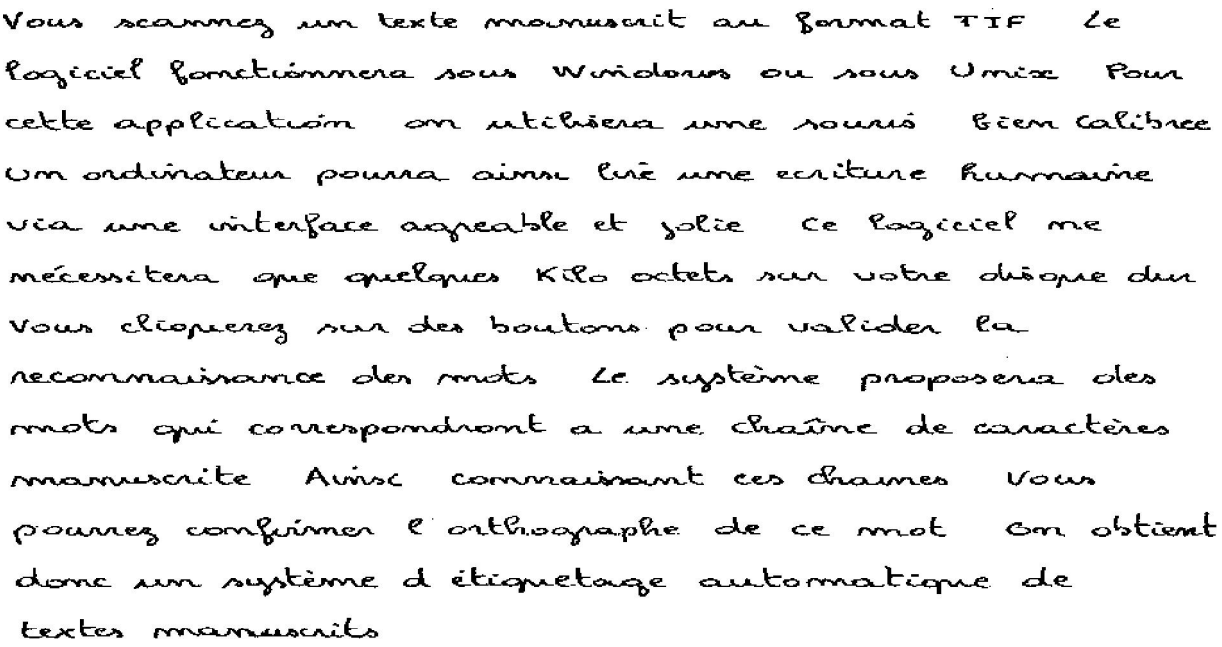


Figure 2.2 : Exemples de textes manuscrits français traités dans [NOS 02] et [MAR 01a]. Certaines contraintes d'espacement inter-lignes et inter-mots ont été imposées aux scripteurs.

Dans ce cas, on ne dispose pas d'un modèle physique de document, mais des connaissances a priori sur les espacements entre les lignes et entre les mots permettent d'effectuer une localisation des mots sans faire appel à la reconnaissance. Dans [MAR 01a] et [NOS 02], les documents exploités ont été produits en imposant deux contraintes aux scripteurs : premièrement, les lignes de texte doivent être suffisamment espacées de telle sorte qu'elles sont parfaitement séparables par une simple recherche de lignes horizontales de pixels blancs. Deuxièmement, on impose aux scripteurs de suffisamment séparer les mots pour que les espaces entre deux mots (espaces inter-mots) soient toujours plus grands que des espaces séparant deux lettres d'un même mot (espaces intra-mots). Sous cette contrainte, la tache de

segmentation d'une ligne de texte en mots consiste à estimer un seuil maximum au-delà duquel les espaces entre deux composantes seront considérés comme espaces inter-mots.

L'approche proposée dans [SRI 93] segmente les lignes de texte en mots. Chaque mot est ensuite soumis à un moteur de reconnaissance dont on conserve les N meilleures propositions. Le treillis de reconnaissance de la ligne est alors exploré en considérant des contraintes linguistiques d'ordre grammatical (nom, verbe, adjectif, ...). Le principal inconvénient de cette approche est d'enchainer les traitements séquentiellement. Ainsi, une erreur de segmentation commise en amont est fatale pour la reconnaissance.

Nous pouvons constater que les travaux mettant en œuvre une localisation des mots sans reconnaissance dans des documents dont la structure physique est inconnue sont limités. Dans [MAR 01a] et [NOS 02], les contraintes imposées aux scripteurs ne sont que rarement respectées dans le cas de documents réels non destinés à être lus par un système de lecture automatique de documents. Dans [SRI 93], une segmentation des lignes de texte en mots sans reconnaissance est également effectuée, et l'exploitation de connaissances grammaticales ne peut corriger toutes les erreurs faites lors de la phase de segmentation.

2.3.6. Localisation/reconnaissance de mots

Pour remédier au problème difficile de segmentation des lignes en mots dans le contexte de textes libres, les approches proposées dans [MAR 01b] et [VIN 04] ne réalisent pas de segmentation préalable de la ligne en mots. Dans les deux cas, les méthodes développées sont testées sur la base IAM [MAR 99] comportant des textes relativement propres. Après une étape de segmentation du document en lignes, les lignes de texte sont considérées dans leur intégralité, et une décision globale de localisation/reconnaissance sur l'ensemble de la ligne est effectuée. La localisation/reconnaissance est réalisée par une approche à segmentation implicite. Des modèles de lignes de textes sont réalisés grâce à des HMM de lettres, concaténés pour former des modèles de mots, eux-mêmes concaténés pour former un modèle de ligne. Lors du décodage, la segmentation et la reconnaissance des mots sur l'ensemble de la ligne de texte sont alors réalisées simultanément par l'algorithme de Viterbi. Ces modèles de ligne considèrent donc qu'une ligne de texte est composée uniquement de mots connus, ce qui impose de travailler avec de grands lexiques : jusqu''a 50 000 mots dans [VIN 04], et plusieurs milliers dans [MAR 01b].

2.4. Catégorisation :

La catégorisation a pour objectif de regrouper les documents similaires, c'est à dire thématiquement proches, au sein d'un même ensemble. Un système de catégorisation va donc, à partir uniquement des mots contenus dans un document, tenter de détecter le ou les thèmes abordés dans celui-ci. L'intérêt d'une telle démarche est d'organiser les connaissances de façon à pouvoir effectuer, par la suite, une recherche ou une extraction d'information efficace.

Un système de catégorisation est constitué de trois composantes principales (***figure 2.3***) [KOC 06] :

– les prétraitements ont pour but de convertir le flux de caractères entrants en un flux de termes. Ces termes sont obtenus après filtrage des mots vides («stopword») et «stemming» ;
– une représentation vectorielle des documents permet de convertir la liste de termes en un vecteur de valeurs décrivant le document ;
– un classifieur va enfin se prononcer sur l'appartenance d'un document à chaque thème.

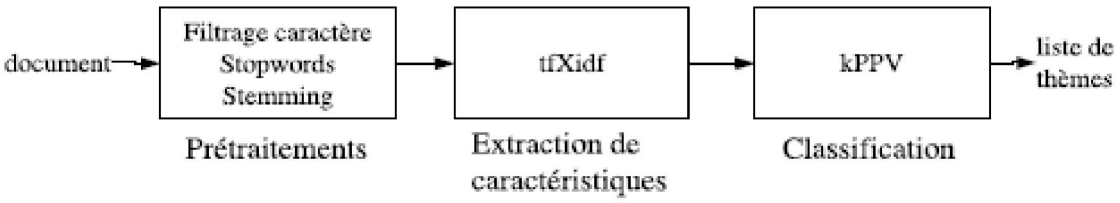


figure2.3 : Processus pour la catégorisation d'un document [KOC 06]

La mise en place d'un système de catégorisation de documents manuscrits n'est pas aussi immédiate que pour les documents imprimés. Contrairement aux documents imprimés, les lettres d'un même mot manuscrit sont reliées entre elles. Les lettres n'étant plus clairement segmentées, leur reconnaissance devient délicate. Cette problématique est clairement énoncée dans [SAY 73] : «pour reconnaître une entité, il faut savoir la localiser ; pour la localiser, il faut tout d'abord la reconnaitre». Ainsi, la segmentation d'un mot en lettres doit être réalisée en même temps que sa reconnaissance. [KOC 06]

A cela vient s'ajouter une mise en page plus souple sur le manuscrit que sur l'imprimé. Les documents manuscrits sont par nature plus variables que les documents imprimés. Ainsi, les lignes de texte, tout en restant globalement horizontales, présentent des fluctuations dans leurs lignes de base : il est courant de voir des lignes légèrement incurvées et inclinées. De plus, si l'on considère la dimension des interlignes proportionnellement à celle des mots, les lignes de

texte des documents manuscrits sont plus serrées que dans les documents imprimés. Il est donc nécessaire de mettre en place des méthodes d'analyse de mise en page dédiées aux documents manuscrits. [KOC 06]

Cette variabilité inhérente aux documents manuscrits se retrouve également dans l'espacement des mots dans une ligne. Dans le cas de l'imprimé, la distance séparant deux mots est toujours supérieure à celle séparant deux lettres d'un même mot. Ceci n'est plus systématique dans le cas du manuscrit, ce qui rend la localisation des mots dans les lignes de texte hasardeuse si l'on ne fait pas intervenir un processus de reconnaissance. [KOC 06]

2.5. Conclusion :

Généralement les humains recherchent de l'information dans des documents quand ils n'ont pas la connaissance et le temps nécessaire à la réalisation d'une tâche. Les mémoires artificielles (les documents) joueraient le rôle de complément aux mémoires naturelles (les connaissances d'un individu, d'un groupe) dans le contexte de la réalisation d'une tâche. La recherche d'information dans les documents pourrait ainsi être considérée comme une alternative à la résolution de problèmes, autre activité mise en œuvre par l'humain quand il manque de connaissances pour réaliser une tâche. Dans ce chapitre nous avons présenté le domaine de la recherche d'images ainsi que la recherche de mot en image de documents. Nous avons donné tout d'abord quelques définitions et notations nécessaires des différents systèmes de récupération d'images puis les différents types de requêtes possibles en ce qui concerne la recherche de mot en image. Ensuite, nous avons passé à l'extraction et la localisation des informations manuscrite dans les documents donnant par la suite quelques systèmes de localisation de l'information manuscrite contraint comme les formulaires et les adresses postale, et enfin nous avons donné un aperçu sur le système de catégorisation.

Chapitre 3
Conception

Pour rechercher des informations il faux un processus de recherche, dans une base de documents, des documents qui sont considérés pertinents au sens d'un besoin exprimé par l'utilisateur, sous la forme d'une requête. Pour cela, la requête et les documents de la base sont généralement représentés dans un même espace de caractéristiques. De ce fait, le choix des caractéristiques est particulièrement primordial. Comme les documents doivent être décrits de façon à pouvoir répondre à tout type de requête, on ne peut en général faire intervenir une quelconque étape de sélection de caractéristiques.

En effet, la reconnaissance optique de caractères ne peut être appliquée que sur des documents imprimés ou des documents manuscrits avec un lexique limité. Dés que les documents soient dégradés, complexes aux points de vue structure et disposition spatiale du contenu textuel et graphique et avec un lexique plus large, l'OCR devient inefficace. Dans le cas des documents arabes manuscrit qui sont le cœur de notre travail, la difficulté s'augmente à cause des caractéristiques morphologiques de l'écriture arabe qui compliquent de plus la tâche de l'OCR à différents niveaux du traitement.

Nous proposons une approche de recherche de mots dans des images de documents manuscrits arabes sans recourir à une reconnaissance du contenu.

Le présent chapitre est consacré à la présentation de notre approche. Tous d'abord, nous présentons quelques caractéristiques de l'écriture arabes pour montrer la complexité du traitement de ce genre de documents. Dans le reste du chapitre, nous décrivons la méthodologie développée pour la conception du système proposé, Avant de conclure.

3.1. Caractéristiques de l'écriture arabe :

L'arabe s'écrit de la droite vers la gauche. L'écriture est semi-cursive soit sous forme imprimée ou manuscrite. Le concept de majuscule et minuscule en écriture arabe n'existe pas. L'alphabet arabe comporte 28 lettres (voir le ***tableau 3.1***) La forme des lettres dépend de leur position dans le mot. Certaines lettres prennent jusqu''a 4 formes différentes : par exemple le

(ع ـعـ ـع) ou le (ه ـه ـه). Mais pour la plupart des lettres, les formes début/milieu et fin/isolé sont identiques à la ligature près. La présence d'une ligature avec la lettre précédente ou avec la lettre suivante ne modifie pas la forme de la lettre de manière significative (pas plus que dans l'´ecriture manuscrite cursive latine). Les ligatures se situent toujours au niveau de la ligne d'´ecriture, c'est `a dire qu'il n'existe pas de lettre à liaison haute comme le 'o' ou le 'v' en alphabet latin. [MEN 08]

Certains caractères arabes incluent une boucle qui peut avoir différentes formes (***figure 3.1***).

Figure 3.1 *: Exemple de différentes formes de la boucle*

Dans l'alphabet arabe, 15 lettres parmi les 28 possèdent un ou plusieurs points. Ces signes diacritiques sont situés soit au-dessus, soit en dessous de la forme à laquelle ils sont associés, mais jamais les deux à la fois. La ***figure 3.2*** illustre la variabilité des styles d'´ecriture des points ou groupes de points en écriture manuscrite arabe. Un groupe de deux points peut ainsi s'´ecrire sous forme d'une seule, ou de deux composantes connexes. On remarque la très forte similarité entre deux points reliés par un trait, et une voyelle de type 'A' ou 'I' dont les exemples sont donnés ***figure 3.3***. Un groupe de trois points peut donner lieu à une, deux ou trois composantes connexes, en fonction du style d'´ecriture. **[MEN 08]**

En général on ne représente pas les voyelles, sauf dans les manuels scolaires. L'absence de voyelles peut toutefois être source de confusions. Comme le rappellent Y. Bahou et al dans [203], un mot peut avoir plusieurs voyellations possibles et par conséquent plusieurs catégories grammaticales. Dans certains cas, une phrase peut donc avoir deux voyellations différentes, ce qui nous donne deux structures syntaxiques possibles. Par exemple, dans le cas : يخشى الأستاذ الطلبة qui peut se voyeller des deux manières suivantes **[MEN 08]** :

يَخْشَى الأُسْتَاذُ الطَلَبَةَ : L'enseignant craint les ´étudiants.

يَخْشَى الأُسْتَاذَ الطَلَبَةُ : Les étudiants craignent l'enseignant.

Etiquettes des lettres	isolé	début	milieu	fin	Etiquettes des lettres	isolé	début	milieu	fin
Alif	ا		ا		Daad	ض	ضـ	ـضـ	ـض
Baa	ب	بـ	ـبـ	ـب	Thaaa	ط	طـ	ـطـ	ـط
Taaa	ت	تـ	ـتـ	ـت	Taa	ظ	ظـ	ـظـ	ـظ
Thaa	ث	ثـ	ـثـ	ـث	Ayn	ع	عـ	ـعـ	ـع
Jiim	ج	جـ	ـجـ	ـج	Ghayn	غ	غـ	ـغـ	ـغ
Haaa	ح	حـ	ـحـ	ـح	Faa	ف	فـ	ـفـ	ـف
Xaa	خ	خـ	ـخـ	ـخ	Gaaf	ق	قـ	ـقـ	ـق
Daal	د		د		Kaaf	ك	كـ	ـكـ	ـك
Thaal	ذ		ذ		Laam	ل	لـ	ـلـ	ـل
Raa	ر		ر		Miim	م	مـ	ـمـ	ـم
Zaay	ز		ز		Nuun	ن	نـ	ـنـ	ـن
Siin	س	سـ	ـسـ	ـس	Haa	ه	هـ	ـهـ	ـه
Shiin	ش	شـ	ـشـ	ـش	Waaw	و		ـو	
Saad	ص	صـ	ـصـ	ـص	Yaa	ى	يـ	ـيـ	ـى

(a) [SLI 09]

caractère	initiale	médiane	finale	isolé
Ta marbouta			ـة	ة
Lamalif			ـلا	لا

(b) [BEN 09]

caractère	initiale	médiane	finale	isolé
Alif+~			ـآ	آ
Alif+ء			ـأ	أ
			ـإ	إ
Waw+ء			ـؤ	ؤ
Ya+ء	ـئـ	ـئـ	ـئ	ئ

(c) [BEN 09]

Tableau 3.1 : *(a) – l'alphabet arabe dans ses différentes formes. (b) – les caractères additionnels (c) – Hamza et Madda et les positions qu'elles occupent en association avec Alif, Waw et Ya*

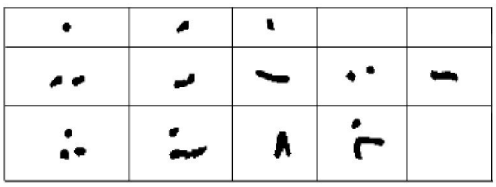


Figure 3.2 : *Points en arabe: un, deux ou trois points.* **[MEN 08]**

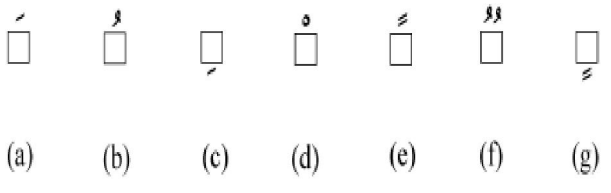
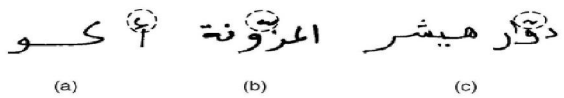


Figure 3.3 : *Voyelles en arabe : (a) A, (b) OU, (c) I (d) -, (e) AN, (f) OUN, (g) IN.* **[MEN 08]**



(a) (b) (c)

Figure 3.4 : *Autres signes diacritiques : (a) hamza, (b) chadda, (c) madda* **[MEN 08]**

Comme dans l''ecriture latine, l''ecriture arabe contient des ascendants et des descendants (voir ***figure 3.5***). En arabe, les descendants peuvent se prolonger horizontalement sous la bande de base, ce qui introduit une superposition verticale entre la lettre qui comprend le descendant et la lettre suivante. **[MEN 08]**

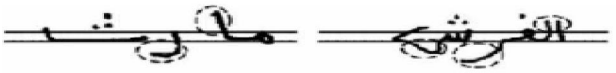


Figure 3.5 : *Les ascendants et descendants sont entourés. La bande de base est donnée à titre indicatif.* **[MEN 08]**

Une ou plusieurs composantes connexes par mot arabe, 6 lettres ne sont pas liées à leur successeur : أ, و, ز, ر, ذ, د. Ces lettres introduisent donc une coupure dans le mot. Un pseudo-mot est une unité connexe regroupant une ou plusieurs lettres sous forme d'une séquence. Un mot peut être composé d'un ou plusieurs pseudo-mots. **[MEN 08]**

En manuscrit, l'espacement entre les différents pseudo-mots d'un même mot n'est pas forcément systématiquement supérieur à l'espacement entre deux mots différents, ce qui pose parfois des problèmes de segmentation. Lorsqu'une des 22 lettres (28 moins les 6 qui ne se lient pas avec la suivante) apparaît dans sa forme "fin de mot" ou "isolée", cela signifie obligatoirement que l'on arrive à la fin d'un mot. Remarque : les lettres (ظ et ط) sont les seules parmi les 22 à ne pas prendre une forme différente lorsqu'elles sont isolées ou en fin de mot. **[MEN 08]**

Par ailleurs, les articles (le, la, les) font partie du mot auquel ils sont rattachés. La séquence ال (un pseudo mot qui contient la lettre ا isolée, suivi d'un autre pseudo-mot qui commence par la lettre ل) correspond nécessairement au début d'un mot. **[MEN 08]**

Les caractères d'une même composante connexe peuvent être ligaturés horizontalement ou verticalement. Dans certaines fontes, on peut aller jusqu'à quatre caractères ligaturés verticalement. Ceci rend la segmentation à priori en caractères quasi-impossible. La ***figure 3.6*** montre une ligature verticale de trois caractères Laam, Miim et Jiim. Les chevauchements verticaux peuvent se produire par l'intersection des composantes connexes (pseudo-mots, voir

figure 3.6) ou des mots pour quelques combinaisons de caractères. Les chevauchements et ligatures dépendent de la fonte utilisée. **[SLI 09]**

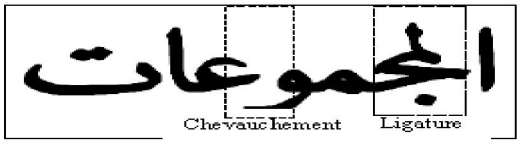


Figure 3.6 : *Exemple de mot arabe présentant une ligature verticale et un chevauchement* **[SLI 09]**

L'écriture Arabe est connue pour sa richesse en fontes et styles, elle varie selon les milieux et les régions. Il existe environ 450 fontes d'écritures différentes dont seulement quelques-unes sont couramment utilisées dans le monde arabo-musulman, nous citons à titre d'exemple : le Neskhi, Thoulthi, Roqa, Diwani, Koufi, Farsi… Le Neskhi demeure aujourd'hui la fonte la plus utilisée pour l'écriture imprimée. Chaque style arabe est régi par des lois particulières. D'un style à un autre, les proportions d'une même lettre et son dessin peuvent changer considérablement. La ***figure 3.7*** montre un exemple de différents styles graphiques de l'écriture arabe.

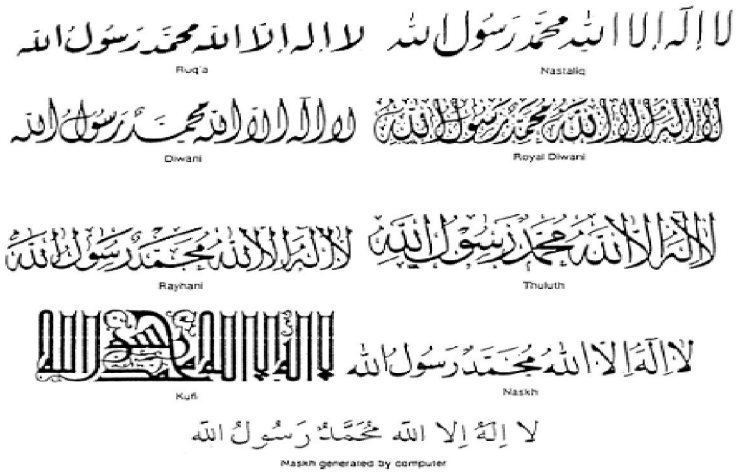


Figure 3.7: *Différents styles et fontes pour l'écriture arabe*

3.2. Architecture du système proposé :

L'objectif que nous nous sommes assignés s'articule autour du développement d'un système de recherche de mots dans des images de documents arabes manuscrits sans recourir à une reconnaissance du contenu.

Afin de réaliser notre système on procède comme suit : nous commençons par une phase de traitement et d'analyse des images de documents et ensuite nous passons à une phase de recherche dont laquelle on exploite les résultats de la phase précédente pour répondre à une requête de l'utilisateur.

3.2.1. La 0

Cette phase a pour objectif d'analyser les images composent notre base afin d'en extraire leurs caractéristiques. Elle commence par l'acquisition des documents (images d'enveloppes postales algériennes), les prétraiter, les segmenter en lignes et en mots et détecter leurs contours. Ces derniers seront codés par le code de Freeman avec l'extraction du nombre de formes par la suite et enregistrés pour former une base de fichiers de codes. La phase de recherche se fait sur cette base de codes et non pas sur les documents eux-mêmes. Cette première phase est illustrée par la figure suivante.

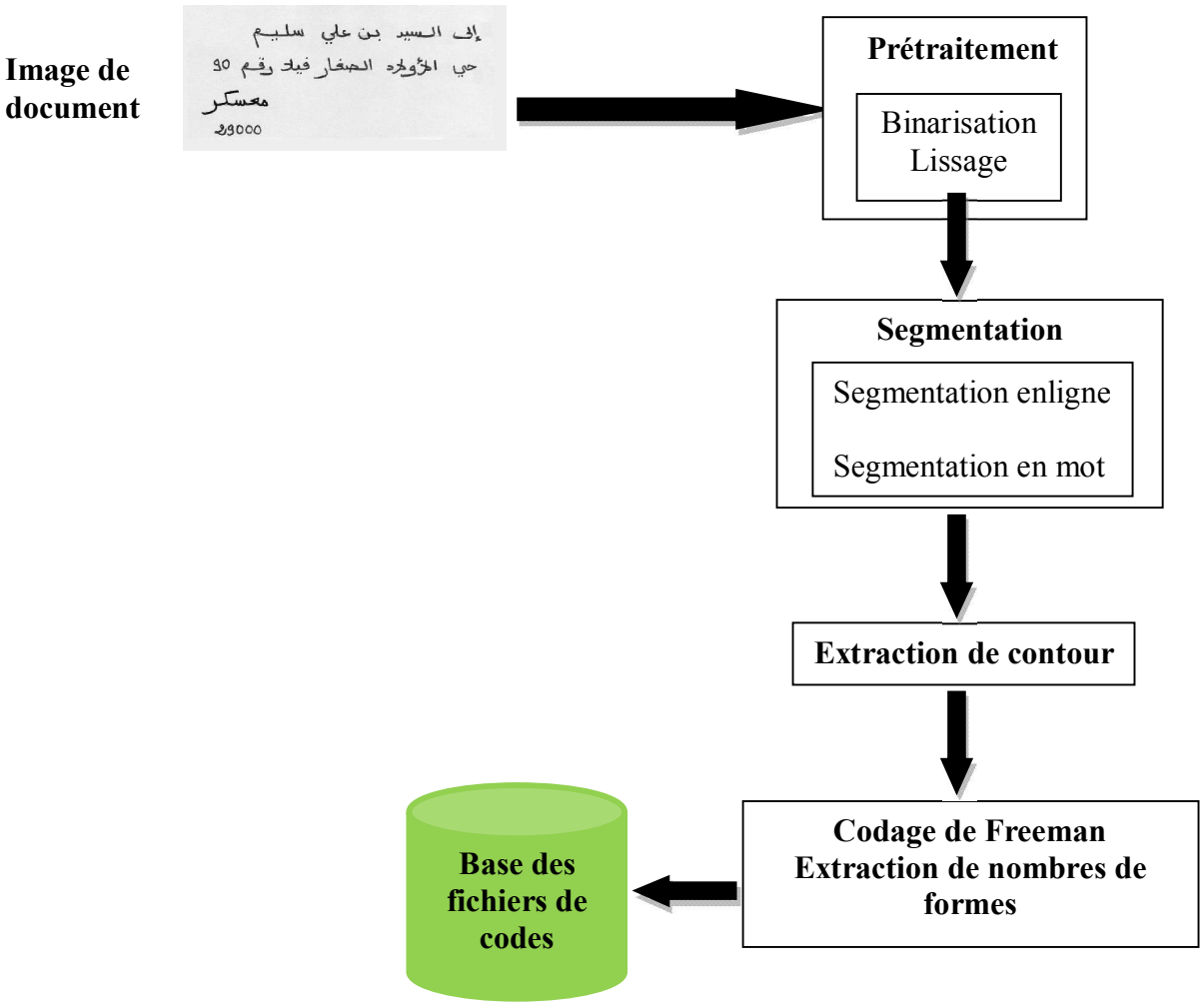


Figure 3.8: *schéma générale de la phase d'analyse*

3.2.1.1. Prétraitement :

Le prétraitement inclut toutes les fonctions effectuées avant de commencer le traitement pour produire une version ≪ nettoyée ≫ de l'image d'origine afin qu'elle puisse être utilisée directement et efficacement. Ainsi dans notre système le prétraitement comprend la binarisation et le lissage. Comme nous avons scanné nos images en niveau de gris, on saute la première étape qu'est la transformation de l'image on niveau de gris.

3.2.1.1.1. Binarisation :

La binarisation consiste à convertir l'image numérisée en une image binaire c'est-à-dire qu'elle sera représentée que par des pixels noirs et des pixels blancs, plus simple à traiter. Elle consiste en une comparaison du niveau de gris des pixels composant l'image avec un seuil, et le problème réside dans le choix de ce seuil puisque la qualité des résultats en recherche de mot dépend de la qualité de la binarisation. Nous avons effectué une comparaison entre deux méthodes de seuillage à savoir : le seuillage global fixe, et la méthode locale de Nick.

-Pour le seuillage global fixe, c'est la technique de binarisation la plus simple, elle consiste à comparer le niveau de gris de chaque pixel x_i de l'image avec un seuil global fixe t (par exemple 127) On note bi la nouvelle valeur du pixel, le seuillage est donné par l'expression suivante : $bi = 255$ si $xi \geq T$ et $bi = 0$ si $xi < T$.

-Pour la méthode locale de Nick, le calcul du seuil est réalisé comme suit :

$$\mathbf{T=m+k\sqrt{\sum (pi^2-m^2)/NPT}}$$

Avec :

k = - 0.2

pi = niveau de gris du pixel i

m = moyenne des valeurs de gris

NPT = nombre de pixels

Un seuil local est recherché pour de petites fenêtres (19 x 19) dans l'image. Tout pixel ayant un niveau de gris inférieur à ce seuil sera considéré comme un point noir et sera représenté par la valeur « 0 », les autres pixels ayant une valeur supérieure au seuil seront considérés comme des pixels blancs. Le pseudo code de l'algorithme utilisé est le suivant:

Algorithme de Binarisation
Entrée : Image I en niveaux de gris

Sortie : Image I' bimodale (1 et 0)

Début

1. la formule de seuil est :

$T=m+k\sqrt{\sum (pi^2-m^2)/NPT}$

2. **pour** chaque pixel p de I **faire**

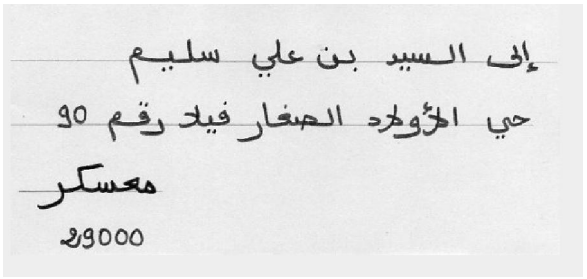
Si $p \leq T$ **alors** p :=0 // rendre le pixel noir

Sinon p :=1 //rendre le pixel blanc

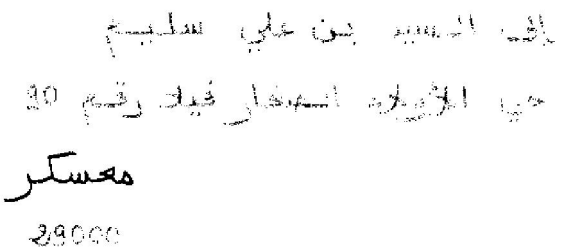
Fin pour

Fin.

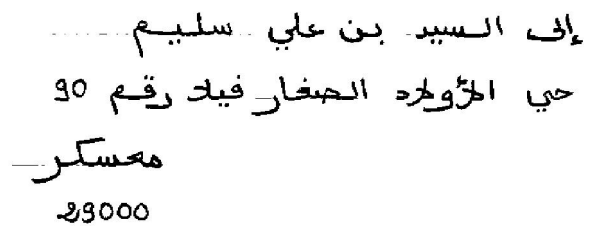
Après avoir testé les deux algorithmes nous constatons que la méthode locale de Nick donne de meilleurs résultats que le seuillage globale fixe comme on le voit dans la (***figure3.9***), c'est pour sa que nous avons choisie de binariser nos images avec la méthode locale de Nick.



(a)



(b)



(c)

Figure3.9 : *résultat de la binarisation, (a) image en niveau de gris, (b) image binarisée par le seuillage globale fixe, (c) image binarisée par la méthode locale de Nick*

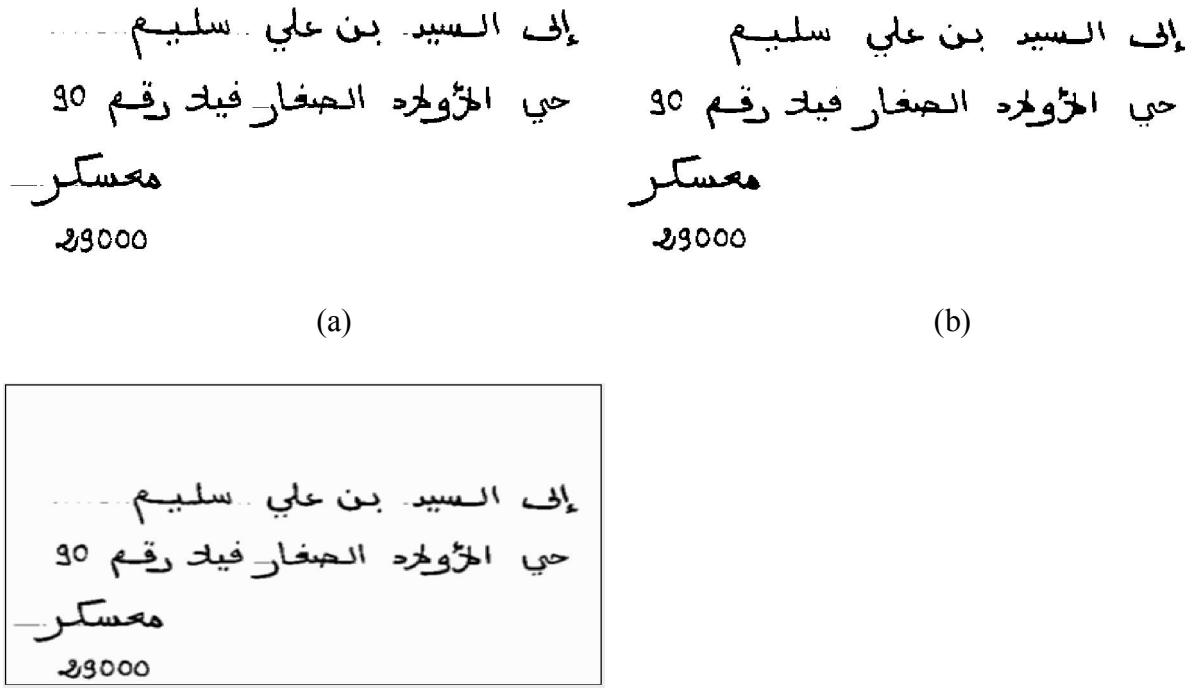
3.2.1.1.2. Lissage :

Le processus d'acquisition ou de binarisation peut introduire des bruits dans l'image, qui se traduisent en particulier par la présence d'irrégularités le long des tracés des caractères ce qui peut dégrader les performances de notre système. Pour pallier à ce problème, nous procédons

à un lissage. Nous avons essayé 2 algorithmes de filtrage linéaire et non linéaire qui sont respectivement le filtre moyenneur et le filtre médian, mais après avoir testé les 2 filtres nous avons opté pour le filtre médian parce que :

Le filtre moyenneur rend l'image moins bruitée mais plus floue et étale les contours (***figure 3.10***). C'est un des défauts des filtres linéaires. Son principe consiste à remplacer un pixel par une moyenne calculée dans son voisinage. La taille du kernel que nous avons utilisé est 3*3

Alors que le filtre médian diminue efficacement le bruit et de plus les caractères gardent leur taille originale donc pas de perte d'informations. Son principe consiste à classer les pixels voisins par ordre croissant des valeurs puis affecter la valeur médiane au pixel courant (central). Nous avons choisi de travailler avec un voisinage 3*3, alors pour ce voisinage on fait le trie des 9 valeurs et on sélectionne la 5ème. La figure 3.10 montre le résultat d'application de cette méthode



(a) (b)

(c)

Figure 3.10 : *résultat du lissage, (a) image binarisée, (b) image lissée avec le médian, (c) image lissée avec le moyenneur.*

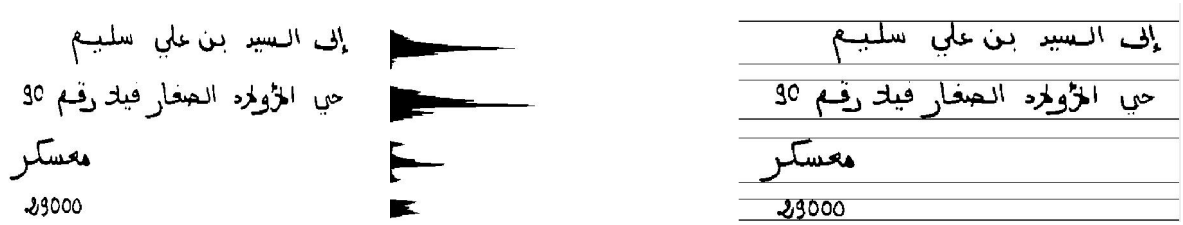
3.2.1.2. Segmentation :

Après le prétraitement, nous devons extraire les mots du texte. Dans ce projet, nous supposons que les documents ne contiennent pas de parties graphiques. L'extraction des zones textuelles

des images sort donc du cadre de notre travail. Pour extraire les mots, nous devons tous d'abord segmenter le texte en lignes, et par la suite, segmenter chaque ligne en sous-mots (composantes connexes).

3.2.1.2.1. Segmentation en lignes :

Pour extraire les lignes du texte, nous procédons à une méthode basée sur la technique de projection horizontale. Cette méthode consiste à faire la somme de tous les pixels noirs sur chaque ligne et de construire l'histogramme correspondant qui sera constitué de pics et de vallées, représentent respectivement les lignes du texte et les espaces entre elles (***figure 3.11***).



(a) (b) (c)

Figure 3.11 : *segmentation en lignes, (a) image prétraitée, (b) son histogramme de projection, (c) image segmentée en lignes*

3.2.1.2.2. Segmentation en mots :

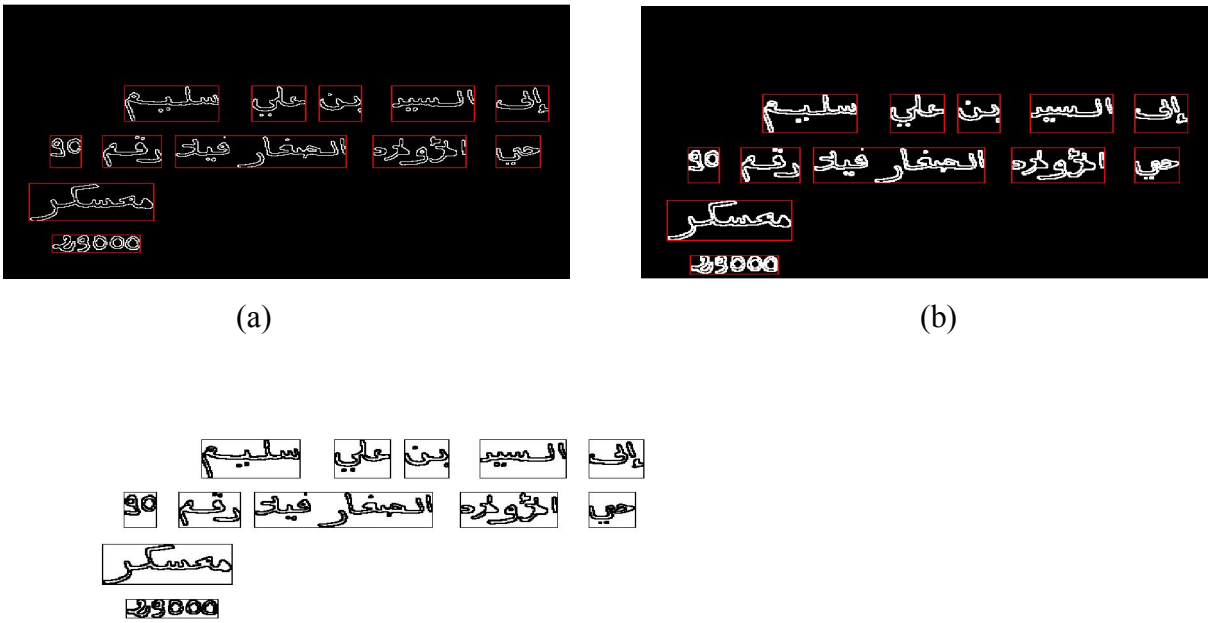
Pour la segmentation en mots d'une image nous avons supposé que les mots sont suffisamment séparer pour que les espaces entre deux mots (espaces inter-mots) soient toujours plus grands que des espaces séparant deux lettres d'un même mot (espaces intra-mots). Sous cette contrainte, la tache de segmentation d'une ligne de texte en mots consiste à estimer un seuil maximum au-delà duquel les espaces entre deux composantes seront considérés comme espaces inter-mots. A la fin de cette étape, chaque mot sera délimité par un rectangle englobant rouge **(*figure 3.12*).**

Figure 3.12 : *segmentation en mots*

3.2.1.3. Suivi du contour et codage

Avant que nous passions au suivi du contour nous détectons en premier le contour de l'image segmentée avec un filtre. Après comparaison des deux filtres de Sobel et Roberts nous avons opté pour le filtre de Roberts, Il s'agit d'un des opérateurs les plus simples qui donne toutefois des résultats corrects. Pour faire simple, l'opérateur calcule le gradient de l'intensité de chaque pixel. Ceci indique la direction de la plus forte variation du clair au sombre, ainsi que le taux de changement dans cette direction. On connaît alors les points de changement soudain de luminosité, correspondant probablement à des bords, ainsi que l'orientation de ces bords.

Après la détection du contour nous effectuons un autre algorithme afin d'obtenir l'image négative de l'image traitée c.à.d. conversion des pixels noir en blanc et vis versa **(*Figure 3.13*).**



(a) (b)

(c)

Figure 3.13 : *Détection du contour (Roberts) (a), Détection du contour (Sobel) (b), Image négative(c)*

Le suivi de contour est communément utilisé pour l'extraction de caractéristiques structurelles des caractères. La chaîne de Freeman est la méthode la plus utilisée de description des contours dans les images. C'est une technique de représentation des directions de contour (on code la direction le long du contour dans un repère absolu lors du parcours du contour à partir d'une origine donnée). Les directions peuvent se présenter en 4 connexités ou en 8 connexités. Les codes des contours sont donnés par la ***figure 3.14*** suivante.

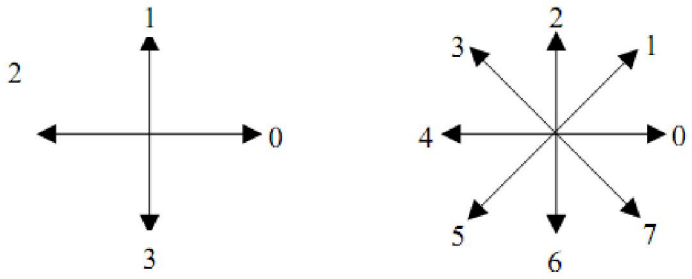


Figure 3.14: *Le code de Freeman en 4-connexités (à gauche) et en 8 connexités (à droite)*

Chaque contour est codé en spécifiant un point de départ suivi par une chaîne ou une séquence de codes de la chaîne de Freeman. Suivant ce codage, chaque image est représentée sous forme d'une liste de contours.

Une des caractéristiques extraites lors de l'implémentation du code de Freeman est le nombre de formes de l'image qui va nous aider à minimiser le taux de comparaison dans la recherche.

Afin d'expliquer notre technique d'extraction du nombre de formes après implémentation du code de Freeman sur une image de tests en vous donne cet exemple illustratif (***figure 3.15)***.

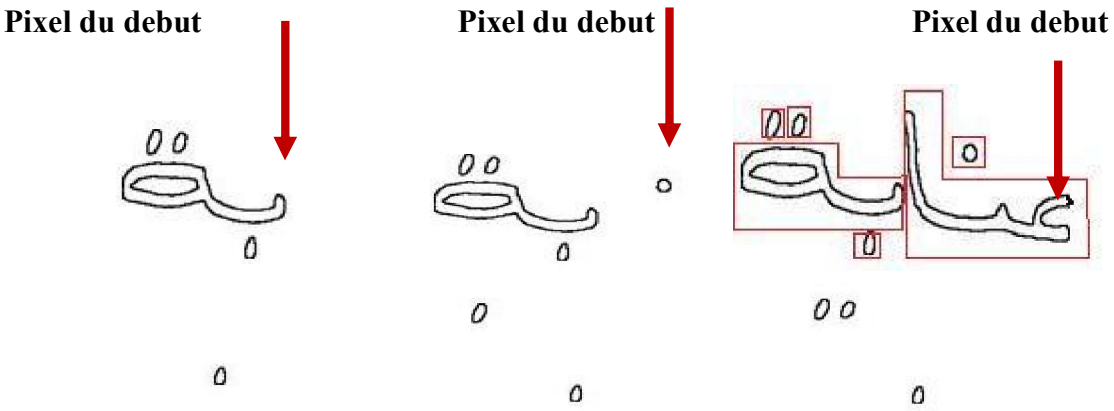


Figure 3.15: *exemple illustratif du calcul du nombre de formes*
(Les six formes sont englobées en rouge)

En applique le codage de Freeman sur le premier pixel de la première forme après le suivi du contour et le parcoure en retourne au pixel initial et en enregistre le code de la forme en passant après à la forme suivante enfin en concatène les chaines de caractères et en aura à la fin la chaine complète de 6 séparateurs c.à.d. 6 formes qui va limiter la recherche après.

Exemple :

5322323323333455656665654433321222222233454545454455675700.7544444544432323323355666766654322223345666654322223355 . . . .

Lors de l'enregistrement en enregistre les images traitées et codées en donnant la main à l'utilisateur de saisir leurs noms en latin dans la base.

3.2.2. La deuxième phase : recherche du mot en dans les images

Le système répond à la requête en retournant un ensemble d'images jugées pertinentes pour l'utilisateur. En utilisant un algorithme de recherche approximative. La phase de recherche est illustrée dans la figure suivante.

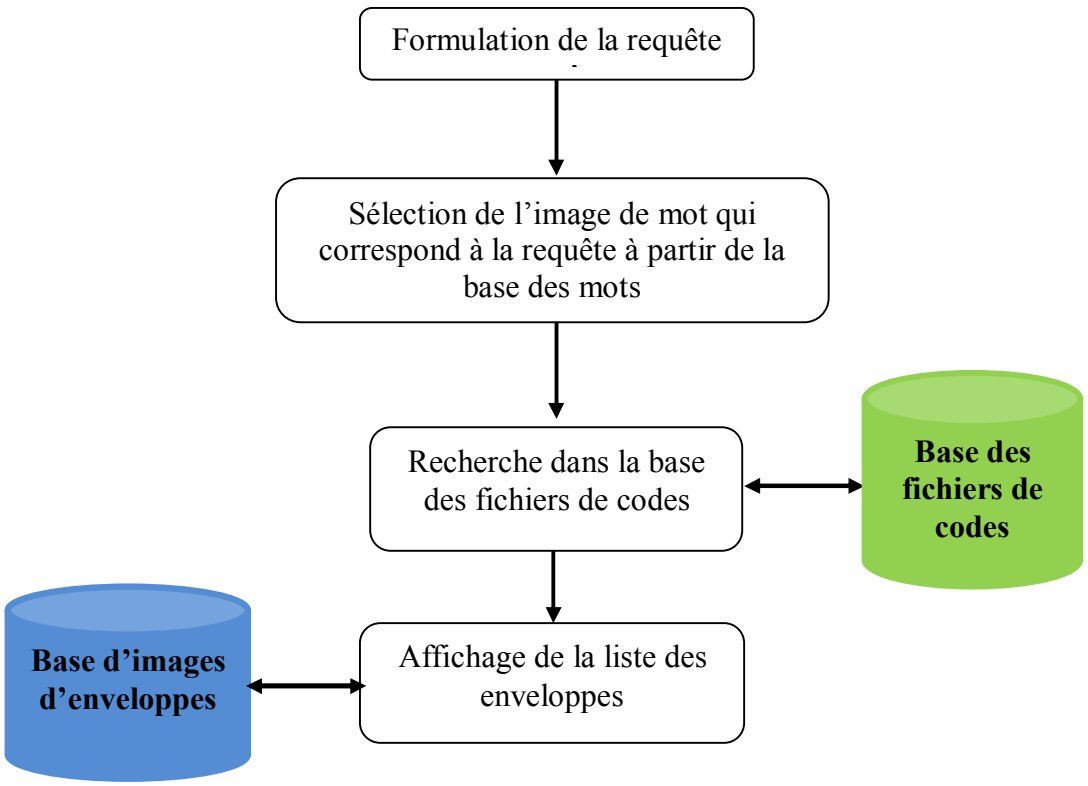


Figure 3.16 : *Schéma général de la phase de recherche*

3.2.2.1. Algorithme de recherche

Chaque document ainsi que la requête est décrit dans le même espace de caractéristiques : une mesure de similarité entre chaque document et la requête est nécessaire afin d'ordonner les documents selon leur pertinence, nous avons utilisé la distance de Levenshtien.

La distance de Levenshtien mesure la similarité entre deux chaînes de caractères. Elle est égale au nombre minimal de caractères qu'il faut supprimer, insérer ou remplacer pour passer d'une chaîne à l'autre. Prenons deux cas :

1) m1 = "maison" et m2 = "maison"
2) m1 = "maison" et m2 = "meison"

Dans le premier cas, d (m1, m2) = 0, car m1 et m2 sont strictement identiques : aucune opération n'est réalisée pour passer d'une chaîne à l'autre. Dans le second, la distance est non nulle, puisque les termes sont différents. J'ai substitué un 'e' au 'a', une seule opération à été effectuée, donc un coût. D'où d (m1, m2) = 1.

Cette distance est d'autant plus importante que le nombre de différences entre les deux chaînes est grand.

L'algorithme de Levenshtein est un algorithme de programmation dynamique (solution de type du bas en haut), qui utilise une matrice de dimension où n et m sont les dimensions (n+1) * (m+1) des deux chaînes de caractères. Dans le pseudo code suivant, la chaîne chaine1 est de longueur longueurChaine1 et chaine2, de longueur longueurChaine2. Cet algorithme renvoie un entier positif ou nul. Il renvoie 0 si les chaînes 1 et 2 sont égales. Si les chaînes 1 et 2 sont très différentes, la fonction renverra au maximum la plus grande longueur des deux chaînes**.**

Pseudo code de Levenshtein

```
entier DistanceDeLevenshtein (caractere chaine1 [1..longueurChaine1], caractère chaine2
  [1..longueurChaine2])

entier d [0..longueurChaine1, 0..longueurChaine2]

entier i, j, coût          // i et j itèrent sur chaine1 et chaine2
Pour ( i = 0 ) jusqu'à longueurChaine1 faire
      d [i, 0] := i
Pour ( j = 0 ) jusqu'à longueurChaine2 faire
      d [0, j] := j
Pour ( i = 1) jusqu'à longueurChaine1 faire
      Pour ( j = 1) jusqu'à longueurChaine2 faire
      Si chaine1 [i] = chaine2 [j] alors coût := 0
```

```
        Sinon coût := 1
d [i, j] := minimum (d [i-1, j] + 1,     // effacement
                     d [i, j-1] + 1,     // insertion
                     d [i-1, j-1] + coût   // substitution     )
            Retourner d [longueurChaine1, longueurChaine2]
```

L'invariant est qu'on peut transformer le segment initial chaine1 [1..i] en chaine2 [1..j] en utilisant un nombre minimal de d [i, j] opérations. L'algorithme achevé, la solution est contenue dans la dernière position à droite de la rangée du bas de la matrice.

3.2.2.2. La saisie du mot de test

L'utilisateur peut saisir un mot ou une lettre alphabétique en latin dans la zone de saisie en cliquant sur le bouton une liste de mots correspondant à la requête s'affichera dans la partie droite de l'interface. Après il choisi l'une des images de mots afficher dans la partie de droite, il la sélectionne ensuite la recherche se déclenche en affichant les résultats dans la partie gauche

3.3. Conclusion :

Dans ce chapitre, nous avons présenté un système permettant la recherche des occurrences d'un mot dans les images de documents arabes sans recourir à une reconnaissance du contenu afin d'éviter le coût élevé et l'effort ardu de l'OCR. L'idée de base de notre système consiste à transposer le problème de recherche d'image de documents du domaine de l'analyse de documents au domaine de la recherche d'information. Le système proposé opère en deux phases : une phase de traitement et d'analyse qui s'effectue hors ligne, dans laquelle chaque document de notre collection est représenté par un code. Le code nous permettra d'employer facilement une technique plus aboutissante de recherche d'information à savoir la technique de recherche approximative. On s'intéresse principalement dans ce projet aux composantes textuelles des documents Arabes. L'extraction des zones textuelles des images sort du cadre du travail demandé. Les discussions sur l'application ainsi qu'au fonctionnement seront l'objet du prochain chapitre.

Chapitre 4

Implémentation et résultats

Le but de ce chapitre est de présenter les différentes interfaces et fonctionnalités de notre système ainsi que la discussion à propos des résultats obtenus, organisé de la manière suivante : dans la section 1 nous illustrons l'environnement expérimental de notre système et le contexte des expérimentations, ensuite dans la section 2 nous présentons la base de données utilisées pour évaluer toutes les étapes de notre système. Dans la section 3, nous décrivons les différentes fenêtres de l'application ainsi que les options. Enfin dans la section 4 nous donnons des exemples de résultats obtenus des expérimentations. La conclusion sera dans la dernière section.

4.1. Environnement d'expérimentation :

Notre système a été développé en langage de programmation Java, avec l'environnement de développement Eclipse Platform (voir ***figure 4.1***) qui est un environnement de développement intégré libre basé sur le concept de programmation orientée objet. C'est un environnement qui permet aux développeurs, même non expérimentés, de créer facilement des interfaces homme/machine d'aspect Windows, Web…etc.

Le choix du langage JAVA est motivé principalement par sa portabilité et son ouverture sur Internet, en plus de l'existence de plusieurs bibliothèques en java qui réalisent des traitements dont on a besoins.

Tous les tests effectués sont exécutés sur un PC disposant d'un processeur Intel core2 duo cadencé à 4.0 GHz, de 3 Go de mémoire centrale et d'un disque local de 250 Go.

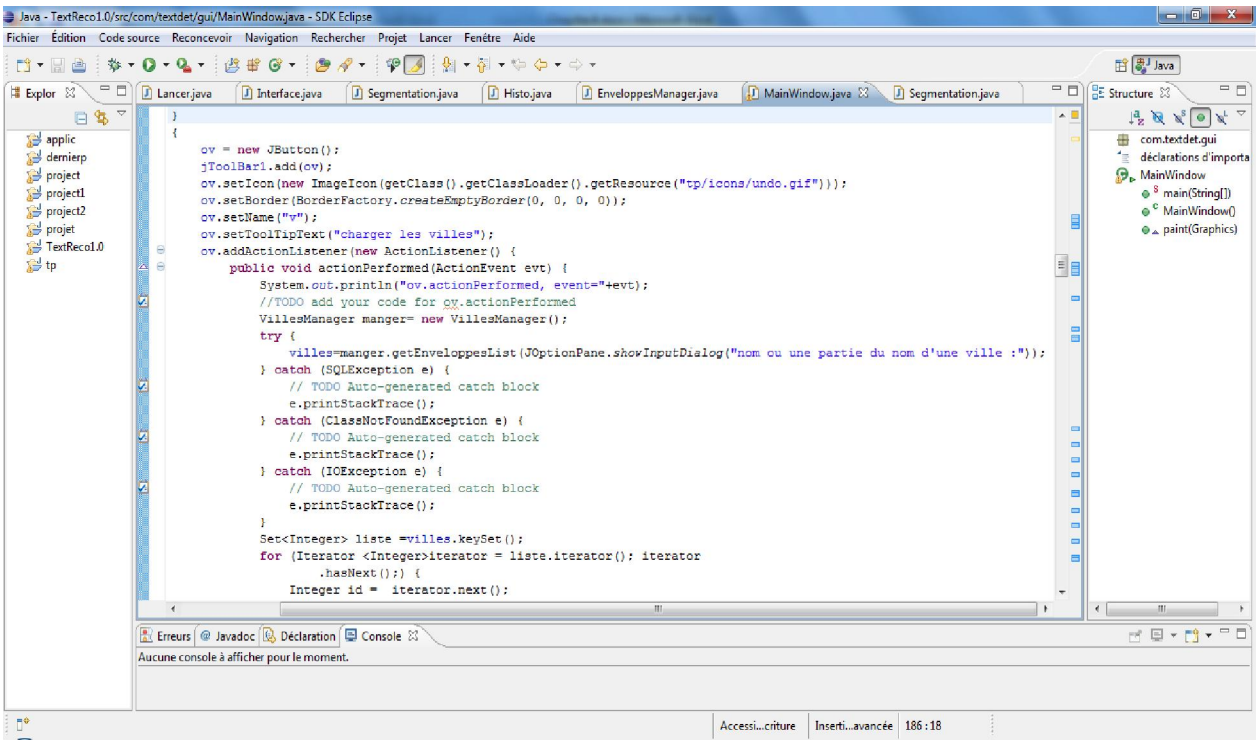


Figure 4.1 : *Interface de l'environnement de développement Eclipse*

4.2. Base de données :

Pour estimer les performances de notre système (les performances des deux parties : analyse et recherche) nous avons besoin d'une base de données. Vu l'absence d'une base de données standard des documents arabes pour pouvoir l'utiliser dans la validation, une base locale de 100 images d'enveloppes postales algériennes et de mot de test a été construite par nous même. Les images de la base sont en niveau de gris et elles sont sous le format JPEG avec une dimension de 551x380 et scannées dans les mêmes conditions. La ***figure*** suivante montre quelques exemples de documents de cette base :

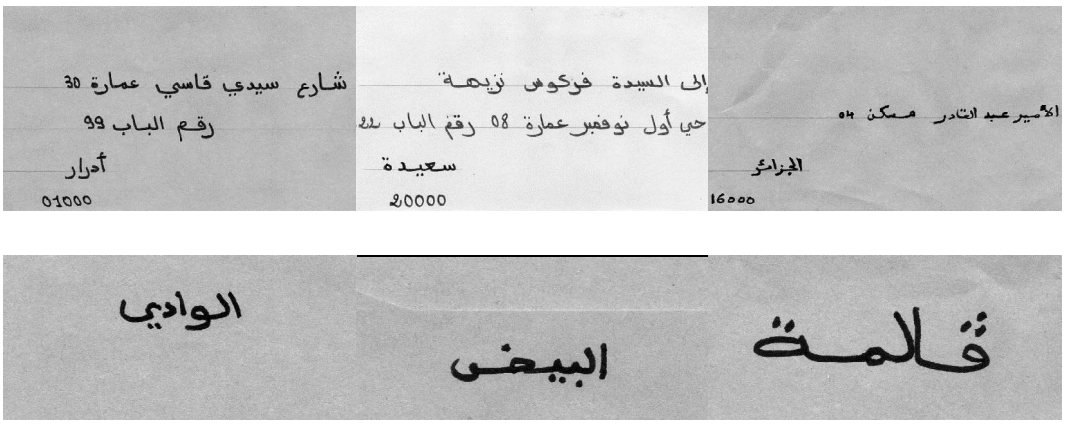


Figure 4.2 : *Exemple de documents de la base (enveloppes et mots de test)*

Notre base de données est en MySQL qui est accessible et utiliser conjointement par le langage java. Afin de visionner la base en utilise MySQL Query Browser qui est un outil graphique fourni par MySQL pour la création, l'exécution et l'optimisation des requêtes dans un environnement graphique qui nous aide à interroger et d'analyser les données stockées dans notre base de données MySQL (***Figure 4.3***).

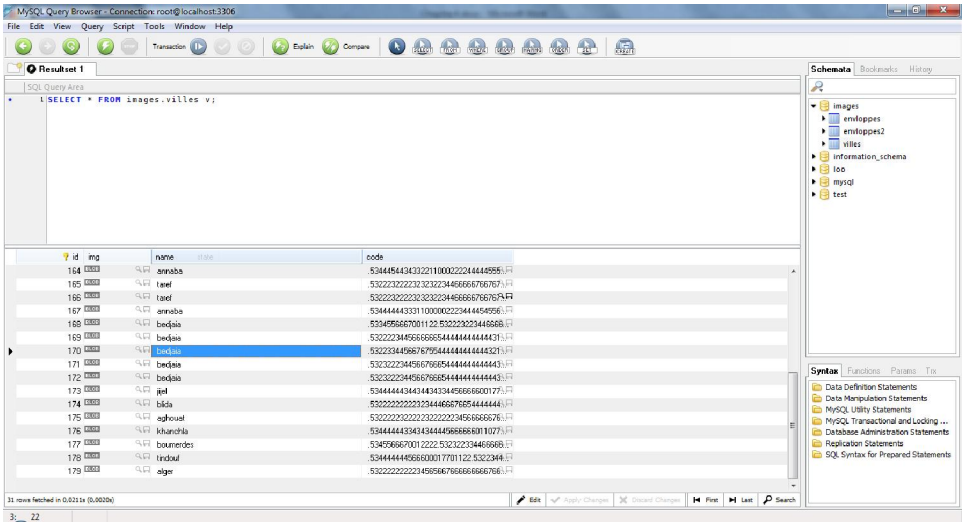


Figure 4.3 : *interface de MySQL Query Browser*

4.3. Description de l'application

4.3.1. Vue générale de l'interface de l'application

Lors du lancement de l'application la fenêtre principale s'affiche (***Figure.4.4***) contenant les menus et les raccourcis facilitant ainsi l'accès aux fonctions de l'application. Notre interface est devisée en trois parties, la partie gauche pour afficher les images des enveloppes non traités que contient la base lors du chargement ainsi qu'aux enveloppes traitées quand t'on fait la recherche, la partie gauche pour l'affichage des mots de tests lors de la recherche, la partie central pour afficher les images des enveloppes en taille réel afin de voir les traitements et l'enregistrement, et enfin une petite fenêtre pour visionner le code de Freeman générer lors de l'enregistrement.

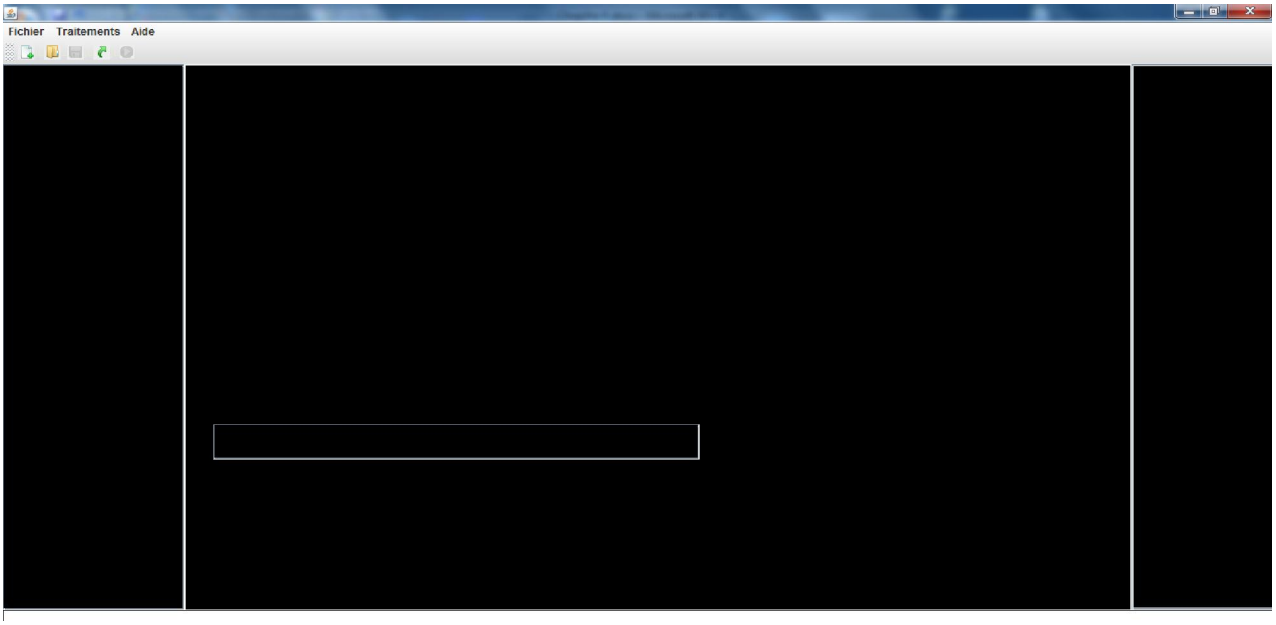


Figure 4.4 : *Interface principale de l'application*

4.3.2. Menu de l'application et prétraitement

Le menu de l'interface principale donne accès aux fonctionnalités du logiciel

4.3.2.1. Menu Fichier

Le menu Fichier contient les fonctions du chargement, d'ajout et d'enregistrement des images traitées ainsi la possibilité de quitter l'application.

En peut aussi utiliser le raccourci  pour l'ajout de nouvelles images et le raccourci  pour le chargement des images pour le traitement

4.3.2.2. Menu Traitement

Le menu Traitement contient les fonctions de prétraitement c.à.d. binarisation par Nick, lissage, détection de ligne, segmentation en mots, détection de contour et l'image négative. Notant que chaque traitement s'active un par un lors du traitement sur une image.

En peut lancer le traitement aussi en cliquant sur le raccourci qui lance les traitements successivement à chaque clique jusqu'à ce qu'il soit désactiver Pour l'enregistrement d'une image traité en clique sur le raccourci qui s'active après la désactivation du raccourci traitement, une fenêtre avec un champ de saisi s'affiche à l'écran afin de donner un nom à l'enveloppe ainsi qu'au code de Freeman généré.

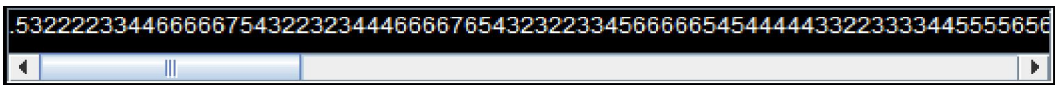


Figure 4.5 : *enregistrement de l'image traité avec le code de Freeman*

4.3.3. La recherche

La méthode de recherche de notre application est très simple, en cliquant sur le raccourci une fenêtre avec un champ de saisi s'affiche aussi pour donner le nom de la ville à chercher en latin dans les images des enveloppes préenregistrer dans la base ***Figure 4.6*** **(a)** et lors de l'affichage des mots de test à droite dans l'interface, un clique directe sur l'un des mots affiché engendre la recherche à partir de la base en comparant son code de Freeman avec les codes des enveloppes préenregistrer dans la base. Affichant ainsi les images des enveloppes à la partie gauche qu'on peut les visionner au milieu de l'interface **Figure 4.6 (b)**

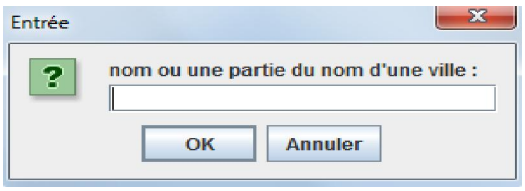


Figure 4.6 (a) : *Champ de saisi du mot à chercher*

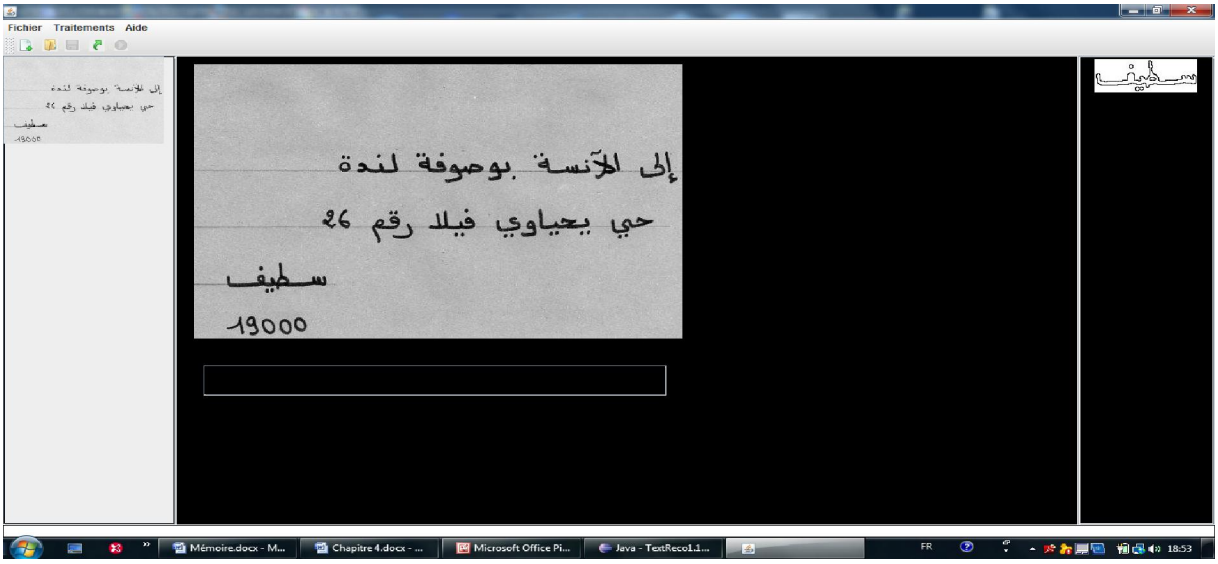


Figure 4.6 (b) : *Exemple de recherche du mot de la Wilaya de (Stif)*

4.4. Résultats et discussions :

Dans cette section, nous exposons les différentes expérimentations effectuées, les résultats obtenus de ces expérimentations, et nous discutons ces résultats.

Les expérimentations adressées dans cette section concernent à la fois la partie d'analyse de documents et la partie de recherche pour pouvoir les évaluer séparément et pour montrer les performances de notre système de recherche sur la base de données présentée précédemment.

4.4.1. Partie d'analyse :

Pour la partie d'analyse, nous avons effectués plusieurs expérimentations afin de tester les performances de chaque étape de cette partie, et pour choisir la meilleure méthode à appliquer dans chaque étape, en se basant principalement sur des critères visuelles, à cause de l'absence des critères standards d'évaluations.

4.4.1.1. Binarisation :

Il existe plusieurs méthodes d'évaluation de la binarisation, et pour choisir la meilleure méthode à utiliser dans notre système, nous avons effectué une comparaison entre une méthode de binarisation globale et l'autre locale qui sont respectivement : le seuillage global fixe, et la méthode locale de Nick.

Nous avons choisi la méthode de Nick qui donne visuellement de meilleurs résultats, comme nous l'avons montré dans le chapitre précédent.

4.4.1.2. Segmentation en lignes et en mots :

La segmentation en lignes est utilisée sur des images de documents non inclinés, pour éviter l'étape de la correction de l'inclinaison puisque nous avons utilisé la méthode de la projection horizontale qui est sensible à l'inclinaison. Les résultats obtenus sont très satisfaisons comme nous l'avons vu dans le chapitre précédent.

Pour la segmentation en mot la méthode que nous avons utilisée consiste à estimer un seuil maximum au-delà duquel les espaces entre deux composantes seront considérés comme espaces inter-mots est qui donne des résultats satisfaisons.

4.4.1.3. Extraction de contour :

Pour l'extraction de contour nous avons testé deux algorithmes, l'algorithme de Sobel et l'algorithme de Roberts.

Nous avons choisi le filtre de Roberts car il donne un contour plus fin que le filtre de Sobel qui donne l'avantage de parcourir moins de pixels noir lors de la génération du code de Freeman et qui permet bien sûre de donner un code moins grand.

4.4.1.4. Code de Freeman et calcul de nombre de formes :

L'utilisation du code de Freeman avec le calcul du nombre de formes permet de caractériser les documents et les sauvegarder ainsi pour facilité la recherche comme nous l'avons expliqué dans le chapitre précédent.

4.4.2. Partie de recherche :

4.4.2.1. Rappel et précision :

On mesure l'efficacité d'une technique de recherche d'informations en utilisant deux mesures distinctes :

Rappel : Le rappel est défini par le nombre de documents pertinents retrouvés au regard du nombre de documents pertinents que possède la base de données. Si cette adéquation entre le questionnement de l'utilisateur et le nombre de documents présentés est importante alors le taux de rappel est élevé.

Le Rappel est calculé à partir du nombre de document Pertinent retourné attribuées a une requête diviser par le nombre totale de documents pertinent a la même requête appartenant au fond documentaire.

Précision : La précision est le nombre de documents pertinents retrouvés rapporté au nombre de documents total proposé par le système de recherche pour une requête donnée.

Le principe est le suivant: quand un utilisateur interroge une base de données, il souhaite que les documents proposés en réponse à son interrogation correspondent à son attente. Tous les documents retournés superflus ou non pertinents constituent du bruit. La précision s'oppose à ce bruit documentaire. Si elle est élevée, cela signifie que peu de documents inutiles sont proposés par le système et que ce dernier peut être considéré comme "précis".

La précision est calculée a partir du nombre de document pertinent retournés attribués a la requête divisé par le nombre totale de documents retournés par le système.

Nous allons donner maintenant des exemples de résultats obtenus qui nous permettent de justifier la précision du système de recherche proposé lord du test sur 6 noms de wilayas.

Le tableau suivant montre les résultats obtenus.

Critères / ***Nom de la ville***	***Nombre d'apparition (Nombre apparut/ Nombre d'images enregistrer)***	***Précision***	***Rappel***
Taraf	5/5	1	1
Annaba	3/3	1	1
Bejaïa	4/5	1	0.8
Guelma	2/3	1	0.7
Constantine	2/2	1	1
Mostaganem	1/1	1	1

A partir du tableau précédent, nous constatons que les résultats obtenus du système proposé montrent une précision moyenne de 100% de la recherche et ca avec l'algorithme de Levenshtien. Ceci revient principalement à la bonne qualité des enveloppes et la taille réduite

de la base en plus en a utilisé un seul mot de test qui est le nom de la Wilaya dans la recherche, parce que c'est l'information la plus évidente dans les enveloppes.

Les villes comme Bejaïa et Guelma donnent des résultats moins bons on terme de rappel, n'oublions pas surtout les facteurs de dégradation du taux de précision tel que l'acquisition de ces images en premier lieu, les résultats des prétraitements de ces images et la génération du code sur celles-ci.

On remarque que le système présente un taux de rappel moyen de 0.8. En effet quelque documents pertinents en été ignoré lors de le recherche parce que la recherche du mot s'effectue sur la base des fichiers de codes et non pas sur une base d'indexes.

L'utilisation d'une base d'indexe permet d'augmenter ce taux parce que il est connue que les indexes ne peuvent pas contenir tout les mots.

4.5. Conclusion :

Dans ce chapitre, nous avons présenté notre système de recherche de mots arabes dans les documents manuscrit

Le système a été testé sur une base contenant 100 images d'enveloppes postales algériennes et d'images de test. La base nous l'avons construit en l'absence d'une base standard

Les résultats de la recherche obtenus sur la base des images documents, sont à notre avis très satisfaisants en précision et pour le temps de réponse, la recherche des documents comme les résultats obtenus montrent la faisabilité et robustesse de la démarche employée.

Conclusion générale et perspective

Le travail adressé dans ce mémoire s'inscrit dans le cadre du traitement automatique du document et leur indexation. Nous nous sommes intéressés dans ce mémoire par les images d'enveloppes postales algériennes. Notre contribution concerne essentiellement le développement d'un système de traitement et de recherche de mots dans les images de documents manuscrits arabes sans recourir à une reconnaissance du contenu afin d'éviter le cout élevé et l'effort ardu de la reconnaissance.

En effet, la recherche d'images de documents par le contenu (RIDC) nécessite une analyse détaillée de la structure physique et logique conjointement avec une description des objets figurent sur le document. Nous nous sommes focalisés dans cette première ébauche sur les parties textuelles des documents arabes manuscrits. Une variabilité et diversité des formes des mots et caractères sont observés rendant l'automatisation de la tâche très ardue. Une recherche et indexation par le contenu s'avèrent également un défi majeur pour les recherches actuelles dans le domaine de la RIDC.

Le système développé est composé de deux parties essentielles, la première partie regroupe plusieurs techniques tirées principalement du domaine de l'analyse de documents : prétraitements, segmentation, extraction de contours et codage, ayant comme but de représenter chaque document de notre collection par un ensemble de codes. Ces codes nous permettrons d'employer facilement les techniques les plus aboutissantes de recherche d'information à savoir la techniques de recherche approximative

La deuxième partie de notre système qui est la partie de recherche est représentée par une interface, permettant à l'utilisateur de chercher les documents qu'il veut, en formulant une requête textuelle en langue latine. Le système répond à cette requête en retournant un ensemble de documents jugés pertinents pour l'utilisateur. Pour ce faire, nous procédons en une technique issue du domaine de la recherche d'informations. Ensuite, nous procédons à une recherche dans les fichiers de codes correspondants aux documents. Nous avons proposé d'employer une recherche approximative, parce que les fichiers de codes ne sont pas précis et dans certains cas ils sont incomplets, à cause de plusieurs facteurs : non discrimination des

caractéristiques employées, inefficacité des algorithmes de traitements utilisées sur les images de mauvaises qualité…etc. Nous avons utilisé une mesure de distance entre les chaines : la distance de Levenshtein.

Afin d'évaluer les performances de notre système, nous avons effectué plusieurs expérimentations et tests visant à évaluer chaque étape des deux parties de notre systèmes séparément. Les tests sont réalisés sur une base que nous avons crée à cause de l'absence d'une base de données standard d'enveloppes postales algériennes, et elle est composée de100 images d'enveloppes écrit par 3 scripteurs.

Comme tous les travaux de recherche, plusieurs extensions sont envisageables pour améliorer le système proposé dans ce mémoire :

- Tester l'approche sur une base de données réelle et très large d'images de documents arabes anciens contenants différents types de dégradations, complexité de structure et variations d'écriture.
- Elargir le système pour qu'il regroupe d'autres étapes de traitements et utiliser d'autres techniques plus performantes de segmentation et d'ajouter une technique de correction de l'inclinaison.
- Utilisation de documents qui contiennent des parties graphiques et textuelles et l' ajouter d'un autre module permettant la séparation entre le texte et le graphique, et l'extraction des caractéristiques de ce dernier, pour permettre un recherche mixte texte-graphique.
- Utiliser les caractéristiques structurelles de l'écriture, par exemple les boucles, les jambes, les hampes et la séparation entre les points diacritiques.
-
- Tester d'autres algorithmes de recherche qui peuvent apparaître performants et les combiner avec les mesures utilisées ici.

Bibliographies

[ANT 97] ANTONACOPOULOS A., «Local Skew Angle Estimation from Background Space in Text Regions», Proceedings of the 4th International Conference on Document Analysis and Recognition, Germany, August 18–20, 1997, vol. 2, pp. 684–688.

[Al- 95] B. Al-Badr, S.A. Mahmoud, «Survey and bibliography of Arabic optical text recognition». Signal processing, vol. 41, pp. 49-77, 1995.

[ABD 06] Abdulkader A., «Two tier approach for Arabic offline handwriting recognition», The Proceedings of IWFHR'06, 10th International Workshop on Frontiers in Handwriting Recognition, La Baule, France, pp. 161-166, October 2006.

[ALM 02a] Al-Ma'adeed S., Higgins C., Elliman D., «A database for Arabic handwritten text recognition research», Proceedings of IWFHR'02, 8th International Workshop on Frontiers in Handwriting Recognition, pp. 485-489, Ontario, Canada, August 2002.

[ALM 02b] Al-Ma'adeed S., Higgins C., Elliman D., «Recognition of off-line handwritten Arabic words using hidden Markov model approach», Proceedings of ICPR'02, 16th International Conference on Pattern Recognition, Vol. 3, pp. 481-484, Quebec City, Canada, August 2002.

[ALM 04] Al-Ma'adeed S., Elliman D., Higgins C., «Off-line recognition of handwritten Arabic words using multiple hidden Markov models», Knowledge-Based Systems, Vol. 17, N°. 2-4, pp. 75-79, May 2004

[ALM 06] Al-Ma'adeed S., « Recognition of off-line handwritten Arabic words using neural network», Proceeding of GMAI'06, International Conference on Geometric Modeling and Imaging, pp. 141-114, London, England, July 2006.

[AMI 96] AMIN A., FISCHER S., PARKINSON T., SHIU R., "Fast algorithm for skew detection", SPIE Proceedings, Vol. 2661, 29-30 Janvier 1996.

[AND 99] J. André, M.A. Chabin. «Numériser les documents anciens : et après ? », 1999, Les documents anciens, numéro spécial de Document Numérique, vol. 3, no1-2, pp. 7-11.

[ARI 01] Arica, N. et Yarman-Vural, F. T. (2001). « An overview of character recognition focused on off-line handwriting ». IEEE transactions on systems, man and cybernetics - part C: Applications and reviews, 31(2): 216-233.

[AZO 95] Antoine Sourou AZOKLY, «Une approche uniforme pour la reconnaissance de la structure physique de documents composites fondée sur l'analyse des espaces», Thèse de doctorat, Institut d'Informatique Université de Fribourg Suisse, 1995.

[BAC 98] Bachimont B. « Bibliothèques numériques audiovisuelles. Des enjeux scientifiques et techniques ». Revue Document numérique, 2:3-43-4, 2 19-242, Hermès, 1998.

[BAG 97] BAGDANOV A., KANAI J., "Projection Profile Based Skew Estimation Algorithm for JBIG Compressed Images", In: Proceedings of the 4th International Conference on Document Analysis and Recognition, Germany, 1997, p. 401-405.

[BAI 91] H.S. Baird. H. Bunke. K. Yamamoto. « Structured document image analysis », Springer Verlag. Berlin/ Heidelberg, 1991..582p.

[BAP 98] Frédéric BAPST, « Reconnaissance de documents assistée: architecture logicielle et intégration de savoir-faire », Thèse de doctorat, Université de Fribourg (Suisse), 1998

[BEL 92] Belaid A., Belaid Y., « Reconnaissance des formes : méthodes et application », InterEdition 1992.

[BEL 01] Abdel Belaïd, « Reconnaissance automatique de l'écriture et du document », Pour la science, 2001. 22 p. (Article dans une revue de vulgarisation.)

[BEL 02] A. Belaid, «Analyse et reconnaissance de documents, Cours INRIA: le Traitement électronique de Documents», Collection ADBS, 3-7 octobre, Aix-en-Provence, 2002.

[BEL 06] Abdel Belaıd and Christophe Choisy. «Human reading based strategies for off-line arabic word recognition». Summit on Arabic and Chinese Handwriting Recognition 2006 - SACH'06, 2006.

[BEN 99] A.BENNASRI, A.ZAHOUR, B. TACONET, « Extraction des lignes d'un texte manuscrit arabe », CIFED, 2002.

[BEN 02] N. Benahmed, «Optimisation de Réseaux de Neurones Pour la Reconnaissance des Chiffres Manuscrits Isolés, Sélection et Pondération des Primitives par Algorithmes Génétiques», Thèse pour l'obtention de la Maîtrise en Génie de la Production Automatisée, Montréal, Mars 2002.

[BEN 08] Abdallah BENOUARETH, «Reconnaissance de Mots Arabes Manuscrits par Modèles de Markov Cachés à Durée d'Etat Explicite», Thèse de Doctorat d'Etat, Labo. LRI, Département d'informatique, Université d'Annaba, Algérie, 2008

[BER 98] BERGLER S., KHOURY S., SUEN B. C. Y., WAKED B., "Skew Detection, Page Segmentation and Script Classification of Printed Document Images", IEEE International Conference on Systems, Man, and Cybernetics, October 1998, pp. 4470-4475.

[BIM 99] Del Bimbo A, «Visual Information Retrieval». San Francisco : Morgan KaufmannPublishers, 1999.

[BOS 08] Denis BOSSY, « Le traitement d'image dans l'analyse de documents anciens », Séminaire de recherche du groupe DIVA Université de Fribourg, 29 mai 2008

[BOU 06] Riadh BOUSLIMI, «Système de reconnaissance hors-ligne des mots manuscrits arabe pour multi-scripteurs, mémoire de magistère, UNIVERSITE DE JENDOUBA (TUNISIE). 2006

[BUN 97] BUNKE H., WANG P. S. P, "Handbook of character recognition and document analysis", Edition World Scientific Publiching, 1997.

[CHA 06] Clément Chatelain, « Extraction de séquences numériques dans des documents manuscrits quelconques », thèse de doctorat, Université de Rouen, 5 décembre 2006.

[CHE 06] S. Chevalier, M. Lemaître, E. Geoffrois, «Étude de primitives spectrales pour la reconnaissance de caractères manuscrits dans le cadre d'une approche markovienne 2D», Actes 15ème Congrès Francophone AFRIF-AFIA Reconnaissance des Formes et Intelligence Artificielle (RFIA'2006), Tours, France (CDROM), 2006.

[DAN 99] DANCE C., NEWMAN W., TAYLOR A., TAYLOR S., TAYLOR M., ALDHOUS T., "CamWorks: A Videobased Tool for Efficient Capture from Paper Source Documents", IEEE International Conference on Multimedia Computing and Systems, Volume 2, June 07-11, 1999.

[DAR 94] P. Dargenton, «*Contribution à la Segmentation et à la reconnaissance de l'écriture manuscrite*», Thèse de Doctorat, 1994.

[DEF 95] O. DEFORGES, P. PIQUIN, C. VIARD-GAUDIN, D. BARBA, « Segmentation d'images de documents par une approche multirésolution. Extraction précise des lignes de texte », Traitement du Signal 1995 - Volume 12 – n° 6

[DER 09] DERDOUR Khedidja, « Reconnaissance de formes du chiffre arabe imprimé : Application au code à barre d'un produit», mémoire de magistère, Université HADJLAKHDAR –BATNA. 2009

[DEV 82] P.A. Devijver & J. Kittler. « Pattern recognition, a statistical approach ». Englewood Cliffs, London, 1982.

[DOS 00] Philippe DosCR, « Un environnement pour la reconstruction 3D d'édifices à partir de plans d'architecte», Thèse de doctorat, Université Henri Poincaré - Nancy 1, 30 juin 2000.

[DRI 07] Fadoua Drira, « Contribution à la Restauration des Images de Documents Anciens », Thèse de doctorat, L'institut National des Sciences Appliquées de Lyon, 2007.

[DUO 05] Jean DUONG, « Etude des Documents Imprimés : Approche Statistique et Contribution Méthodologique», Thèse de doctorat, L'institut National des Sciences Appliquées de Lyon, 2005.

[ELL 90] D.G. Elliman. « A review of segmentation and contextual analysis techniques for text recognition ». Pattern Recognition. 1990. vol. 23. n° 3. pp. 337-3-16.

[EMP 03] Hubert EMPTOZ, Franck LEBOURGEOIS, Véronique EGLIN, Yann LEYDIER, « La reconnaissance dans les images numérisées : OCR et transcription, reconnaissance des structures fonctionnelles et des méta-données », La numérisation des textes et des images : techniques et réalisations», éditions Presse de Lille 3 France, ISBN 2-84467-050-4, 16-17 janvier 2003, pp. 105-129.

[FAU 04] Julien FAUQUEUR «Contributions pour la Recherche d'Images par Composantes Visuelles», Thèse de doctorat, l'Université de Versailles Saint-Quentin en Yvelines. 2004.

[FEL 01] Feldbach M., Tönnies K.D., « Line detection and segmentation in Historical Church registers », Actes de ICDAR'01, Seattle, septembre 2001, pp. 743-747.

[GUS 99] Gusnard de Ventadert (nom collectif), « Les documents anciens », Document numérique, Hermès, Vol. 3, no 1-2, juin 1999, pp. 57-73.

[HAD 06] Karim HADJAR, « Une étude de l'évolutivité des modèles pour la reconnaissance de documents arabes dans un contexte interactif », Thèse de doctorat, Université de Fribourg (Suisse), 2006.

[HIL 04] Xavier HILAIRE, « Segmentation robuste de courbes discrètes 2D et applications à la rétroconversion de documents techniques», Thèse de doctorat, Institut National Polytechnique de Lorraine, 2004.

[HUL 98] HULL J. J., TAYLOR S. L., "Document Analysis System II", World Scientific Edition, 1998.

[ING 99] INGLIS S. J., "Lossless Document Image Compression", these de doctorat, Université de Waikato, New Zealand, Mars 1999.

[JAI 96] JAIN A., YU B., "A Robust and Fast Skew Detection Algorithm for Generic Documents". Pattern Recognition, Volume 29 No. 10, October 1996, p. 1599-1629.

[JAI 00] Jain A.K., Duin R.P.W., Mao J., «Statistical pattern recognition: A review», IEEE Transactions on PAMI, Vol. 22, N°. 1, pp. 4-37, January 2000.

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

[JUN 04] Keechul Jung, Kwang In Kim, and Anil K. Jain. «Text information extraction in images and

video: a survey». Pattern Recognition, 37(5):977–997, 2004.

[KAV 02] KAVALLIERATOU E., FAKOTAKIS N., KOKKINAKIS G., "Skew angle estimation for printed and handwritten documents using the Wigner-Ville distribution", Image and Vision Computing, 2002, pp.813–824

[KEF 09] Abderahmane Kefali, Toufik Sari et Mokhtar Sellami, « Implémentation de plusieurs techniques de seuillage d'images de documents arabes anciens », journées Gestion Electronique de Documents & Réseaux de Recherche en sciences et Technologies d'information, GED'09, Université de Annaba, Algérie, les 20-21 Mai 2009.

[KHU 09b] Khurram KHURSHID, Claudie FAURE, Nicole VINCENT, « Recherche de mots dans des images de documents par appariement de caractères », Actes du dixième Colloque International Francophone sur l'Écrit et le Document.

[KOC 06] G. Koch. «Catégorisation automatique de documents manuscrits : application aux courriers entrants». Thèse de doctorat, Université de Rouen, 2006.

[LAR 86] Librairie Larousse, editor. « *Petit Larousse Illustré* ». Larousse Diffusion, Paris (France), 1986.

[LAR 89] Librairie Larousse, editor. « *Larousse de Poche* ». Larousse Diffusion, Paris (France), 1989.

[LE 94] LE D. X., THOMA G., WESCHLE H., "Automated Page Orientation and Skew Angle Detection for Binary Document Images", Pattern Recognition, Volume 27, Number 10, October 1994, pp. 1325-1344.

[LEE 01] Lee, S.W., Ryu, D.S., « Parameter-Free Geometric Document Layout Analysis », IEEE Tran.on PAMI., Vol. 23, No. 11, p1240-1256, 2001.

[LEY 06] Leydier Y., LeBourgeois F., Emptoz H., «Contribution à la création d'un moteur de recherche sémiotique : application aux manuscrits latins médiévaux». CORESA, pp 104-109, Caen, France, 9 et 10 novembre 2006.

[LIK 94] Likforman-Sulem L., Faure C., « Extracting lines on handwritten documents by perceptual grouping », in Advances in Handwiting and drawing : a multidisciplinary approach, C. Faure, P. Keuss, G. Lorette, A. Winter (Eds), pp. 21-38, Europia, Paris, 1994.

[LIK 95] Likforman-Sulem L., Hanimyan A., Faure C. « A Hough Based Algorithm for Extracting Text Lines in Handwritten Documents », Actes de Int. Conf. On Document Analysis and Recognition ICDAR'95, Montréal, pp. 774-777.

[LIK 03] Likforman-Sulem L. « Apport du traitement des images à la numérisation des documents manuscrits anciens », *Document numérique*, Vol. 7, n° 3-4, 2003, pages 13 à 26.

[LEM 07] Mélanie Lemaitre, «Approche markovienne bidimensionnelle d'analyse et de reconnaissance de documents manuscrits». Thèse de doctorat, Université Paris 5 René Descartes, 2007

[MAR 99] U. Marti & H. Bunke. «A Full English Sentence Database for Off-Line Handwriting Recognition». ICDAR, pages 705–708, 1999.

[MAR 01a] U. Marti & H. Bunke. «Text Line Segmentation and Word Recognition in a System for General Writer Independent Handwriting Recognition». ICDAR, page 159, 2001.

[MAR 01b] U.V. Marti & H. Bunke. «Using a statistical Language model to improve the performance of an HMM-based cursive handwriting recognition system». IJPRAI, vol. 15, pages 65–90, 2001.

[MEH 05] Mehennaoui Z., Benouareth A., Sellami M., "Reconnaissance des caractères arabes par Support Vector Machines", Proceedings of Text Image and Speech Recognition Workshop, pp. 80-87, Annaba, Algérie, Décembre 2005.

[MEN 08] Farès Menasri «Contributions à la reconnaissance de l´´ecriture arabe manuscrite», Thèse de Doctorat, Université Paris Descartes. 2008

[NAG 00] Nagy G., « Twenty Years of Document Image Analysis in PAMI, *Transactions on Pattern Analysis and machine Intelligence*», Vol. 22, No 1, January 2000.

[NEM 09] Nemouchi Soulef, Farah Nadir « Reconnaissance de l'Ecriture Arabe par Systèmes Flous», 2009.

[NIC 06] Stéphane Nicolas, Thierry Paquet, Laurent Heutte, «Un panorama des méthodes syntaxiques pour la segmentation d'images de documents manuscrits», Laboratoire PSI CNRS FRE 2645 - Université de Rouen Place E. Blondel UFR des Sciences et Techniques F-76 821 Mont-Saint-Aignan cedex. 2006.

[NOS 02] A. Nosary. «Reconnaissance automatique de textes Manuscrits par adaptation du scripteur». Thèse de doctorat, Université de Rouen, 2002.

[OFF 06] Office québécois de la langue française, « *Le grand dictionnaire terminologique* », (en ligne), septembre 2006. Disponible sur : http://www.granddictionnaire.com.

[PAQ 00] Paquet T., "Reconnaissance de l'écriture manuscrite: des modèles aux systèmes", Habilitation à diriger les recherches, Université de Rouen, Décembre 2000.

[PAR 96] PARKER J. R., "Algorithms for image processing and computer vision", Wiley, John & Sons, Incorporated, 1996.

[PEC 03] M. Pechwitz, V. Märgner,"HMM based approach for handwritten Arabic word recognition using the IFN/ENIT–database", Proceeding of ICDAR'03, 7th International Conference on Document Analysis and Recognition, Vol. 2, pp. 890-894, Edinburgh, Scotland, 2003.

[PIT 90] I.Pitas, A.N.Venetsanopulos, « Digital nonlinear filters », Kluwer Academic Press, 1990.

[ROB 02] PICOCHE J. « Dictionnaire étymologique du français ». Paris : Le Robert, Edition Gilles Firmin, 2002, collection « les usuels

[ROD 09] José A. Rodriguez-Serrano, Florent Perronnin «Handwritten word-spotting using hidden Markov models and universal vocabularies». ». Journal Pattern Recognition. Volume 42 Issue 9, September, 2009 Elsevier Science Inc. New York, NY, USA

[ROU 07] Laëtitia Rousseau, «Reconnaissance d'écriture manuscrite hors-ligne par reconstruction de l'ordre du tracé en vue de l'indexation de documents d'archives». Thèse de doctorat, l'Institut National des Sciences Appliquées de Rennes, 2007.

[SAF 00] SAFABAKHSH R., KHADIVI S. "Document Skew Detection Using Minimum-Area BoundingRectangle", The International Conference on Information Technology: Coding and Computing, March 27 - 29, 2000, pp. 253-258.

[SAU 00] J. Sauvola, M. Pietikainen, « Adaptive document image binarization », Pattern Recognition 33 (2) (2000) 225–236.

[SAY 73] K.M. Sayre. «Machine recognition of handwritten words : A project report». Pattern Recognition, vol. 5, pages 213–228, 1973.

[SEH 00] SEHAD A., HOCINI H., SEHAD M., AMEUR S., "Analyse des documents par la méthode des k plus proches voisins", CVA, 18-20 Novembre 2000, TIZIOUZOU, Algérie.

[SEH 04] A. Sehad, L. Mezai, M.T. Laskri, M. Cheriet, «Détection de l'inclinaison des documents arabes imprimés», 8ème Colloque International Francophone sur l'Ecrit et le Document, CIFED'2004, La Rochelle, France, 21-25 Juin, 2004.

[SOU 02] Souad SOUAFI, « Contribution à la reconnaissance des structures de documents écrits : Approche probabiliste », Thèse de doctorat, L'institut National des Sciences Appliquées de Lyon, 21 septembre 2002, Numéro d'ordre: 02 ISAL 0043.

[SOU 06] Souici-Meslati L, «Reconnaissance des mots arabes manuscrits par intégration neuro-symbolique», Thèse de Doctorat d'Etat, Labo. LRI, Département d'informatique, Université d'Annaba, Algérie, Février 2006.

[SRI 93] Rohini K. Srihari & Charlotte M. Baltus. «Incorporating Syntactic Constraints in Recognizing Handwritten Sentences». IJCAI, pages 1262–1267, 1993.

[SRI 05] Pavithra Babu Sargur Srihari, Harish Srinivasan and Chetan Bhole. «Handwritten arabic word spotting using the cedarabic document analysis system». In 2005 Symposium on Document Image Understanding Technology, The Marriott Inn and Conference Center, University Maryland University College, Adelphi, Maryland, November 2-4, 2005.

[SRI 06] Pavithra Babu Sargur Srihari, Harish Srinivasan and Chetan Bhole. «Spotting words in handwritten arabic documents». In In Procs. of SPIE, San Jose, CA, USA, Jan 2006.

[STE 95] Mémoire du monde : Principes directeurs pour la sauvegarde du patrimoine documentaire /document élaboré pour l'UNESCO au nom de l'IFLA par Stephen Foster, Jan Lyall, Duncan Marshall et Roslyn Russel. - Paris : UNESCO, 1995.

[SYI 06] Syiam M., Nazmy T. M., Fahmy A. E., Fathi H., Ali K., "Histogram clustering and hybrid classifier for handwritten arabic characters recognition", Proceedings of the 24th IASTED International Multi-Conference Signal Processing, Pattern Recognition, and Applications, pp. 44-49, Innsbruck, Austria, 15-17 February 2006.

[TAB 98] TABIZA Mohammed, « Filtres Lp : Etude des propriétés et Application en Traitement d'images », thèse de doctorat, Spécialité: Electronique-Electrotechnique-Automatique, université de Savoie, 16 Mars 1998

[TRI 95]I. Trier, A. Jain et T. Taxt, «*Feature extraction methods for character recognition – A survey*», Pattern Recogntion, 1995.

[VIN 04] A. Vinciarelli, S. Bengio & H. Bunke. «Offline Recognition of Unconstrained Handwritten Texts Using HMMs and Statistical Language Models». IEEE Trans. on PAMI, vol. 26, no. 6, pages 709–720, 2004.

[WAK 98] Waked, B. Bergler, S. Suen, C.Y. Khoury, S. « Skew Detection, Page Segmentation, and Script Classification of Printed Document Images », in: IEEE Systems, Man, and Cybernetics, 1998.

[WOL 06] C. Wolf. « Document Ink bleed-through removal with two hidden Markov random fields and a single observation field », Technical Report RRLIRIS2006-019. Lyon: LIRIS, INSA Lyon, 2006.

[WON 82] Wong K., R. Casey, F. Wahl (1982), « Document analysis system », I.B.M. Journal of Research and Development, 26, no 6.

[YIN 01] YIN P.Y., "Skew Detection and Block Classification of Printed Documents", Image and Vision Computing, 2001, pp. 567-579.

[ZAH 04] Abderrazak Zahour, Bruno Taconet, Saïd Ramdane, « Contribution à la segmentation de textes manuscrits anciens », CIFED, 2004