République Algérienne Démocratique et Populaire
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique
Université Cheikh Larbi Tebessi –Tébessa
Faculté des Sciences et Technologie
Département Informatique

Mémoire de Magister

Ecole **D**octorale en **S**cience et **T**echnologie de l'**I**nformation et de la **C**onnaissance (INI)
Spécialité : **Informatique**
Option : **SIC**

Catégorisation de Textes Arabes

Présenté par : **Samir KHEDIRI**

Dirigé par : **Nadir FARAH**

Jury composé des personnes suivantes :

Président	**Pr.**	**Hamid SERIDI**	**Université de Guelma**
Rapporteur	**Dr.**	**Nadir FARAH**	**Université de Annaba**
Examinateur	**Dr.**	**Med Tarek KHADIR**	**Université de Annaba**
Examinateur	**Dr.**	**Djamel MESLATI**	**Université de Annaba**

2009

Résumé

Grâce aux progrès des technologies numériques, la préservation et la valorisation de notre patrimoine documentaire est devenue un enjeu majeur mais, par la suite, elle a posé des difficultés d'accès à l'information et à son organisation. L'analyse des documents peut apporter une solution, mais les méthodes classiques ne sont pas suffisamment souples pour s'adapter à la variabilité rencontrée. Pour palier à ce problème, nous proposons comme objectif : la catégorisation automatique des textes manuscrits arabes, par une technique d'appariement approximatif des chaînes de caractères, pour contourner les difficultés des méthodes de classification classiques basées sur l'apprentissage. Afin de contribuer à la sauvegarde et à la valorisation de l'énorme héritage culturel constitué de milliers d'ouvrages et de documents manuscrits anciens dont nous disposons.

Mots-clés : Catégorisation automatique de texte, valorisation du patrimoine, analyse et traitement d'images, reconnaissance de l'écriture manuscrite arabe, distance d'édition, classification.

Abstract

With the progress of numerical technologies, the safeguarding and the valorisation of our documentary inheritance became a major stake but, thereafter, raised difficulties of access to information and have its organization. The analysis of documents can bring a solution, but the traditional methods are not sufficiently flexible to adapt to the variability. To solve this problem, we propose, the automatic categorization of the Arab handwritten texts by Approximate String Matching to avoid the difficulties of the classification based on the machine learning (training). In order to contribute to the valorisation of the enormous cultural heritage made up of thousands of works and old handwritten documents that we have.

Keywords: Automated text categorization, patrimonial document valorisation, image analysis and processing, Arabic handwriting recognition, edit distance, classification.

ملخص

مع تطورالتكنولوجيا الرقمية ، أصبح من الواجب المحافظة وتعزيز هذا التراث الوثائقي الذي أصبح يشكل تحديا كبيرا و الذي لاحقا ، أصبح يشكل صعوبات في الحصول علـــى المعلومات وتنظيمها. تحليل الوثائق يمكن أن تعتبر حلا لهده المشكلة، ولكن الطرق التقليدية ليست مرنة بما فيه الكفاية للتكيف مع التغيرات المصادفة و للتغلب على هذا العائق، نقترح التصنيف الآلي للنصوص على المخطوطات العربية من خلال تقنية حساب المسافات التقريبي للحروف للتغلب على صعوبات التصنيف التقليدية و للمساهمة في حفظ و تثمين التراث الثقافي الهائل المتكون من الآلاف من الكتب والمخطوطات القديمة التي بحوزتنا .

الكلمات الدالة : التصنيف الآلي للنصوص ، تثمين التراث الوثائقي ، تحليل ومعالجة الصور، التعرف على خط اليد العربي ، تعديل المسافة ، التصنيف.

Table des matières

Liste des tableaux

Liste des figures

A la mémoire de mon défunt Père et celui de mon encadreur.

A la plus belle mère du monde et à mes frères
Zidane, Adlen et Said.

Remerciements

Je suis reconnaissant envers Dieu le tout puissant, qui ma permit d'accomplir et d'achever ce modeste travail.

Je tiens tout d'abord à remercier mon directeur de recherche M.Nadir FARAH, qui a encadré ce travail, pour ses nombreux conseils, suggestions et sa gentillesse remarquable.

Je remercie également M.Hamid SERIDI, M.Med Tarek KHADIR et M.Djamel MESLATI, qui ont accepté de faire partie du jury.

Un immense merci à ma famille et à mes amis M.Kardache, Naoufel, Abdelatif, Rabie, Fethi, Nourddine, Abdeljalil, Abdelhakim, pour leurs soutiens et encouragements. Un merci tout particulier à ma belle Sarah, discrète mais toujours perspicace sans oublier mon cher ami Abdelhakim pour son aide précieuse.

Enfin, un grand merci à tout ceux qui m'ont aidé, de prés ou de loin, lors de la réalisation de ce mémoire.

Introduction générale

Actuellement, les bibliothèques et musées du monde détiennent des recueils de grand intérêt et regorgent d'ouvrages riches en informations qui malheureusement ne peuvent être accessibles du fait de leur valeur et de leur état de conservation. Avec l'essor des technologies du numérique, il est possible maintenant de valoriser ce patrimoine culturel en proposant des répliques numériques de bonne qualité, ce qui permet de partager l'accès à l'information tout en préservant les documents originaux, dans un but de consultation d'archives ou de parcours culturel ou scientifique.

La mémoire collective de nos sociétés et civilisations est aujourd'hui en danger de perdition. Il est primordial de préserver ce bien inestimable (images de documents anciens ou historiques, des formulaires, des archives gouvernementales ou commerciales, des publications scientifiques ou techniques,…). C'est dans cette perspective que des démarches rapides doivent être menées, dans le but d'une conservation permanente de ces précieuses ressources, pour pouvoir léguer aux générations suivantes la connaissance accumulée depuis des siècles sous différentes formes. Ainsi, avec le développement des technologies numériques et l'éclatement récemment des projets de valorisation du patrimoine au moyen de la numérisation [SAPCCA[1] 04], [MADONNE[2] 06], [DEBORA[3] 99]. Des campagnes de numérisation des collections à grande échelle ont été lancées pour une éventuelle sauvegarde des données dans tous les domaines. On se retrouve par la suite avec des quantités phénoménales de documents, le plus souvent des manuscrits, représentant des mines d'or de savoir et de connaissance enfuis dans des bibliothèques publiques ou dans des réserves inaccessibles au grand public du à leur rareté et à leur état de détérioration. Se pose alors, le problème de la valorisation et de l'indexation de ces biens numérisés et la réutilisation de ces contenus [Bensefia 03]. L'analyse d'images de documents peut apporter une solution à l'indexation et à la gestion des masses de données issues de la numérisation. Cependant, un nombre restreint de travaux se sont intéressés à l'analyse des structures dans les documents manuscrits. Ajoutons à cela, le problème non résolu de l'analyse de

[1] www.lri-annaba.net/sapcca/index.htm
[2] l3iexp.univ-lr.fr/madonne/index.html
[3] debora.enssib.fr/

pages complètes d'écriture. En effet, le traitement des documents patrimoniaux soulève de nombreuses difficultés. Ces difficultés sont liées à la grande variété des problèmes et des types de documents rencontrés, de plus les documents considérés sont souvent des documents dégradés et fortement bruités, caractérisés par une très grande hétérogénéité, à la fois dans les contenus, mais également dans la présentation et dans l'écriture.

Dernièrement, l'avancée des technologies numériques notamment dans le cadre d'applications spécifiques telles que la Gestion Electronique de Document (GED), et le développement des scanners à haut débit et à faible coût [Kirtas[4] 01], on a pu industrialiser des applications adaptées à des domaines précis comme le traitement automatique des chèques, le tri du courrier, la lecture d'adresses sur les enveloppes postales ou encore de formulaires. Malencontreusement, pour le traitement des ouvrages anciens le problème reste posé.

Une des réponses à cette problématique d'accès et de gestion de cette gigantesque masse de documents numérisés est : la catégorisation automatique ou classification supervisée du texte [Sebastiani 02]. En effet, l'accès à l'information textuelle a motivé depuis de nombreuses années les travaux des chercheurs issus de différentes communautés, malheureusement, les résultats sont mitigés pour le traitement du manuscrit et encore plus pour la langue arabe, du fait de la très grande hétérogénéité qui rend le traitement des documents patrimoniaux très pénible. Ces difficultés sont liées à la grande variété des problèmes et des types de documents rencontrés, ces documents patrimoniaux regroupent à la fois des documents anciens et des documents plus récents, des documents imprimés et des documents manuscrits. Additionnons à cela, les documents sont dégradés et fortement bruités. Ces difficultés demeurent en manipulant des documents hétérogènes comportant des informations de natures différentes (texte, images, graphiques) et des mises en page complexes. Dans le domaine de la reconnaissance de l'écrit, les efforts se sont principalement concentrés sur la reconnaissance de mots ou de phrases. Cependant l'analyse de pages complètes d'écriture reste encore un problème non résolu de manière acceptable.

Ce champ apparu relativement récent dans la recherche d'information, qui est la catégorisation, donne un espoir de réduire le travail humain de façon caractéristique, voire de le remplacer dans une proportion limitée pour gérer les immenses parcs d'informations ou de documents électroniques dans le monde.

Evidemment, s'agissant de catégorisation automatique des documents issus du patrimoine, qui est le but de notre étude, les données initiales sont constituées d'images numérisées de documents manuscrits. On peut distingué déjà, deux catégories, les archives de types textuels et les archives exclusivement de type graphique. Notre travail se focalisera sur le premier cas.

Les systèmes de catégorisation de documents permettent de détecter le ou les thèmes abordés dans ceux-ci. Cette catégorisation est réalisée en grande partie grâce à des outils issus de la recherche d'information. Ces modèles sont importants car ils sont à l'origine de nombreux modèles de catégorisation de textes, la distinction entre ces deux domaines n'est pas toujours aussi facile qu'on le pense.

Dans ce mémoire, nous allons aborder la catégorisation automatique d'anciens textes manuscrits issus du patrimoine culturel arabe de manière différente des techniques classiques entreprises dans ce domaine. L'idée est d'utiliser une approche d'appariement de chaînes de caractères pour arriver à une mesure de similarité acceptable selon un seuil à déterminer pour enfin, pouvoir se prononcer sur la décision de catégoriser tel document ou autre dans sa catégorie adéquate. On parle alors de la distance d'édition ou distance de Levenshtein.

Le plan de cette étude consacrée à la catégorisation automatique des manuscrits se fera comme suit : après cette introduction générale, d'abord, nous parlerons dans un premier temps dans le premier chapitre de la structure générale des documents et de leurs caractéristiques, dans le second chapitre, nous présenterons les grandes étapes qui composent une chaîne de

[4] www.kirtas-tech.com

reconnaissance de l’écriture manuscrite ainsi qu'une définition approfondie de l'écriture arabe et de ses aspects contraignants pour la tâche de reconnaissance. Nous énumérerons dans le chapitre numéro trois les traitements infligés aux documents numérisés, en commençant par la numérisation, les techniques de prétraitements. Vient ensuite, une tâche très délicate : l'analyse d'image de documents, que nous expliquerons par les différentes méthodes de segmentation. Nous consacrerons le quatrième chapitre à la description de la catégorisation automatique de textes et particulièrement à l'approche d'appariement approximatif des chaînes de caractères qui est le but de notre démarche. Le chapitre numéro cinq est réservé à l’expérimentation et aux résultats des tests. Enfin, nous achèverons l’étude, dans le chapitre six, par une conclusion, ainsi que les perspectives des futurs travaux.

CHAPITRE 1

STRUCTURE DES DOCUMENTS

1 Structures des documents

1.1 Introduction

La définition traditionnelle d'un document peut être donnée par rapport à son contenu et son support physique, l'exemple classique est celui du livre présenté sur support papier sous la forme d'une suite de pages. Avec le développement des nouvelles technologies dans le domaine de la numérisation et les différents formats de stockage, le support électronique représente l'information sous forme d'un codage en pages numériques (Postscript, pdf, …) ou d'images de pages scannées.

Cette composition graphique et spatiale d'un document est appelée structure ou forme, elle a pour fonction d'organiser le contenu de façon à en simplifier la lecture et la compréhension [Ramel 06].

1.2 Types de document

Depuis le temps que l'humain a découvert l'écriture, il utilisé des documents de différentes sortes comme les contrats, les récits, les écritures comptables,… A l'égard des éléments distincts du contenu que peuvent contenir les documents, nous conduit a différencier le document par le suivi ou non de règle d'élaboration. On trouve le document structuré, le document image et le document principalement graphique.

1.2.1 Document structuré

Le document structuré est un document à forte teneur textuelle dont les contenus respectent une règle de présentation et d'organisation qui permet à leur compréhension par le lecteur. On discerne différentes classes de documents structurés [Haou 08] :

- Document textuel complexe : comportant trop de complexité pour identifier une structure unique, comme les pages de journaux, des anciens documents et en général tout documents susceptible d'être organisé en deux dimensions.
- Document textuel incorrectement structuré : type de documents comprenant une structure non cohérente et peu de règles sont respectées, la structure peut varier au long d'un même document.
- Document textuel respectant des règles de structuration : documents respectant des bases de structuration comme les fiches de catalogues de bibliothèque.
- Document à structure rigide : document produit par des systèmes de gestion de base de donnée comme les formulaires.

1.2.2 Document image

Le document image peut être vu comme des photographies, des dessins, des écrits manuscrits, mais aussi tout autres documents divers qui sont numérisés (les télécopies, contrats, déclarations, ...). Ces documents ne sont caractérisés par aucune structure interne mais souvent accompagnés d'une fiche descriptive structurée.

1.2.3 Document graphique

Le document graphique ou à forte teneur graphique est utilisé comme support de représentation d'information. Les domaines où on peut trouver ce genre de document sont étendus comme les cartes géographiques, topologiques, plans cadastraux, plans d'accès, les schémas, plans d'architectures, plans d'évacuation, les organigrammes, plans de réseaux téléphoniques ou électriques.

Les documents graphiques comprennent une quantité importante d'informations, peuvent être d'une grande complexité et leur domaine d'application est très vaste. Ils sont constitués de lignes, de régions pleines, de régions hachurées, de texte, ... (figure 1-1) [Ramel 06].

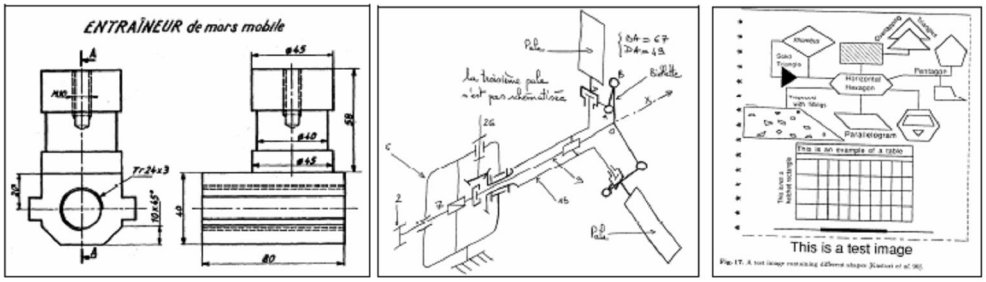


Figure 1-1 : Exemple de documents graphiques

1.3 Structuration d'un document

La structure d'un document correspond à différents niveaux d'abstraction de l'information. Elle identifie chaque éléments composants ce dernier. En effet, on distingue généralement deux types de structures : la structuration physique et la structuration logique [Chatelain 06].

Structure logique d'un document : La structuration logique est une abstraction qui décrit l'organisation du contenu du document, la manière d'éclaircir l'information, ne tenant compte ni de son support ni de sa présentation. C'est une description hiérarchique et logique du contenu d'un document au moyen d'entités logiques telles que les sections, les chapitres, les paragraphes, les titres, ...

Structure physique d'un document : La structure physique décrit au biais d'entités physiques telles que les caractères, les mots, les lignes de texte, les blocs, les objets graphiques et typographiques, ... le modèle de représentation graphique des entités logiques, c'est à dire la mise en page du document.

Il existe donc une relation entre la structure logique et la structure physique ce qui peut amener à confondre les deux. Cependant, même si toutes les deux peuvent conduire à un découpage du document similaire, la manipulation des éléments et l'accès à chacune demeure très différent.

1.4 Le document imprimé

Il existe une relation fortement identifiable entre la structure logique du document, et sa structure physique, dans le cas des documents imprimés. En effet, nous avons dis que la structure physique est la traduction de la mise en forme appliquée à la représentation logique du document par un suivi de règles de typographie et de mise en page, qui ont pour but de faciliter la compréhension par le lecteur.

Le processus d'analyse dans le cas des documents imprimés, est un processus inverse du processus qui a conduit à la création du document. Il s'agit de ce fait de retrouver l'information que l'auteur du document a voulu transmettre. Pour cela on cherche généralement à déterminer la structuration physique à partir de l'image du document, puis à déduire de la structure physique les règles typographiques permettant de retrouver la structure logique de ce document [Nicolas 06].

1.5 Le document manuscrit

Les documents manuscrits sont des documents en majorité textuels, et dont la structure est relativement simple contrairement aux documents imprimés. Cependant cette structure est caractérisée par une très forte variabilité spatiale qui se traduit par des lignes de texte inclinées et fluctuantes, des chevauchements entre les lignes, des espaces irréguliers entre les mots, ... Les documents manuscrits sont des documents non contraints qui ne suivent pas nécessairement des règles de structuration explicitement définies. Néanmoins, la plupart du temps, un corps de texte occupe la majorité de la page avec des notes en marge du texte (figure 1-2). La page peut aussi contenir des illustrations graphiques de différentes tailles, des ornements et des lettrines…

En analyse des documents manuscrits, Contrairement aux documents imprimés, on ne cherche pas à déterminer la structure physique et la structure logique du document, mais on

considère plutôt une phase de segmentation et une phase de reconnaissance de l'écriture [Heutte 03].

Résultat, les notions de structure physique et de structure logique se rapportent plus aux documents structurés et imprimés, qu'aux documents manuscrits.

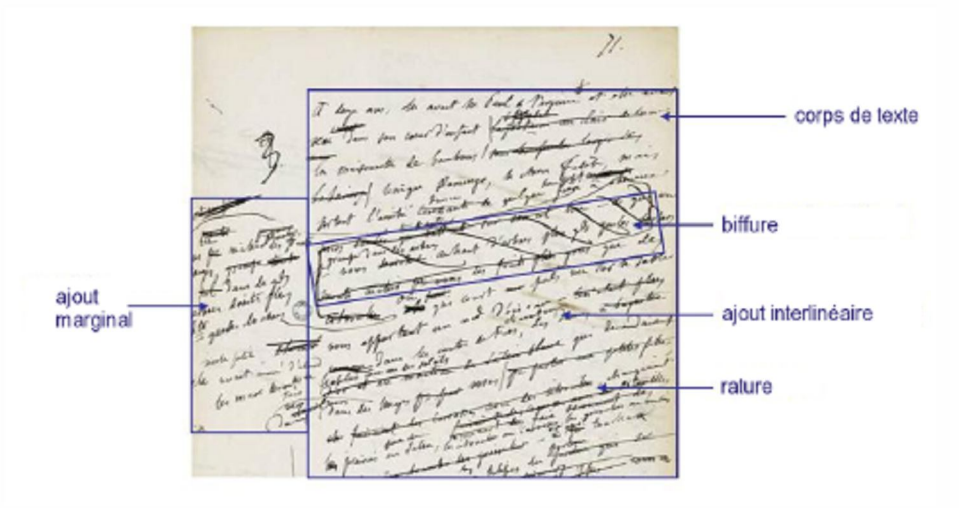


Figure 1-2 : Quelques caractéristiques de la structure d'un manuscrit

Les irrégularités des mises en page sont beaucoup plus nombreuses que dans les ouvrages actuels à cause d'imprécisions ou de libertés prises par l'imprimeur.

1.5.1 Les documents manuscrits anciens

Le terme *manuscrit*, est attesté en français pour la première fois en 1594 avec la généralisation de l'imprimerie qui a joué un rôle dans l'apparition du terme. Le manuscrit est pris dans le sens de livre manuscrit, à l'exclusion des documents, lettres ou autres papiers écrits à la main. En arabe, le mot *makhtut*, désigne la même signification [Humbert 02]. Se sont des documents rédigés à la main, qui peuvent être vues sous l'angle de leurs contenus textuels et/ou graphiques.

Figure 1-3 : Grande diversité des manuscrits du moyen Age au 18ème siècle

Il y a dans le monde environ cinquante mille manuscrits grecs et un demi million de manuscrits latins pour les manuscrits en écriture arabe, aux dires de certains spécialistes, il faudrait multiplier ce dernier chiffre par cinq ou six [Humbert 02]. La bibliothèque de la Süleymaniye, à Istanbul, conservait elle seule, environ cent mille unités, dont soixante-dix mille manuscrits arabes, le reste étant en persan ou en turc (figure 1-3).

Cette immense richesse des collections de manuscrits arabe a été produite sur plus d'un millénaire, constituant par la suite la forme la plus familière du livre. Ces des textes assez nombreux, encore mal réunis et peu étudiés. Ce qui nous conduit a parlé du métier de copiste qui n'est autre que le créature de ces ouvrages manuscrits, on l'appelle *kâtib*, *nassâkh*, *khattât* ou encore *warrâq*, et qui est parfois aussi *mudhahhib* (doreur), *naqqâsh* (enlumineur) ou *mujallid* (relieur). On peut distinguer le copiste amateur du professionnel par le style d'écriture soignée, une écriture professionnelle et une écriture calligraphiée [Humbert 02]. Beaucoup de travail doit être fait aussi sur les copies des manuscrits en Afrique concernant la production littéraire arabo-africaine [MARA 96].

1.5.2 Les caractéristiques des documents anciens

Le traitement des documents patrimoniaux soulève de nombreuses difficultés vu la variabilité de leurs structures et contenus (figure 1-4) [Likforman-Sulem 06], [Ramel 06] :

- La grande variabilité entre les styles d'écriture
- Faible espacement entre les lignes provoquant des contacts entre les caractères.
- Les différences de mise en page.
- La non régularité ou manque de similarités entre les formes.
- Usage fréquent d'ornements (zones non textuelles).
- Disposition fluctuante des illustrations graphiques.
- Les déformations dues à la courbure naturelle des pages.
- La présence de plis et de déchirures.
- La présence d'inclinaisons, des bombages du papier.
- La présence de tâches d'humidité absorbée par le papier.
- Transparence du recto sur le verso.
- La variation d'éclairage.

Ces défauts engendrent une perte d'information et augmente ainsi les difficultés pour la reconnaissance de la structure des documents et par la suite leur analyse devient encore plus complexe. D'où la nécessité de mettre en œuvre une chaîne de numérisation et prétraitement pour mieux contrer ces obstacles.

Figure 1-4 : Quelques caractéristiques des documents anciens

1.6 Conclusion

De manière générale on peut distinguer les documents selon leur nature (imprimé, manuscrit), selon leur contenu (données textuelles, graphiques ou hétérogènes), selon leur niveau de structuration (documents structurés ou documents non contraints) ou encore suivant leur niveau de dégradation (documents anciens ou bruités, et documents faiblement dégradés). Cette variabilité dans certains documents rend la tâche de reconnaissance de l'écriture manuscrite encore plus difficile comme nous allons le voir dans le chapitre suivant.

CHAPITRE 2

LA RECONNAISSANCE DE L'ECRITURE MANUSCRITE

2 La reconnaissance de l'écriture manuscrite

2.1 Introduction

L'écriture est un moyen de communication encore essentiel dans nos civilisations. Elle peut avoir différents styles, imprimée ou manuscrite. L'imprimé est généralement codé selon une fonte particulière qui rend plus facile le développement des systèmes automatiques de traitement. Le manuscrit renvoi une écriture plus personnelle et intime, spécifique à son auteur.

Le but de la reconnaissance de l'écriture est de transformer un texte écrit en une représentation compréhensible par une machine et facilement reproductible par un logiciel de traitement de texte. Cette tâche n'est pas ordinaire car les mots possèdent une infinité de représentations. D'une part, chaque personne produit une écriture qui lui est propre, et d'une autre part, il existe de nombreuses polices de caractères pour l'imprimé avec de nombreux styles (gras, italique, souligné, ombré etc.) et des mises en page différentes et complexes. Suivant le type d'écriture qu'un système doit reconnaître (manuscrit, cursif ou imprimé), les opérations à effectuer et les résultats peuvent varier considérablement. Surtout concernant l'écriture manuscrite, son traitement soulève de très nombreux problèmes décrits dans la section 2.5 [Caillault 05].

2.2 Reconnaissance de l'écriture manuscrite

Beaucoup de travaux se sont attaqués au domaine de la reconnaissance de l'écriture manuscrite pour la langue latine: l'anglais, le français, l'italien, l'espagnol et même pour les langues asiatiques comme le chinois et le japonais [Vinciarelli 01] ; récemment, quelques recherches ont vu le jour, centrées sur l'écriture arabe qui est complexe et contraignante, plus que celles des autres langues, son traitement soulève de très nombreux problèmes [Lorigo 06], [Farah 05], [Amin 98], [Gumah08].

Cette tâche de reconnaissance n'est pas aisé car les mots possèdent une infinité de représentations, chaque personne produit une écriture qui lui est propre, des styles différents et des mises en page différentes et complexes. Ajoutons à cela, une complexité accrue de l'écriture arabe qui est une langue sémite, de nature cursive, comportant 28 lettres fondamentales [BenAmor 06]. Le sens d'écriture va de droite à gauche. Il n'y a pas de différence entre les lettres manuscrites et les lettres imprimées; les notions de lettre capitale et lettre minuscule n'existent pas en arabe. En revanche, la plupart des lettres s'attachent entre elles, même en imprimerie, et leur graphie diffère selon qu'elles soient précédées et/ou suivies d'autres lettres ou qu'elles soient isolées (variantes contextuelles)[Amin 98].

Il existe deux types de reconnaissance du manuscrit suivant le mode de saisie de l'écriture: la reconnaissance dynamique ou en-ligne et la reconnaissance statique, différée ou hors-ligne, cette dernière qui nous intéresse débute après l'acquisition du document [Arrivault 02], [Lorigo 06].

2.3 Les secteurs du manuscrits

Il existe deux secteurs d'applications en reconnaissance du manuscrit suivant le mode d'acquisition ou de saisie de l'écriture [Arrivault 02] :

2.3.1 La reconnaissance dynamique ou en-ligne

La reconnaissance en ligne ou en temps réel est une reconnaissance dynamique se déroulant pendant l'écriture. Un léger retard, d'un mot ou d'un caractère, permet à la reconnaissance de ne pas empiéter sur la saisie. La réponse en continue du système permet à l'usager de corriger et de modifier son écriture de manière directe et instantanée. Pour écrire ou dessiner, on utilise un crayon ou un stylo, C'est en partant de ce postulat que sont mis au point des systèmes informatiques basés sur l'interaction stylo : l'utilisateur dessine avec un stylo électronique sur une surface tactile, comme il le ferait sur une feuille de papier.

Par rapport à l'utilisation d'un clavier et d'une souris, l'interaction stylo présente plusieurs avantages. Ce mode d'interaction utilise un matériel moins encombrant et plus mobile qu'un ordinateur classique. Il est donc adapté aux besoins des utilisateurs éloignés de leur poste de travail ou plongés dans des milieux où les outils informatiques habituels ne sont pas utilisables. De plus, l'interaction stylo permet un dialogue bien plus naturel, intuitif et convivial entre l'utilisateur et le système informatique. En effet, le stylo électronique et la surface tactile reproduisent la métaphore du papier-stylo. Ceci facilite la prise en main des systèmes informatiques, offrant à un public très large un accès à ces technologies.

Différents types de matériels basés sur une interaction stylo ont émergé ces dix dernières années : certains de grande taille définition comme les tablettes graphiques, les ordinateurs portables de type « *Tablet PC* », d'autres plus petits comme les assistants personnels numériques (PDA) ou les « *smartphones* » (figure 2-1) [Carbonnel 05].

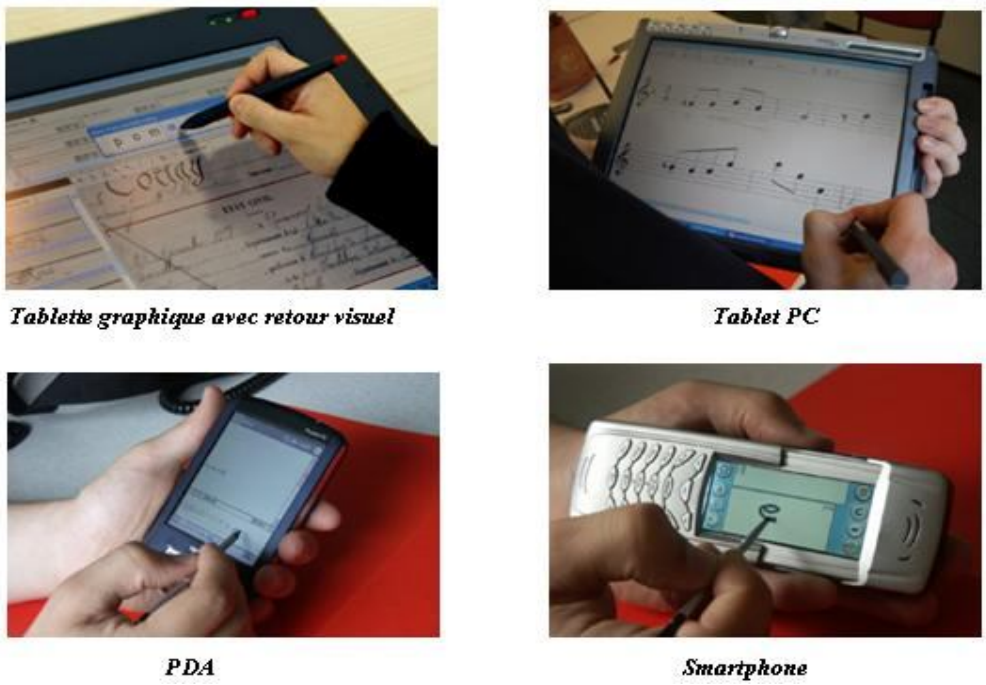


Figure 2-1 : Outils de saisie d'écriture en mode dynamique

Ces différents matériels nécessitent des logiciels adaptés. L'équipe IMADOC[5] de l'IRISA privilégie le développement d'approches génériques, pouvant être adaptées aux différentes applications.

2.3.2 La reconnaissance statique ou hors-ligne

La reconnaissance hors ligne ou en différé démarre après l'acquisition du document en entier. Ce type de reconnaissance convient pour les documents imprimés et les manuscrits déjà rédigés. Ce mode permet l'acquisition instantanée d'un nombre important de caractères, mais impose d'effectuer des prétraitements coûteux pour retrouver l'ordre de la lecture. C'est ce secteur qui nous intéresse.

2.4 Les applications

La reconnaissance de l'écriture est mieux connue sous le nom d'OCR (Optical Character Recognition), du fait de l'emploi de procédés d'acquisition optique. L'OCR connaît plusieurs applications pratiques dans plusieurs domaines d'activité parmi lesquels on peut citer :

- Les banques et les assurances pour l'authentification de chèques bancaires (correspondance entre montants et libellé d'une part, et correspondance entre l'identité du signataire et sa signature, d'autre part), et la vérification de clauses de contrats pour les assurances.
- La poste pour la lecture des adresses et le tri automatique du courrier.
- Les télécommunications pour l'échange de fichiers informatiques à distance.
- La police et la sécurité pour la reconnaissance de numéros minéralogiques pour le contrôle routier, l'authentification et l'identification de manuscrits et l'identification du scripteur.
- Les affaires et l'industrie pour la gestion des stocks et la reconnaissance de documents techniques.
- La bureautique pour l'indexation et l'archivage automatique de documents, et pour la publication électronique.

[5] www.irisa.fr

- L'administration pour la reconnaissance de plans cartographiques et la lecture automatique de documents administratifs.

2.5 La problématique de la reconnaissance de l'écriture manuscrite

La complexité de la reconnaissance d'information manuscrite dépend de plusieurs critères (figure 2-2) [Camillerapp 02], [Tappert 90] :

- Contrairement aux documents imprimés, les lettres d'un même mot manuscrit sont reliées entre elles (chevauchement). Les lettres n'étant plus clairement segmentées, leur reconnaissance devient délicate (le paradoxe de Sayre [Sayre73]).
- Une mise en page plus souple sur l'imprimé que sur le manuscrit. En conséquence, si on retrouve globalement les mêmes contraintes implicites à la rédaction d'un courrier, 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 par rapport 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. Cette variabilité inhérente aux documents manuscrits se retrouve également dans l'espacement des mots dans une ligne. A l'inverse 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.
- Complexité des brouillons du manuscrit, on trouve des paragraphes barrés de plusieurs traits, paragraphes inclinés, annotation dans les marges, écriture penchée, lignes surchargées, mots barrés taches, trous, parties manquantes, papier froissé, encre qui traverse le papier, tracés ou écriture partiellement effacés, certaines cases se sont avérées trop petites à l'usage. Les secrétaires ont donc collé de petites feuilles annexes qui masquent largement la structure du document, des tampons qui viennent brouiller l'aspect visuel de la page,…, des éléments qui viennent perturber le bon déroulement des extracteurs.
- Le conditionnement de l'information, l'écriture reconnue peut être plus ou moins conditionnée par la présence de précasé (cas des formulaires, code postal d'une adresse), des cadres (montants de chèques) ou lignes (lignes

d'un bloc adresse, montant littéral d'un chèque), ou bien non conditionnée (documents libres).

- Le nombre de scripteurs : la réduction du nombre de scripteurs potentiels permet éventuellement de réduire la variabilité et d'apprendre les différents styles d'écriture [Nosary 02]. La difficulté s'accroît en contexte omni-scripteur en raison des styles d'écriture très différents de chacun (figure 2-3) [Tappert 90].
- La taille du vocabulaire : les systèmes de reconnaissance de textes sont souvent basés sur un lexique qui facilite grandement la lecture (Kimura 1994), surtout si celui ci possède un faible nombre de mots (cas des montants littéraux de chèques qui contiennent une trentaine de mots). La reconnaissance de mots est d'autant plus aisée que le nombre de mots dans le lexique est faible. Notons que dans le cas de la reconnaissance de séquences numériques, la présence d'un lexique est plus rare (cas de la reconnaissance de codes postaux) [Augustin 01].
- Problèmes d'acquisition, mauvaises conditions d'éclairage, mauvaise orientation du document ou de la caméra, qualité du capteur, présence d'éléments extérieurs, courbure du texte sur les bords due à l'épaisseur du livre, …) [Arrivault 02].

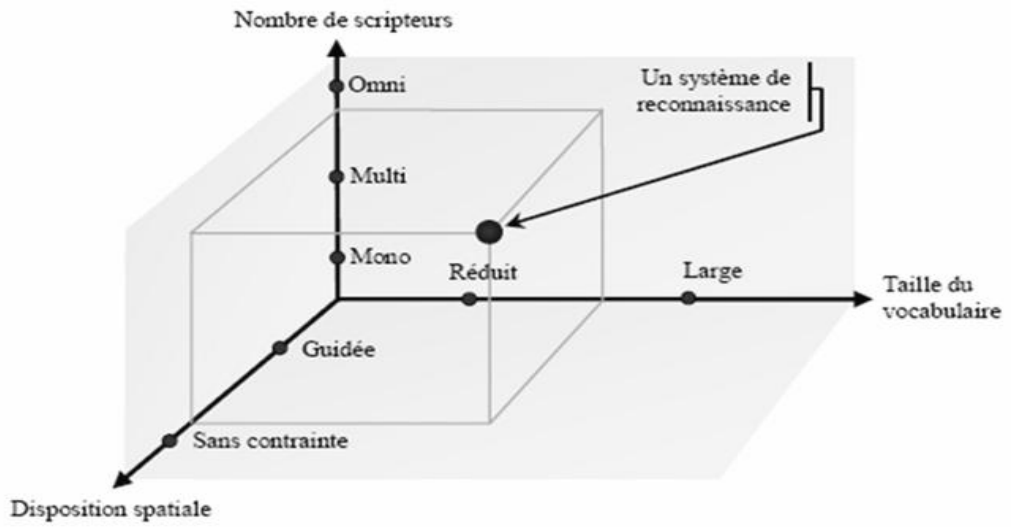


Figure 2-2 : Graphe de complexité des systèmes de reconnaissance du manuscrit

De façon générale cette problématique du manuscrit est présentée comme un dilemme dans [Sayre73] : «pour reconnaître une entité, il faut savoir la localiser et pour la localiser, il faut tout d'abord la reconnaître».

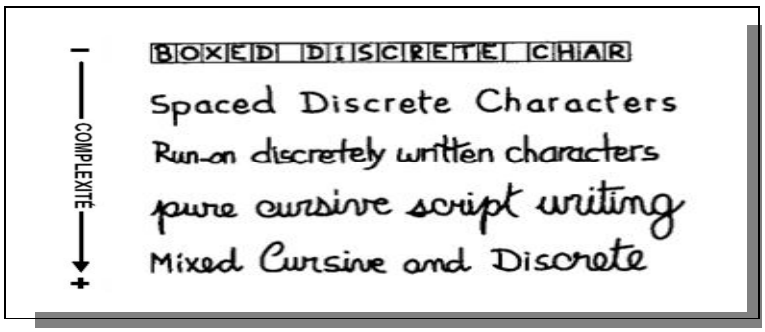


Figure 2-3 : Les cinq différentes classes d'écriture proposées par C.Tappert

2.6 Système de reconnaissance de l'écriture manuscrite

En général, l'objectif de la reconnaissance de l'écriture manuscrite est de développer un système qui se rapproche le plus de l'être humain dans sa capacité de lire. Cependant, cette reconnaissance de l'écriture consiste à extraire d'une forme inconnue (mot, lettres, chiffres) une description plus simple et à établir sur celle-ci une décision.

Cette décision est effectuée généralement en mesurant la ressemblance d'une forme inconnue avec un ensemble de références stockées en mémoire et décrites dans une représentation analogue. Les références sont obtenues lors d'une phase antérieure qualifiée d'apprentissage. Cette phase est très importante dans tout système de reconnaissance de l'écriture. Autrement dit c'est un passage de l'espace observable vers un espace de décision d'appartenance à une classe [Miled 97].

La construction d'un système de reconnaissance de l'écriture comprend plusieurs étapes distinctes (figure 2-4) [Heutte 03]**: Acquisition, prétraitements, extraction des caractéristiques, apprentissage, reconnaissance et post-prétraitements.**

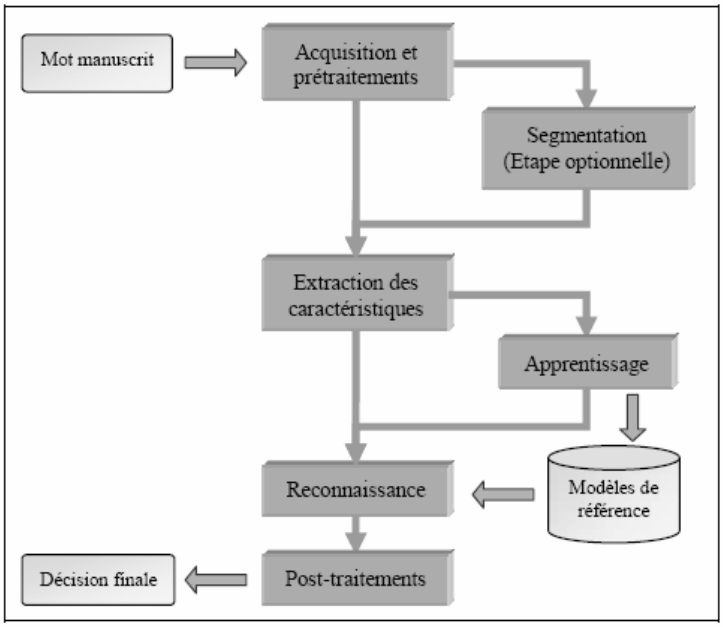


Figure 2-4 : Structure générale d'un système de reconnaissance de mots manuscrits

2.6.1 L'acquisition

C'est la capture et la transformation d'un document en pixels, cette numérisation conduit à un nouvel objet dont les processus de mise à disposition pour les utilisateurs restent encore objet de recherche. *(détaillée dans le chapitre 3 section 3.2)*

2.6.2 Le prétraitement

C'est des techniques qui consistent à améliorer la qualité des images en éliminant les défauts dus à l'éclairage et au processus d'acquisition *(détaillée dans le chapitre 3 section 3.3)*

2.6.3 Extraction des caractéristiques

L'objectif de l'extraction des caractéristiques est la sélection de l'information pertinente depuis une masse d'information globale acquise pour différencier un objet d'un autre, c'est pour

cette raison qu'un système de reconnaissance a besoin que de cette information discriminante. Pour ce faire, une étape d'extraction de caractéristiques est réalisée. C'est une phase cruciale lors de la construction d'un système de reconnaissance due à une perte d'information sensible par la plupart des techniques d'extraction. De ce fait, il faut effectuer un compromis entre quantité et qualité de l'information. Cette perte se trouve au niveau de la production de l'écriture, de l'acquisition, de la réduction et de la segmentation. Plusieurs techniques existent se basant sur le squelette, le pixel ou le contour [Lorigo 06] (Tableau 2-1).

Squelette	Contour	Pixel
Mozaffari 2005	Safabakhsh 2005	El-Hajj 2005
Haraty 2002-2004	Souici-Meslati, Farah 2004	Clocksin 2003
Amin 2003, 1996	Sari 2002	Pechwitz 2003
Khorsheed 2003	Snoussi Maddouri 2002	Al-Badr (morphologie) 1998
Fahmy 2001	Dehghani 2001	Motawa (morphologie) 1997
Abuhaiba 1993-1998	Olivier, Miled 1996, 2001	Al-Yousefi (projection) 1992
Almualim 1987	Romeo-Pakker 1995	

Tableau 2-1 : Techniques utilisées par différents auteurs pour l'extraction des caractéristiques

2.6.3.1 Niveaux des caractéristiques

Les caractéristiques peuvent être extraites à partir des mots, des lettres ou des sous-lettres, donnant ainsi lieu aux trois niveaux de caractéristiques :

Caractéristiques de bas niveaux : extraites à partir des sous-lettres, ayant des formes élémentaires tel que les petites lignes, les courbes, les traits, les barres,…, et des particularités géométriques simples, ce qui fait qu'elles soient très appréciées.

Caractéristiques de niveaux moyen : extraites à partir des lettres, généralement utilisées dans les systèmes de reconnaissance des caractères cursifs basés sur la segmentation explicite,

citons, à titre d'exemple de caractéristiques de niveau moyen, les distributions de transitions entre le fond et l'écriture.

Caractéristiques de haut niveau : ce sont les caractéristiques perceptuelles, facilement visibles, consistant en la détection d'éléments structurels, elles sont indépendantes des styles d'écritures évitant ainsi le problème de la variabilité des formes. Parmi les caractéristiques de haut niveau on peut citer : les boucles, les ascendants, les descendants, en plus des jonctions, les points finaux et traits, les barres des t et les points diacritiques pouvant être utilisés pour trouver une représentation approximative du mot, ceci permet de se débarrasser d'une partie du lexique ou de rejeter un résultat du processus de reconnaissance dont la représentation n'est pas compatible avec celle détectée [Heutte 03], **[**Nosary02].

2.6.3.2 Représentation des caractéristiques

Le schéma de représentation des caractéristiques varie en fonction des contraintes d'implémentation et des éventuelles stratégies utilisées.

Vecteur et matrice : ce schéma est généralement utilisé pour représenter les caractéristiques de niveau bas et intermédiaire. L'image du mot est segmentée en portions dont les caractéristiques sont extraites et représentés par des valeurs booléennes, entières ou réelles dans une matrice ou un vecteur, cette représentation est moins convenable pour représenter les caractéristiques de haut niveau.

Comptage : généralement utilisé pour représenter les caractéristiques de haut niveau en calculant le nombre de caractéristiques existantes (nombre des ascendants, nombre des descendants, nombre des boucles,….).

Séquence : utilisé pour représenter les caractéristiques de niveau haut et intermédiaire qui permettent d'approximer le mot par une séquence de symboles représentant un ensemble de primitives structurelles.

Structure de graphe : l'image est représentée par un graphe où les nœuds correspondent aux différentes caractéristiques, et les relations entre elles sont illustrés par des arcs, cette représentation graphiques est très puissante car elle montre la position des caractéristiques et leurs relations. [El-Hajj 05].

2.6.3.3 Types de caractéristiques

Il existe une grande diversité des caractéristiques utilisées en reconnaissance de mots. Le choix d'un type particulier de caractéristiques se révèle toujours difficile pour le concepteur d'un système de reconnaissance de mots manuscrits car les performances du système dépendent avant tout d'une bonne définition des caractéristiques. Les caractéristiques retenues doivent permettre de décrire de façon non équivoque toutes les formes appartenant à une même classe tout en les différenciant des autres classes. La difficulté du choix de ces caractéristiques réside dans le compromis à établir entre des contraintes telles que la rapidité de détection des caractéristiques, leur facilité de mise en œuvre et leur insensibilité aux distorsions (styles d'écriture différents, bruits dans l'image,…).

Il existe plusieurs types de caractéristiques, parmi les plus utilisées, nous citons les caractéristiques structurelles, et statistiques (Tableau 2-2) [Lorigo 06].

Auteurs	Primitives
Elhajj 2005	Densités de pixel, transitions de densité, et configuration des concavités le long des armatures en respectant les lignes de base.
Mozaffari 2005	Le changement de la moyenne et la variance de X et de Y dans les parties du squelette.
Almaadeed 2004	Ascendants, descendants, primitives structurelles.
Haraty 2002-2004	Les boucles, point extrême/jonction, largeur, hauteur, densités, hauteur des contours.
Souici, Meslati et Farah 2004	Les boucles, points, composantes connexes, ascendants, descendants.
Amin 2003	Les boucles, points, hamza, lignes, les courbes et leur relation
Snoussi maddouri 2002	Ascendantes, descendantes, boucles, points, positions de Pseudo-mot,

	descripteurs normalisés de Fourier.
Miled 2001	Concavités et inclinations pour les ascendantes et les descendantes, densités et forme des diacritiques.
Alyousefi 1992	Statistiques des moments horizontales et verticales des projections.
Almuallim 1987	Points extrêmes, largeurs, armatures, points de croisement, autres.

Tableau 2-2 : Caractéristiques utilisées par différents auteurs

Caractéristiques structurelles :

Dans ce type de caractéristiques, la description d'une forme se fait à partir de sa topologie et géométrie, en donnant sa propriété locale et globale. Il existe une multitude de caractéristiques, parmi les plus utilisées dans la reconnaissance globale des mots arabes [Farah 05], [Amin 98], [Lorigo 06] :

- Nombre de composantes connexes, car dans l'écriture arabe les scripteurs respectent généralement la séparation entre les composantes connexes d'un mot.
- Les points diacritiques, car ils permettent de distinguer les caractères ayant le même corps principal.
- Les hampes et les jambages qui sont généralement des primitives recherchées dans la perception humaine.
- Les boucles et cavités qui permettent la distinction entre les mots qui possèdent des primitives presque similaires.
- Les points extrêmes et points terminaux, les jonctions et croisements, les angularités.
- Les points d'inflexion, les points de rebroussement, …etc.

Caractéristiques statistiques :

Ces caractéristiques sont dérivées de la distribution des pixels appartenant au caractère, dans l'image entière ou dans certaines parties uniquement. Dans [Huette 03] trois familles de caractéristiques sont suggérées telles que : les moments invariants, les projections, et les profils. Elles sont extraites en considérant la distribution des pixels noirs de l'objet (caractère, mot, chiffres). Le processus d'identification de la meilleure méthode d'extraction de caractéristiques n'est pas évident. Par exemple, les moments de Zernike s'appliquent bien sur des images à

niveaux de gris et que la projection s'applique souvent sur des caractères segmentés pour résoudre leur problème de reconnaissance de l'écriture manuscrite. D'autres caractéristiques basées sur des informations extraites à partir de l'image du contour comme les descripteurs de Fourier ou encore la description locale des contours qui est une technique exprimant les frontières discrètes de l'image par une séquence de code comme la chaîne de Freeman. Cependant, il est nécessaire d'effectuer pour chaque problème de reconnaissance une évaluation expérimentale de quelques méthodes d'extraction de primitives les plus prometteuses. Ces expériences permettront de faire un choix judicieux des primitives à extraire car souvent, l'utilisation d'une seule méthode d'extraction de caractéristiques n'est pas suffisante pour obtenir de bonne performance du système de reconnaissance [El-Hajj05].

2.6.4 L'apprentissage

Cette étape permet de construire un dictionnaire de prototype. Il s'agit de regrouper en classes plusieurs prototypes dont les caractéristiques se rapprochent. A partir de l'étape de l'apprentissage le système doit ajuster ses paramètres afin de donner une réponse lors de l'étape de reconnaissance. Il existe 2 types d'apprentissages : supervisé et non supervisé [Falou 98].

- L'apprentissage dit superviser lorsque les différentes classes des échantillons sont connues à priori, et aussi parce qu'il est guidé par un superviseur. Ce dernier indique pour chaque échantillon en entrée le nom de la classe qui le contient.
- L'apprentissage dit non superviser ou apprentissage sans superviseur, c'est où le système essaye de construire automatiquement les classes à partir des échantillons de référence et des règles de regroupement sans intervention du superviseur **[**Burrow 04**].**

Il s'agit, lors de cette étape, d'apprendre au système les propriétés pertinentes des caractères du vocabulaire utilisé. L'idéal serait de lui apprendre autant d'échantillons que de formes d'écriture différentes, mais cela est impossible à cause de la trop grande variabilité de l'écriture qui conduirait à une combinatoire ardente des modèles de représentation. La tendance consiste alors à remplacer le nombre par une meilleure qualité des traits caractéristiques et de bonnes séparatrices des classes d'apprentissage. Les procédés d'apprentissage sont différents

suivant qu'il s'agisse de caractères imprimés ou manuscrits, et suivant qu'il s'agisse de reconnaître une ou plusieurs fontes.

Parmi les classifieurs on trouve : symbolic classifiers , probalistic classifiers, exemple-based classifiers, les réseaux de neurones (neural network), les SVM…etc. Récemment, on dénote une nouvelle tendance basée sur des classifieurs hybrides ou des comités de classifieurs [Sebastiani 02], [Feldman 07], [Farah 05], [Lorigo 06] Tableau 2-3.

Auteur	Moteur de reconnaissance
El-Hajj 2005	HMM
Mozaffari 2005	Le plus proche voisin
Farah 2004	Combinaison parallèle de trois classifieurs ANN, Kplus proche vosin , K-NN flou.
Haraty 2002-2004	Multiple ANN
Souici-Meslati 2004	ANN avec règles
Amin 2003	Règles d'apprentissage par programmation de la logique inductive.
Clocksin 2003	SVM
Snoussi Maddouri 2002	Réseau de neurones de quatre couches
Fahmy 2001	Réseau de neurones
Amin 1996	Réseau de neurones de cinq couches
Al-Yousefi 1992	Classifieur bayesien quadratique

Tableau 2-3 : Classifieurs utilisés par différents auteurs

2.6.5 La reconnaissance

Suivant la complexité des informations à reconnaître, une étape préalable de segmentation en entités de plus bas niveau peut être nécessaire notamment dans le cas de la reconnaissance de séquences numériques et de mots cursifs; elle consiste en premier lieu à isoler les zones d'intérêt, les lignes de texte, les mots dans les lignes; les caractères et les mots peuvent ensuite subir différentes normalisations pour faciliter l'étape suivante de reconnaissance [Vinciarelli 03].

On distingue deux grandes familles d'approches pour la reconnaissance de mots manuscrits isolés : La première, qui regroupe les approches dites analytiques, reconnaît les mots en s'appuyant sur la reconnaissance de ses lettres, elle s'appuie sur la segmentation des mots en lettres, étape extrêmement délicate. A l'inverse, la deuxième famille, qui regroupe l'ensemble des approches globales, considère le mot comme une entité indivisible. Il s'agit alors de décrire le mot par l'intermédiaire d'un jeu de caractéristiques regroupées dans un vecteur qui sera ensuite soumis à un classifieur. Ce dernier se prononcera alors en faveur de la meilleure hypothèse de reconnaissance mot. L'apprentissage du classifieur consiste à lui présenter suffisamment d'exemples de vecteurs de caractéristiques de chaque mot. C'est pourquoi ces approches sont dédiées à la reconnaissance dirigée par le lexique.

2.6.6 Le post-traitement

Une étape de post traitement peut être rajouté à un système de reconnaissance de l'écriture, qui à pour but d'améliorer le taux de la reconnaissance, en introduisant des informations contextuelles permettant de lever l'ambiguïté dans la reconnaissance de certains mots ou caractères, parmi ces informations en citant [Heutte 03] :

- Les connaissances pragmatiques sur la longueur moyenne de chacune des lettres, ou sur le nombre de lettres constituant un mot.
- Les algorithmes de correction orthographiques.
- Les connaissances linguistiques de différents niveaux :

  Lexical : pour valider la reconnaissance effectuée en ne retenant que des mots du dictionnaire, et en rejetant les listes de lettres inconsistantes.

  Syntaxique et sémantique : pour réduire la liste des mots candidates et valider ceux qui ont été retenus à l'étape précédente.

2.7 L'écriture arabe

2.7.1 Définition

La langue arabe est parlée par plus de 100 millions de personnes et utilisée dans plus de 22 pays [Burrow 04], [Miled 97]. Elle est également employée dans la plupart des écrits, à l'oral, dans les situations officielles ou formelles (discours religieux, politiques, journaux télévisés,…).

L'arabe littéral se distingue ainsi de l'arabe dialectal, qui est la langue vernaculaire parlée au quotidien et ce depuis l'expansion de l'islam. Cette variété de la langue recouvre plusieurs dialectes locaux pouvant varier assez fortement d'un pays à l'autre. Dans tous les pays arabes, un dialecte national composé par plusieurs dialectes locaux est parlé. Il existe 29 langues utilisant l'alphabet arabe : comme pour l'hébreu et le syriaque, l'espagnol ou le bosniaque [Humbert02].

2.7.2 L'alphabet

L'alphabet arabe comporte 28 lettres (tableau 2-4), pas de notion de majuscules ou de minuscules. L'écriture arabe est cursive que se soit en imprimé ou en manuscrite. Elle s'écrit de droite à gauche.

La forme des lettres dépend de leur position dans le mot. Certaines lettres prennent jusqu'à 4 formes différentes : par exemple le (ح ـحـ ـحـ حـ) ou le (ه ـهـ ـهـ ـه). 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. En arabe, les ligatures se situent toujours au niveau de la ligne d'écriture, c'est-à-dire qu'il n'existe pas de lettre à liaison haute comme le 'o' ou le 'v' en alphabet latin. En écriture arabe, il existe toutefois des ligatures verticales.

Certains caractères ont le même corps, mais la présence ou la position d'un point ou d'un groupe de points, est un trait déterminant pour distinguer ces caractères. La figure 2-5 montre

quelques caractères qui ont le même corps qui se différencient seulement par la présence et le nombre de points au-dessus ou en dessous de leurs corps [Burrow 04].

ض ص – ز ر – ب ت ث – ش س

Figure 2-5 : Caractères qui ont le même corps

N°	Lettre	Isolée	Forme		
			Initiale	Médiane	Finale
1	**Alif**	ا	ـا	ـا	ـا
2	**Baa**	ب	بـ	ـبـ	ـب
3	**Taa**	ت	تـ	ـتـ	ـت
4	**Thaa**	ث	ثـ	ـثـ	ـث
5	**Jeem**	ج	جـ	ـجـ	ـج
6	**Haa**	ح	حـ	ـحـ	ـح
7	**Khaa**	خ	خـ	ـخـ	ـخ
8	**Daal**	د	د	ـد	ـد
9	**Thaal**	ذ	ذ	ـذ	ـذ
10	**Raa**	ر	ر	ـر	ـر
11	**Zaay**	ز	ز	ـز	ـز
12	**Seen**	س	سـ	ـسـ	ـس
13	**Sheen**	ش	شـ	ـشـ	ـش
14	**Saad**	ص	صـ	ـصـ	ـص
15	**Shaad**	ض	ضـ	ـضـ	ـض
16	**Ttaa**	ط	طـ	ـطـ	ـط
17	**Dthaa**	ظ	ظـ	ـظـ	ـظ
18	**Ain**	ع	عـ	ـعـ	ـع
19	**Ghen**	غ	غـ	ـغـ	ـغ
20	**Faa**	ف	فـ	ـفـ	ـف
21	**Qaf**	ق	قـ	ـقـ	ـق
22	**Kaf**	ك	كـ	ـكـ	ـك
23	**Lam**	ل	لـ	ـلـ	ـل
24	**Mem**	م	مـ	ـمـ	ـم
25	**Noon**	ن	نـ	ـنـ	ـن
26	**Haa**	ه	هـ	ـهـ	ـه
27	**Wow**	و	و	ـو	ـو
28	**Yaa**	ي	يـ	ـيـ	ـي

Tableau 2-4 : Différents formes d'un caractère arabe

2.7.3 Signes diacritiques

Le signe diacritique est une composante secondaire d'une lettre, qui vient la compléter ou en modifier le sens. Elles désigneront à la fois points, voyelles et autres signes secondaires (chadda, madda, hamza, ...). Cependant, dans certains travaux, seules les voyelles arabes sont appelées diacritiques (figure 2-6) [Al-Shatnawi 08].

2.7.3.1 Points diacritiques

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. Un groupe de deux points peut ainsi s'écrire sous forme d'une seule, ou de deux composantes connexes. On remarque la très forte similarité entre deux points reliés par un trait. Un groupe de trois points peut donner lieu à une, deux ou trois composantes connexes, en fonction du style d'écriture [Jumari 02].

2.7.3.2 Les voyelles

En arabe, les voyelles ne sont pas des lettres, mais des signes diacritiques associés aux lettres sur lesquelles ils s'appliquent Les voyelles peuvent parfois être mentionnées sur certaines lettres pour lever l'ambiguïté et faciliter la lecture. Mais en général, les scripteurs les omettent purement et simplement, et c'est au lecteur qu'est réservé le soin d'interpréter correctement le sens de la phrase en fonction du contexte. En général on ne représente pas les voyelles, sauf dans les manuels scolaires. L'absence de voyelles peut toutefois être source de confusions. Un mot peut avoir plusieurs voyelles et par conséquent, plusieurs catégories grammaticales (figure 2-6).

2.7.3.3 Autres signes diacritiques

Les autres signes diacritiques sont la hamza, la chadda et la madda. La chadda est une accentuation de la lettre (c'est l'équivalent d'une consonne doublée). Hamza et madda suivent des contraintes morphosyntaxiques plus complexes (figure 2-6) [Jumari 02].

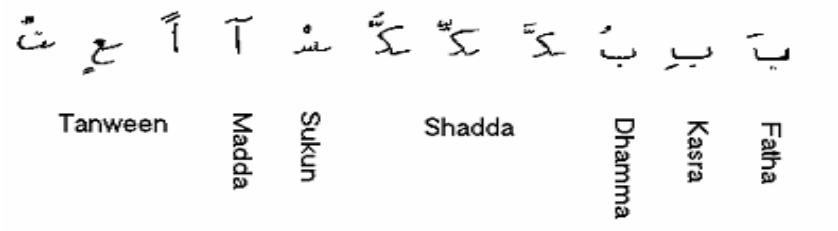


Figure 2-6 : Signes diacritiques dans l'écriture arabe

2.7.4 Ascendants et descendants

Comme dans l'écriture latine, l'écriture arabe contient des ascendants et des descendants. 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 (figure 2-7) [Al-Shatnawi 08].

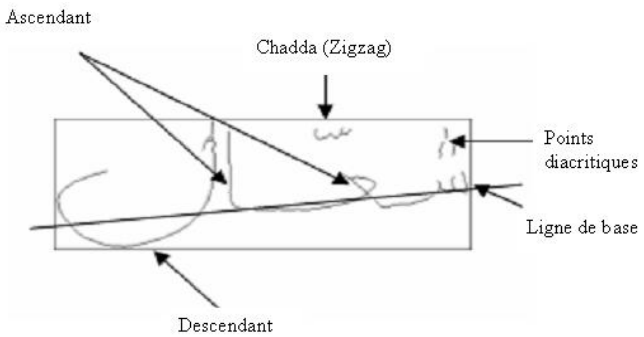


Figure 2-7 : Ascendants et descendants dans la langue arabe

2.7.5 Ligatures verticales

En écriture arabe, il n'y a pas de liaisons hautes comme le 'v' ou le 'o' en latin : les ligatures se situent au niveau de la ligne support de l'écriture (ligne de base). En revanche, les scripteurs sont libres de constituer certains groupes de deux ou trois lettres liées verticalement en début de pseudo-mot, très complexes à segmenter (figure 2-8) [Jumari 02]

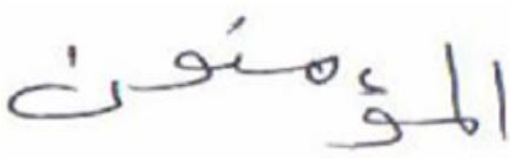


Figure 2-8 : Pseudo-mot dans l'écriture arabe

Comme l'écriture arabe est une écriture calligraphique, six grands styles graphiques différents sont utilisés : ***Thuluth, Naskh, Requeh, Dewan, Farci, et Coufique*** (figure 2-9) **[**Burrow 04**] [**Atanasiu03]**.**

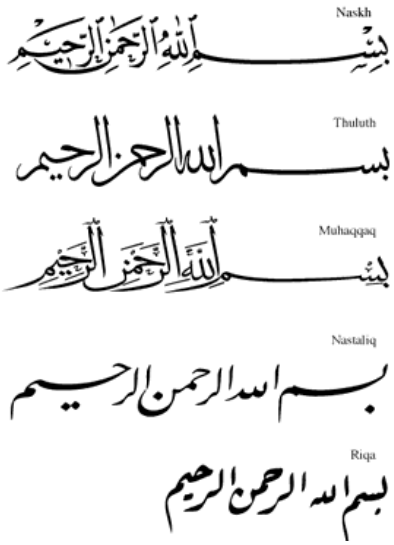


Figure 2-9 : Quelques styles calligraphiques en écriture arabe

2.7.6 Difficultés de la reconnaissance de l'écriture arabe

Il existe principalement deux types de difficultés dans le domaine de la reconnaissance de l'écriture arabe, le premier est lié au manque d'outils de test et de validation, alors que le second est lié à la morphologie de l'écriture arabe [Amin 98], [Lorigo 06].

Les problèmes d'infrastructures sont dus principalement au manque d'applications porteuses. En plus, les outils de test et de validation tels que les bases de données d'images servent à valider les résultats obtenus et à évaluer de manière unifiée les divers travaux réalisés dans des environnements différents. D'autres, tels que les lexiques et les dictionnaires de validation sont indispensables pour améliorer les performances des systèmes de reconnaissance, suite à leur absence, le post-traitement (lexical, syntaxique ou sémantique) est très peu utilisé dans les systèmes de reconnaissance de l'écriture arabe.

L'écriture arabe a un certain nombre de particularités telles que la forte dépendance de la calligraphie du caractère de son contexte, la complexité et la multiplicité des graphies des lettres, la variabilité des liaisons inter-caractères ou des ligatures horizontales et verticales ainsi que la présence de chevauchements. Ces particularités compliquent les tâches de choix des procédures de prétraitement, de la segmentation de textes arabes, de la sélection et de l'extraction des primitives (figure 2-10 à figure 2-19) [Atanasiu 03]**.**

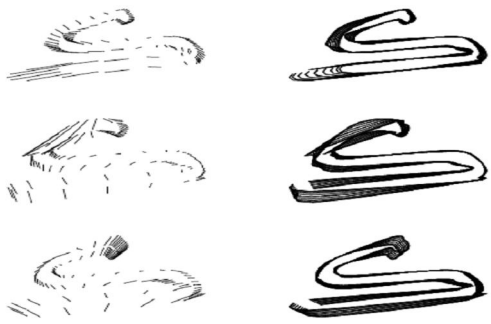


Figure 2-10 : Présentation des principales directions de variation du *kaf* (Ibn Wahid)

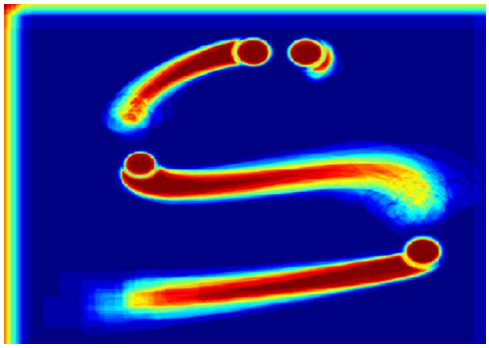


Figure 2-11 : Segmentation de la lettre *kaf* (Ibn Wahid)

Figure 2-12 : Calligraphie persane

Cette calligraphie contemporaine persane montre une utilisation du ya’ long inouïe pour l'époque médiévale.

Figure 2-13 : Les allographes ya’ (A) et ya’ long diagonal (B)

Utilisation par Ibn Wahid au cours de la même ligne de deux allographes pour la lettre ya’.

Figure 2-14 : Méthodes supplémentaires de justification

Ibn Wahid met à profit le crénage (haut), le rapetissement des caractères et l'incourbation de la ligne de base (bas) comme méthodes alternatives à l'utilisation des *kashida*-s.

Figure 2-15 : Idéaux et réalités graphiques

Pour tracer une descendante les traités proposent comme matrice le cercle, cependant aucun spécimen n'épouse à merveille cette forme géométrique parfaite. Les traités proposent aussi la forme de l'oeuf de poule, mieux adaptée aux réalités graphiques, mais certaines extrêmes s'éloignent même de ce modèle naturaliste.

Figure 2-16 : Fluctuation des formes des descendantes

L'appartenance des descendantes à une classe précise n'est pas toujours évidente : la forme peut changer pour la même lettre, pour des positions différentes de la même lettre dans le segment graphique, ou d'une lettre à l'autre.

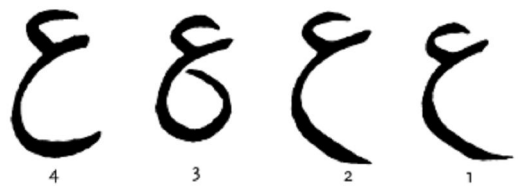


Figure 2-17 : Les allographes des descendantes

On distingue quatre allographes pour les descendantes :
(1) « *murassala* coulée » ; (2) «*musabbala* pendante », légèrement plus ouverte que la première ;
(3) « *makhlufa* retournée » et (4) « *makhlufa* retournée », variante de la précédente et a ne pas confondre avec la descendante fermée.

Figure 2-18 : Régularité et qualité calligraphique

Une écriture très irrégulière sera considérée comme une « note » (haut) ; une écriture moyennement régulière on verra la graphie d'un copiste (milieu) ; et une écriture très régulière sera vite appelée «calligraphie» (bas).

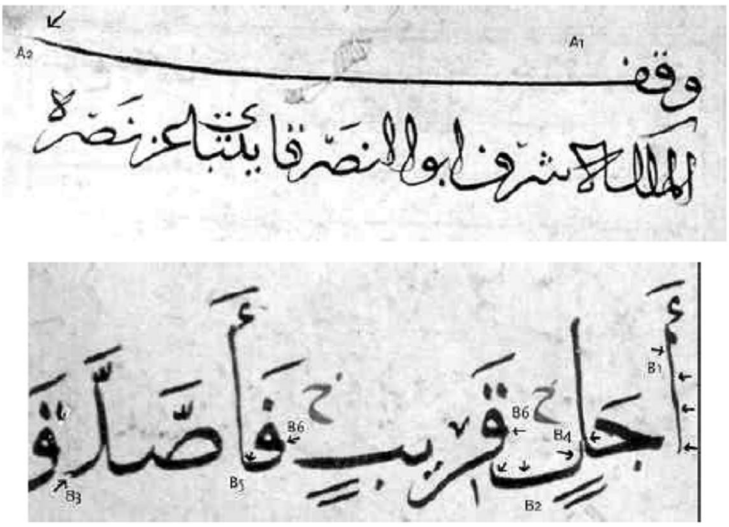


Figure 2-19 : Différence de qualité des traits graphiques

2.8 Conclusion

Malgré, les difficultés de l'écriture arabe, on peut avoir des points de similitude avec l'écriture latine, c'est-à-dire lors de la construction d'un système de reconnaissance des mots avec une base d’images de mots arabes : manuscrite ou imprimée, on suit les même étapes que celle utilisées pour le latin, sauf qu'il existe des différences capitales surtout au niveau de l'extraction des caractéristiques. Ce facteur variant et cet aspect d’hétérogénéité dans l’écriture manuscrite rend la phase de traitement et d’analyse plus fastidieuse. Nous allons décortiquer l’analyse des documents dans le chapitre suivant.

CHAPITRE 3

TRAITEMENT ET ANALYSE DES IMAGES

3 Traitement et Analyse des images

3.1 Introduction

Le traitement d'image est une étape importante avant l'étape de l'analyse des documents. Les traitements permettent de restaurer ou nettoyer l'image, l'analyse permet d'extraire les différentes structures du document (illustrations, éléments graphiques et textuels,...).

Le traitement d'image dans les documents anciens va donc permettre la préparation de ces documents, c'est-à-dire améliorer la qualité visuelle. Ensuite viens l'étape cruciale d'analyse pour rechercher les informations directement dans les images et d'en dégager la structure. Tout au long du processus de traitement, différentes images intermédiaires seront créées.

3.2 L'acquisition (numérisation)

3.2.1 Définition

La numérisation se définit comme la transformation sous forme numérique des informations analogiques. Ces informations analogiques se présentent en général sous forme de documents, en l'occurrence des textes ou des images, sur feuilles de papier ou autres supports analogiques, tandis que leurs équivalents numériques peuvent être stockés sur ordinateur ou sur d'autres supports informatiques tels que les disquettes, les CD-ROM, les mémoires flash, etc. L'appareil qui permet de réaliser cette numérisation est communément appelé un scanner.

Un scanner est équipé d'un capteur, lui-même composé de photodiodes, et a pour rôle principal de transformer la lumière en signal électrique, qui sera par la suite transféré à un processeur pour y être analysé. La lumière peut être réfléchie par les couleurs pour les documents opaques, mais elle peut aussi être directe, car elle traverse alors les documents transparents. Il

existe sur le marché différents types de scanners qui sont surtout à différencier par leurs caractéristiques techniques respectives [Agfa-Gevaert 94], [Mermet 08].

3.2.2 Son rôle

Le processus de la numérisation est primordial, car c'est a partir de cette étape que dépendra les autres étapes de prétraitement, segmentation et reconnaissance. Son rôle consiste à la duplication rapide et économique des ouvrages, la sauvegarde et préservation des documents originaux très empruntés et malmenés, facilite le partage des connaissances à un plus large publique par un accès multiple et efficace aux documents et ouvrages incommunicables du fait de leur rareté.

3.2.3 Les normes de numérisation

La numérisation dépend de l'outil de captation, du format de stockage et du mode de numérisation. Cet assortiment technologique va déterminer la qualité des images et influencer les traitements en aval.

3.2.3.1 Outils de numérisation

Le scanner est l'outil principal de la numérisation. En fait, elle est rendue possible grâce aux capteurs du scanner sensibles à la lumière rediffusée par les couleurs des documents.

Numériser ne signifie donc pas acquérir ou gérer des documents électroniques, mais consiste à transformer l'image papier ou tout autre type de support traditionnel de documents, en image électronique. Le document numérisé devient alors document électronique [Lesk 97].

C'est grâce aux différents capteurs du scanner que la numérisation est rendue possible, il existe plusieurs types (figure 3-1) [Haou 08]:

- Scanner à tambour : ces scanners à tambour recommandés pour la photogravure (également coûteux). La numérisation se fait point par point

sur tout le document dans la technique de numérisation en mode point. Les scanners à tambour utilisent cette technique qui permet d'avoir une très haute résolution, mais avec une faible vitesse de numérisation. Elle n'est pas non plus adaptée à tous les types de documents et est surtout utilisée dans les arts graphiques.

- Scanner à plat : L'un des outils indispensables à l'acquisition d'images numériques est le numériseur à plat. Le principe de fonctionnement de cet appareil ainsi que les différents réglages nécessaires à son emploi seront étudiés et expliqués de façon simple, en fonction des buts à atteindre. Des essais à partir de documents opaques et transparents, noir et blanc et couleur permettront de se familiariser avec cet outil aux applications nombreuses. La technique de numérisation linéaire fonctionne par balayage du document, lequel est analysé ligne par ligne par les photodiodes disposés en rangées. Cette technique est utilisée par les scanners à plat et offre une bonne qualité de numérisation. C'est la technique employée généralement par les scanners de bureau et les appareils photographiques numériques professionnels.
- Scanner à livre ouvert : correspondent bien aux besoins des bibliothèques car leur vaste surface de numérisation assure le traitement des grands formats. Lors du traitement, le livre est ouvert, texte dirigé vers le haut, le dispositif de numérisation se trouvant au-dessus. Parfois, un plateau ajustable compense la différence de hauteur une fois le livre ouvert.
- Scanner ou appareils photographiques numériques : Un autre outil pour l'acquisition d'images numériques est l'appareil photo numérique. Il permet en effet de "photographier" directement en numérique. Il ressemble à n'importe quel appareil photo traditionnel, depuis le compact jusqu'à la chambre professionnelle. Mais les fonctions offertes sont-elles identiques à celles d'un appareil argentique, les résultats sont-ils à la hauteur de nos

attentes, de nos besoins, son utilisation est-elle la même? Toutes ces questions seront traitées parallèlement à la prise en main et à l'utilisation d'appareils numériques. Le document à numériser est saisi en une seule fois dans la technique matricielle. Le capteur étant fixe, le temps d'exposition est devenu plus rapide, mais cette technique n'offre pas en général une très bonne résolution. C'est celle utilisée par les appareils photographiques numériques d'entrée de gamme.

- Scanner pour microformes : les microfiches, microfilms, diapositives, etc. sont numérisés par des machines spécifiques adaptées à chacun de ces supports et permettant une prise de vue automatique ou semi-automatique [Mermet 08].

Scanner de bureau

Scanner à plat

Scanner de microfiches

Copy book

Kirtas

Figure 3-1 : Exemples d'outils de numérisation

3.2.3.2 Formats de stockage

Le Formats de stockage est déterminant par l'usage que pourra être faits des types de documents. On choisi entre format image brute et image traitée en mode texte (texte linéaire et hypertexte) [BnF 09a].

Format image

Les fichiers numériques se distinguent par leur format que l'on identifie grâce à l'extension du nom du fichier. Il existe environ 70 formats de fichiers pour les images bitmap, mais il est déconseillé d'adopter des formats propriétaires de peur de ne pouvoir communiquer les fichiers en dehors d'un réseau restreint doté du même équipement.

Les formats le plus souvent retenus pour les fichiers images en noir et blanc et en couleurs sont :

TIFF (Tagged Image File Format), conçu par Aldus et Microsoft pour l'acquisition et la création d'images, est fréquemment proposé comme format par défaut dans des logiciels de numérisation. Ce format propriétaire est devenu un standard de fait. Il gère toutes les profondeurs de couleurs et intègre des informations de correction gamma. Il comporte de nombreuses variantes (les en-têtes de fichiers varient), il faut prendre garde que les visualiseurs et les logiciels de retouche ne puissent traiter la version choisie de ce format.

JFIF (JPEG File Interchange Format). C'est le format adapté aux images compressées en JPEG.

GIF (Graphics Interchange Format) l'un des plus courants pour les images. Cependant, il ne code pas plus de 256 couleurs par pixel, au-delà les images subissent une perte de qualité. GIF est très répandu sur l'internet.

PNG (Portable Network graphics, prononcé "ping"). Ce format récent améliore la vitesse et la qualité d'affichage et il est bien adapté à une diffusion sur le web. Il comporte également de nouvelles fonctions : la "signature électronique" inscrit dans le fichier le nom de l'auteur ou celui de l'oeuvre.

Les résultats obtenus avec le mode **JPEG** sont variables. C'est pourtant le mode de compression des images fixes le plus utilisé. La perte d'informations est paramétrable selon les accès prévus; elle n'est pas visible à l'œil nu (à l'impression) mais il sera impossible de restituer ce qui aura été perdu. Il est possible de choisir une perte minimale d'informations, avec une restitution, si le document s'y prête, à 80% et un taux de compression de 10.

JPEG2000 compression généralement utilisée pour des images entières à ton continu en couleur ou en échelle de gris, avec ou sans perte. Cet algorithme est plus avantageux que le JPEG car il procède par dégradation sélective de certaines zones moins stratégiques de l'image.

DjVu technique prometteuse qui code chaque élément séparément, un document constitué de texte et d'images est traité en deux parties. La comparaison avec les autres méthodes connues fait apparaître une perte d'informations inférieure et une qualité finale 5 à 10 fois supérieure.

Format texte

HTML c'est un format qui code le rendu à l'écran du document, et non pas sa structure logique en utilisant des balises de formatage pour une éventuelle lecture des documents sur internet sur différentes machines.

SGML repose sur la séparation complète entre la structure logique du document (balises) et sa mise en page qui dépend su support de présentation.

XML est une simplification du format SGML. Son but est de faciliter l'échange automatisé de contenus entre systèmes d'informations hétérogènes, on parle d'interopérabilité.

PDF est une évolution du format de description de page Postscript d'Adobe (utilisé par les imprimeurs) en permettant de représenter des documents en deux dimensions quels que soient le logiciel de traitement et le système d'exploitation (Windows, Unix, Linux, etc.) PDF permet la création de fichiers plus compacts qu'en Postscript, utilisant les normes de compressions habituelles (JPEG, LSW, CCITT) facilitant le transfert de gros fichiers sur les réseaux. Il peut comporter des liens hypertextes, des fonctions de recherche et de navigation, des formulaires. Il peut englober d'autres formats et supporte les métadonnées encapsulées.

De plus cette représentation en format texte des images de documents permet la génération d'index nécessaires à toute recherche d'information dans les documents mais qui demeure cependant très fastidieuse avant toute analyse du texte [BnF 09b].

3.2.3.3 Modes de numérisation

Apres l'acquisition du document, il est nécessaire de la sauvegarder pour pouvoir s'en servir postérieurement.

- Mode binaire (bitonal) : dans le cas ou les documents présentent des caractères bien séparés et de bonnes qualités, la numérisation est faite directement en binaire.
- Mode en niveau de gris : les images en niveau de gris permettent des traitements sophistiqués pour la segmentation des caractères dégradés, de séparer le texte des images, réparer et restaurer les images. Ce mode est conseillé pour les documents qui ne peuvent pas être binariser directement par le scanner.
- Mode en couleur : ce mode est utile pour la conservation en l'état des images en maintenant leurs couleurs d'origine. Toutefois, il est déconseillé en raison du poids énorme résultant de cette sauvegarde en couleur [Mermet 08].

3.3 Le prétraitement

Une fois le document numérisé (minimum 300 dpi[6]), il se présente sous la forme d'une image non structurée et uniquement visualisable. Tout le travail réside donc d'associer aux données "image" des données textuelles qui, par la suite, permettront d'effectuer des recherches informatisées.

La première étape du traitement comprend la correction de l'image en écartant les défauts. L'éclairage et le vieillissement (trous, taches d'humidité,...) du document, sont la source des imperfections lors de la numérisation du document.

3.3.1 Techniques de prétraitement

Consistent à améliorer la qualité des images en éliminant les défauts dus à l'éclairage et au processus d'acquisition. Elles sont différentes suivant que les caractères soient donnés sous forme d'image ou de suite de segments.

A cause de la forte granularité de l'échantillonnage et des divers problèmes d'éclairage et de saisie, l'image du caractère peut subir des manquements ou des empâtements. Il convient de corriger si possible ces problèmes avant toute étape d'analyse. Par ailleurs, il n'est pas toujours utile d'utiliser tous les points de l'image du caractère pour extraire les propriétés caractéristiques. Une étape de réduction élimine les points redondants [Likforman-Sulem 03]. Les techniques de prétraitement sont les suivantes :

[6] dot per Inch ou point par pouce : valeur numérique d'une image composée d'une juxtaposition d'éléments d'image (pixels) disposés en rangées et en colonnes.

3.3.2 La restauration

La restauration ou suppression du bruit des documents numérisés utilisent des opérations sur les images telles que *modifications d'histogrammes* (La modification d'un histogramme est généralement représentée sur une courbe (appelée courbe tonale) indiquant la modification globale des composantes de l'image avec en abscisse les valeurs initiales et en ordonnées les valeurs après modification.) et *filtrages* (passe-bas, passe-haut et morphologiques) pour corriger la luminosité, réduire les bruits, rehausser les contrastes [Young 98].

Pour la plupart des imperfections, il existe différentes méthodes associées afin de les corriger (tableau 3-1)

Défaut	***Prétraitement***
Luminosité trop faible/forte	Modification d'histogramme
Taches	Filtrages passe haut
Points parasites	Filtrages passe bas Filtrages morphologiques
Rotation légère de l'image	Redressement par re-échantillonnage
Courbure de l'écriture	Re-échantillonnage
Ecriture fragmentée	Filtrages (passe haut/bas, morphologiques)
Contours de l'écriture flous	Filtrages passe haut Filtrages morphologiques
Ecriture du verso apparaissant sur le recto	Combinaison des images recto et verso

Tableau 3-1 : Exemples de méthodes de prétraitement utilisées

Ces traitements doivent être utilisés consciencieusement, car certains filtres peuvent avoir des effets néfastes sur d'autres éléments de l'image [Likforman-Sulem 03] [Likforman-Sulem 06].

3.3.3 La correction géométrique

Les normalisations ont pour objectif de rendre l'écriture la plus indépendante possible du scripteur. Trois techniques de correction géométrique sont présentées [Belaid 06] :

La correction de l'inclinaison des lignes (skew)

La correction de l'inclinaison des lignes de texte, consiste à redresser horizontalement les lignes d'écriture obliques. Plusieurs méthodes sont disponibles. Les deux plus populaires sont la transformée de Hough (appliquée sur les centres de gravité des composantes connexes), et les histogrammes de projection (figure 3-2).

La correction de l'inclinaison des lettres (slant)

Certains scripteurs écrivent leurs lettres de façon inclinée par rapport à l'axe vertical. Les lettres peuvent être inclinées vers la droite ou vers la gauche. Pour la même raison que dans le paragraphe précèdent, il convient de corriger cette inclinaison de l'écriture pour la rendre la plus indépendante possible des spécificités d'écriture du scripteur. Plusieurs approches ont été proposées comme les histogrammes de projection, de la même manière que pour la correction du skew, mais cette fois dans des directions proches de la verticale.

La normalisation en hauteur

La taille d'un caractère peut varier d'une écriture à l'autre, ce qui peut causer une instabilité des paramètres. Une technique naturelle de prétraitement consiste à ramener les caractères à la même taille. Afin de rendre la suite des traitements insensible à la taille des caractères, une étape de normalisation de la taille des caractères est parfois effectuée.

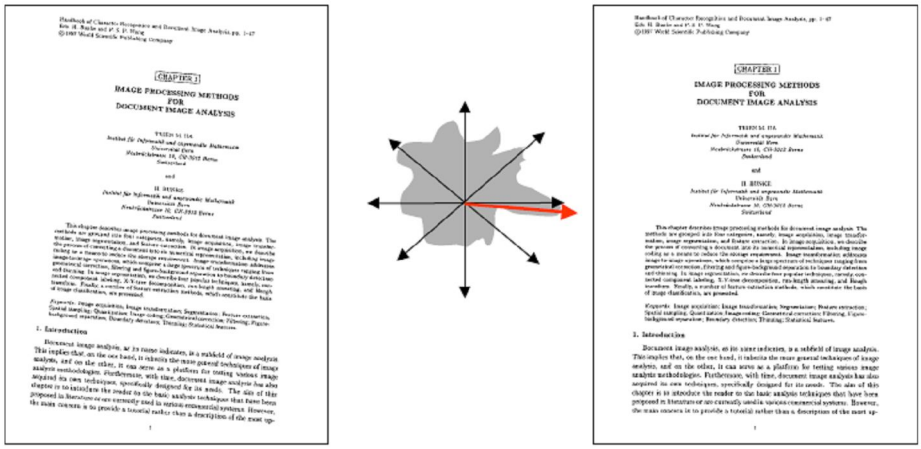


Figure 3-2 : Correction de l'inclinaison

3.3.4 Le seuillage (binarisation)

L'étape suivante consiste à séparer les textes et éléments graphiques du fond de l'image. Elle est appelée *seuillage* ou *binarisation* (figure 3-3).

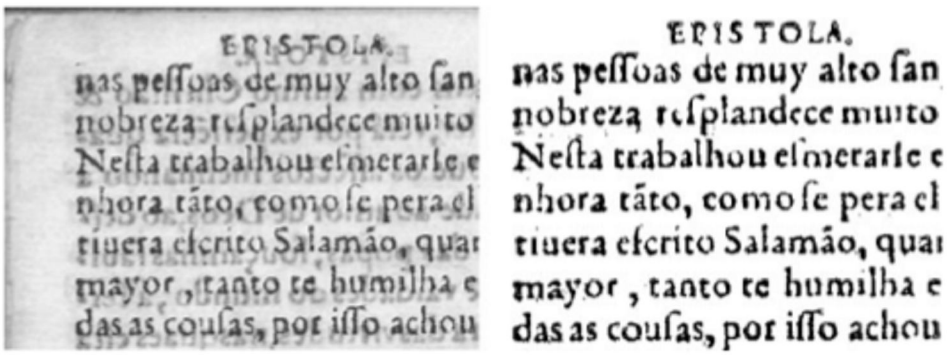


Figure 3-3 : Image avant et après traitement de binarisation

Elle consiste à obtenir une image bitonale (noir pour l'écriture, blanc pour le fond). Le choix d'un algorithme (global ou adaptatif) recherchant un seuil de binarisation doit être effectué et réfléchi avec le plus grand soin pour que le blanc soit attribué au seul fond de l'image et que l'écriture apparaisse bien en noir. C'est dans cet intérêt qu'il est parfois judicieux de régler ce

seuil manuellement au vu de l'histogramme pour de meilleure résultat (figure 3-4) [Boulehmi 08] [Likforman-Sulem 03].

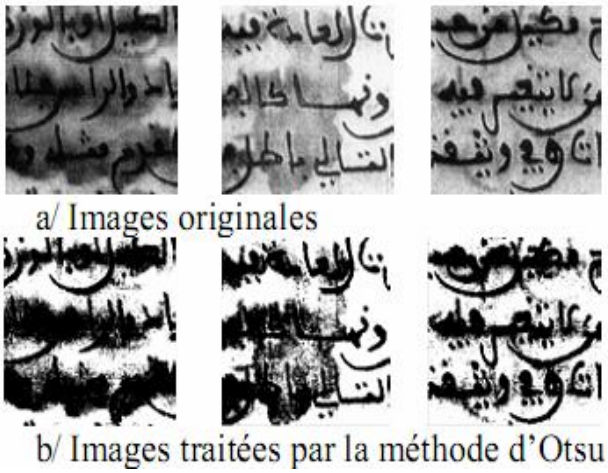


Figure 3-4 : Exemples de binarisation par la méthode Otsu

3.3.5 La compression

La compression de l'image consiste à réduire la dimension du fichier de données avec ou sans perte de l'information. Vu l'immense quantité d'archive de document anciens, il serait très intéressant de sauvegarder toute cette collections dans des formats de compression acceptable (sans perte) pour diminuer l'espace de stockage et permettre un traitement convenable de ces fichiers compressés sans altération sur le contenu. De nouvelles approches de la compression sont proposées est qui sont adaptées à la spécificité des documents patrimoniaux [Nicolas 06].

3.3.6 Amincissement (squelettisation)

La squelettisation est une opération qui permet de passer d'une image à sa représentation en *fil de fer*. Le squelette a un pixel d'épaisseur. C'est une manière de représenter l'information indépendamment de l'épaisseur initiale de l'écriture. Il permet d'extraire des caractéristiques importantes, comme les intersections et le nombre de tracés, leurs positions relatives. Il est

également possible de renormaliser l'épaisseur de l'écriture à partir du squelette. Dans l'approche squelette, nous pouvons noter deux problèmes particuliers: l'apparition de barbules, et le comportement au niveau des intersections [Lorigo 06].

3.3.7 Contours

Le but de la détection de contours est de repérer les points d'une image numérique qui correspondent à un changement brutal de l'intensité lumineuse afin de mettre en évidence la limite, le voisinage et la discontinuité. La détection des contours d'une image réduit de manière significative la quantité de données et élimine les informations qu'on peut juger moins pertinentes, tout en préservant les propriétés structurelles importantes de l'image. D'un point de vue fonctionnel, elle est considérée comme une opération économique permettant d'alléger considérablement le processus de reconnaissance d'objet [Young 98].

3.3.8 Lissage

Les prétraitements peuvent introduire des bruits dans l'image, qui se traduisent en particulier par la présence d'irrégularités le long des contours des lettres. Ces bruits peuvent dégrader les performances de reconnaissance. Des techniques simples et rapides à mettre en oeuvre permettent de diminuer ces bruits comme le lissage (*débruitage* ou *filtre anti-bruit*). On appelle aussi anti-crénelage (*anti-aliasing*) l'opération de lissage spécifique consistant à atténuer l'effet d'escalier produit par les pixels en bordure d'une forme géométrique [Young 98].

Filtrage passe bas : Ce type de filtrage est généralement utilisé pour atténuer le bruit de l'image, c'est la raison pour laquelle on parle habituellement de lissage.

Filtrage passe haut : A l'inverse du filtre passe-bas, il permet notamment d'accentuer les détails et le contraste. Il simplifie l'image en préservant la structure, mais il perd en général de l'information.

On parle alors d'ouverture ou de fermeture :

- L'ouverture élimine les petites composantes, et ouvre les petits isthmes.
- La fermeture bouche les petits trous, et ferme les petits détroits.

3.4 Analyse d'images de documents manuscrits

Dans le cas des documents manuscrits, l'objectif visé par l'interprétation du document est généralement la reconnaissance des mots ou des lignes de texte. Les documents manuscrits sont en effet des documents qui sont majoritairement textuels, et dont la structure est en général relativement simple contrairement aux documents imprimés. Cependant cette structure est caractérisée par une très forte variabilité spatiale qui se traduit par exemple par des lignes de texte inclinées et fluctuantes, des chevauchements entre les lignes, des espaces irréguliers entre les mots.

On considère plutôt une phase de segmentation et une phase de reconnaissance de l'écriture. En effet on est capable aujourd'hui de reconnaître des mots manuscrits isolés avec des taux de reconnaissance corrects, surtout avec un lexique réduit. L'étape de segmentation peut s'intégrer plutôt dans une chaîne de traitements.

3.4.1 La segmentation

Une fois la binarisation effectuée (séparation des formes du fond), on passe à la segmentation. Cette manipulation opère sur les pixels de l'image, elle est donc de bas niveau. Il s'agit tout d'abord de classer les éléments extraits du fond en entités similaires et de distinguer les éléments textuels (caractères, symboles) des éléments graphiques (paraphes, ratures, lettrines, illustrations,...). Après ces différentes opérations effectuées, l'image intermédiaire obtenue est propre, débarrassée d'éléments non textuels et l'écriture est nette (ni fragmentée, ni épaissie) [Boussellaa 06].

3.4.2 Méthodes d'analyse

Trois stratégies différentes d'analyse existes, à savoir la stratégie *ascendante* procédant par agglomération d'entités voisines, la stratégie *descendante* agissant par découpe hiérarchique et la stratégie *mixte* qui est la combinaison des deux stratégies précédentes [Belaid 06], [Likforman-Sulem 06].

- Les stratégies ascendantes ou guidées par les données : se basent sur l'analyse des données pour remonter jusqu'à la structure logique.
- Les stratégies descendantes ou guidées par les modèles : qui exploitent la connaissance a priori de modèles de structuration pour aller rechercher l'information dans l'image du document.
- Les stratégies mixtes : reposent à la fois sur l'analyse des données et sur l'exploitation de modèles.

Stratégie ascendante (RSLA)

L'algorithme RLSA de Wong (1982) est l'un des algorithmes de segmentation physique les plus populaires et les plus anciens, repose sur une stratégie ascendante. Il s'agit d'une méthode itérative basée sur des opérations morphologiques de traitement d'image, qui permet de segmenter des images de documents binaires en blocs, en fusionnant en une seule composante connexe, des composantes de l'image suffisamment proches, selon un seuil fixé. Le principe du RLSA est qu'un lissage est appliqué horizontalement et verticalement sur l'image, produisant deux images. Un « et » logique est appliqué sur ces deux images produisant une image lissée ou image des composantes connexes [Belaid 06] (figure 3-5).

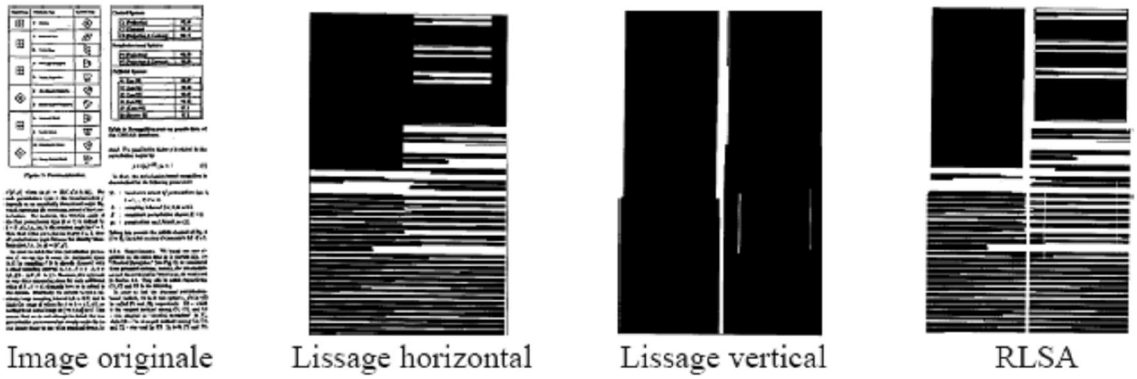


Figure 3-5 : Fonctionnement du RLSA

Stratégie descendante (X-Y Cut)

Un autre algorithme de segmentation très connu et très utilisé est l'algorithme X-Y Cut de Nagy (1984). Cet algorithme s'applique également sur des images binaires. Cependant il s'agit cette fois d'un algorithme descendant qui permet de découper l'image récursivement en zones homogènes de plus en plus petites par analyse des profils de projection horizontaux et verticaux des pixels noirs de l'image [Belaid 06] (figure 3-6).

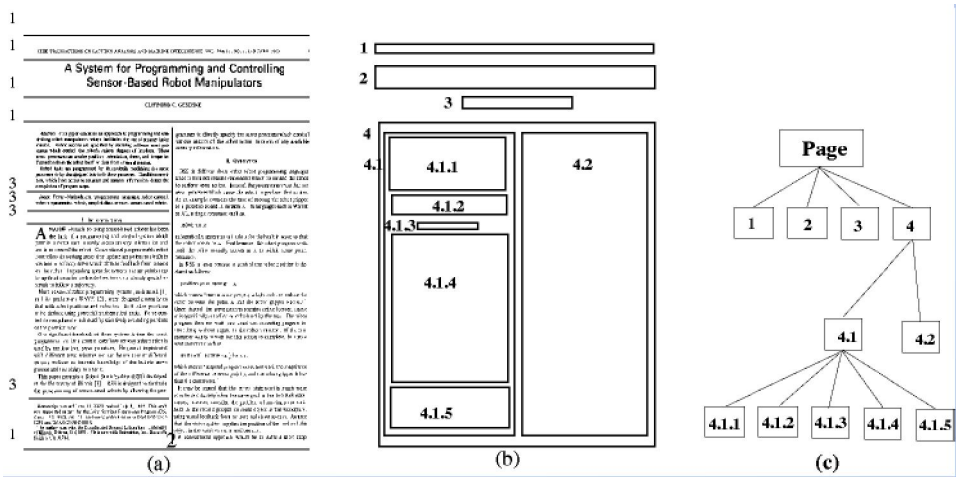


Figure 3-6 : Fonctionnement du X-Y Cut

3.4.3 Segmentation texte/graphique

Les traitements se font sur des images à bas niveau (pixels). La tâche consiste à classer les éléments (formes) en entités similaires en distinguant les éléments textuels des éléments graphiques. On parle alors de segmentation (séparation) texte/graphique [Tan 01], [Yen 04], [Boussellaa 06]. Une analyse en similarité ou en composantes connexes peut aider à cette segmentation en recherchant les ensembles de pixels noirs homogènes ou connectés. Une sélection de type morphologique basée sur la taille ou l'aspect des composantes permet d'éliminer un certain nombre de composantes non textuelles [Likforman-Sulem 03], [Khurshid 08].

3.4.3.1 Segmentation par frontière

Cette segmentation s'intéresse aux contours des objets dans l'image. La plupart de ces algorithmes sont locaux, c'est à dire fonctionnent au niveau du pixel. Des filtres détecteurs de contours sont appliqués à l'image. Le résultat est en général difficile à exploiter sauf pour des images très contrastées. Les contours extraits sont la plupart du temps morcelés et peu précis et il faut utiliser des techniques de reconstruction de contours par interpolation ou connaître a priori la forme de l'objet recherché. Formellement, ce type d'algorithme est proche des techniques d'accroissement de région fonctionnant au niveau du pixel [Baillie 03].

3.4.3.2 Segmentation par classification de pixel

Cette segmentation travaille sur des histogrammes de l'image. Par seuillage, clustering ou clustering flou, l'algorithme construit des classes de couleurs qui sont ensuite projetées sur l'image. La segmentation est implicite puisqu'on suppose que chaque cluster de l'histogramme correspond à une région dans l'image. En pratique, ce n'est pas forcément le cas et il faut séparer les régions de l'image qui sont disjointes bien qu'appartenant au même cluster de couleur. Ces algorithmes sont assez proches des algorithmes de réduction de couleur [Haou 08].

3.4.3.3 Segmentation par région

Dans la segmentation par région on trouve les techniques de decomposition et fusion (split and merge), croissance de régions (growing region), Ligne de partage des eaux (watershed), …etc [Baillie 03], [Bloch 06].

3.4.3.3.1 La segmentation par croissance de régions

Elle correspond aux algorithmes d'accroissement ou de découpage de région. L'accroissement de région est une méthode bottom-up : on part d'un ensemble de petites régions uniformes dans l'image (de la taille d'un ou de quelques pixels) et on regroupe les régions adjacentes de même couleur jusqu'à ce qu'aucun regroupement ne soit plus possible. Le découpage de région est le pendant top-down des méthodes d'accroissement : on part de l'image entière que l'on va subdiviser récursivement en plus petites régions tant que ces régions ne seront pas suffisamment homogènes.

La segmentation par croissance de régions à pour but de décomposer une image en régions homogènes. Différentes méthodes sont employées, parmi elles la croissance de régions par agrégation de pixels. Cette méthode consiste à regrouper les pixels vérifiant un critère d'homogénéité. Ce critère d'homogénéité peut être de différentes natures, le plus simple étant la comparaison des niveaux de gris des pixels selon un seuil.

On définit un germe comme étant le premier pixel que l'on étiquette par un numéro de région suivant son niveau de gris. L'opération de croissance de régions consiste ensuite à regrouper les pixels voisins à ce germe en fonction du critère d'homogénéité. L'exemple suivant illustre l'application de la croissance de régions à une matrice de données image. Dans cet exemple, les germes ont été trouvés par balayage des lignes de gauche à droite et des colonnes de haut en bas.

100	100	50	50	50	50	1	1	2	2	2	2
100	100	50	50	40	40	1	1	2	2	3	3
100	100	50	40	40	40	1	1	2	3	3	3
100	100	40	40	40	40	1	1	3	3	3	3
100	40	40	40	40	40	1	3	3	3	3	3
40	40	40	40	40	40	3	3	3	3	3	3

Image en 256 niveaux de gris et les 3 germes

Résultat de la croissance de région pour un seuil de 5 niveaux de gris

La croissance de régions permet d'analyser la structure de l'image en fournissant des informations sur la topologie des régions de l'image. L'analyse de ces frontières permet d'extraire les relations de voisinage entre régions [Bloch 06].

3.4.3.3.2 Décomposition/Fusion (Split/Merge)

Cette technique enchaîne les deux phases suivantes :

1. Découper itérativement l'image jusqu'à avoir des blocs contenant exclusivement des pixels similaires.
2. Regrouper les blocs voisins s'ils sont similaires.

Les deux phases sont nécessaires afin de garantir que les régions obtenues sont à la fois homogènes et également les plus grandes possibles. Chaque phase étant indépendante de l'autre.

3.4.3.3.3 Ligne de partage des eaux (watershed)

La technique du watershed, contrairement à la méthode de faire progressivement accroître les régions autour de leur point de départ (growing-region). Elle consiste à faire croître simultanément toutes les régions jusqu'à ce que l'image soit entièrement segmentée. Cette méthode tire son nom d'une analogie avec la géophysique. On peut alors considérer les valeurs d'intensité des pixels d'une image en niveau de gris comme une information d'altitude ou un relief topographique. Cette image appelée carte d'élévation est représentée comme un terrain en 3 dimensions. Le principe est de remplir progressivement d'eau chaque bassin du terrain, chaque bassin représente une région et lorsque l'eau monte et que deux bassins se rejoignent, la ligne de rencontre est marquée comme une ligne de frontière entre les deux régions.

D'autres techniques de segmentation plus complexes permettent d'obtenir de meilleurs résultats, notamment :

- La décomposition adaptative (Adaptive Split)
- La recherche de formes géométriques (Model based segmentation)
- L'approche probabiliste (relaxation)
- Le regroupement (k-mean, mean shift)

3.4.4 Segmentation en lignes

Cette étape permettra de trouver les mots à l'intérieur d'images. La procédure se fait en éliminant préalablement les composantes graphiques proches du texte, voire même superposées. Le tableau 3-2 donne quelques exemples de segmentation en lignes par différents auteurs [Likforman-Sulem 06].

Pour ce faire, il est possible de reprendre le concept des composantes connexes ou alors la projection des pixels. En s'aidant des composantes connexes et une direction haut/bas, ces dernières vont progressivement évoluer de caractères en mots, de mots en lignes et de lignes en

paragraphes. Au final, cette manipulation nous permet d'obtenir une approche de la structure physique du document.

Encore une fois, il est utile de rendre sensible sur le fait que la structure des documents anciens est bien plus complexe que les documents actuels et qu'il est très difficile de déterminer leurs structures physiques.

3.4.4.1 Méthode à histogramme

Cette méthode suppose que les lignes de textes soient droites, ou très peu inclinées, et relativement espacées de manière à ce que les hampes et les jambages des lettres se chevauchent au minimum. L'histogramme ainsi obtenu, il est alors possible par extraction des extrema locaux de localiser les lignes de texte du document [Likforman-Sulem 06]. Cette méthode a l'avantage d'être peu coûteuse en temps de calcul et simple dans son principe. Cependant, les résultats obtenus ne sont exploitables que si le document est propre, bruit filtré, inclinaison corrigée, lignes de texte droites et espacées (figure 3-7).

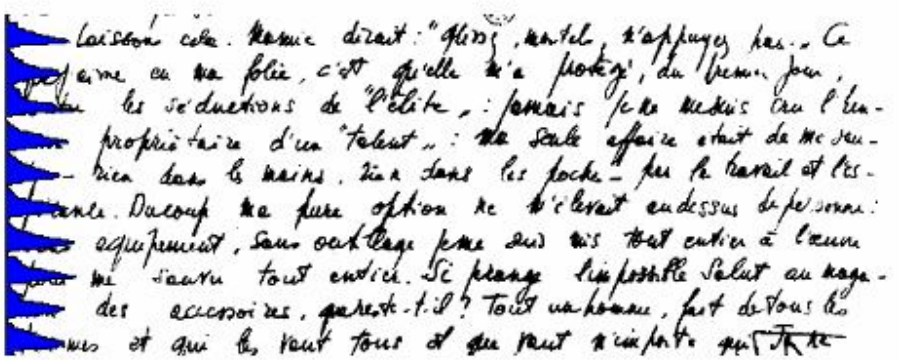


Figure 3-7 : Histogramme de projection sur un texte de Jean-Paul Sartre

3.4.4.2 Méthode par transformée de Hough

La transformée de Hough fait correspondre à une droite dans l'espace de départ (l'image) un unique point dans l'espace d'arrivée (l'espace de Hough). Une méthode de segmentation basée sur cette transformée est présentée dans [Likorman-Sulem 03] (figure 3-8). La première étape est l'extraction des composantes connexes du document. Ensuite, la transformée de Hough est appliquée. Pour chaque droite, le nombre de composantes en intersection avec la droite est mémorisée dans l'espace de Hough. Seules les droites ayant un maximum d'intersections sont retenues comme lignes potentielles. Cependant, certaines droites indésirables peuvent également être détectées (quelques droites verticales dans un document dont les lignes de texte sont principalement horizontales). Les auteurs ont donc ajouté une étape de validation permettant de ne conserver que les lignes de texte véritables.

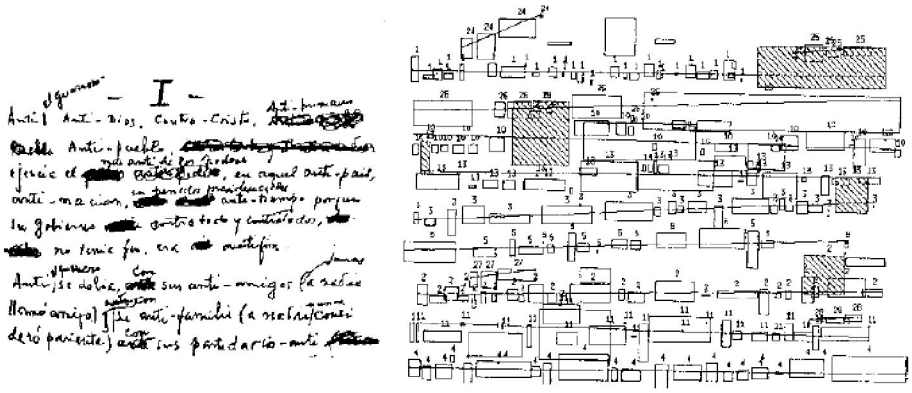


Figure 3-8 : Exemple de segmentation par transformée de Hough

Les résultats obtenus avec ces méthodes sont corrects tant que les lignes de texte du document restent droites et sans fluctuation de la ligne de base. En effet, la transformée de Hough travaillant avec des droites, il est impossible de l'utiliser pour des documents contenant des lignes incurvées ou dont l'orientation varie. Il faut également remarquer que la transformée de Hough est particulièrement coûteuse en temps de calcul [Chatelain 06].

3.4.4.3 Méthode à ombrage (Shading)

Cette méthode est en fait une amélioration de la méthode à histogramme. Au lieu de réaliser l'histogramme de l'image entière, l'image est subdivisée en bandes verticales de taille fixe, puis l'histogramme de chacune des bandes est calculé. Pour chaque bande, on détermine par extraction d'extrema locaux ce que l'auteur appelle des blocs. Une fois l'ensemble des bandes de l'image traitées, il faut reconstruire les lignes à partir des blocs extraits. Pour cela, l'auteur propose d'assigner à un même alignement les blocs des bandes voisines qui présentent un recouvrement vertical. Mais du fait de l'inclinaison du texte, il se peut que certains blocs aient une position ambiguë : ils peuvent être affectés à deux lignes. Les auteurs résolvent ce problème dans certains cas, mais sans en expliquer le principe. Il faut noter également que le choix de la largeur des bandes verticales est important : des bandes trop larges ne présenteraient aucun avantage par rapport à une méthode par histogramme classique, mais une largeur 88 Reconnaissance de documents manuscrits de bande trop faible peut être tout aussi dangereuse : il est possible que certaines bandes ne contiennent que l'espace entre deux mots d'une même ligne. Dans ce cas, la ligne est séparée en deux car l'histogramme des bandes verticales présente un blanc. Les auteurs proposent une largeur de bande de 100 pixels pour des images scannées à une résolution de 300 dpi, sans expliquer ce choix.

D'autres techniques de segmentation en lignes existent comme la méthode par projection partielle, la méthode de suivi de contour partiel et la méthode à regroupement des composantes connexes…,[Likforman-Sulem 06], [Zaidi 08].

Auteurs	***Description***	***Description de la Ligne***	***Type d'écriture***	***Unités***	***Project/ Documents***
Zahour *et al.*, 2004	k-means clusters	Tracé linéaire	Manuscrit Arabe	Blocs de texte	Documents arabes anciens
Antonacopoulos and Karatzas, 2004	Projection de profile	Tracé linéaire	Imprimé Latin	Pixels	Memorial/écrits personnels (2eme guerre mondiale)

He and Downton, 2003	Projection (RXY cuts)	Tracé linéaire	Imprimé et manuscrit Latin	Pixels	Viadocs/ cartes d'histoire naturelle
Feldbach and T?nnies, 2001	Methode de regroupement	Ligne de base	Ecriture cursive	Points min	Registres d'églises (18, 19siècle)
Lebourgeois et al., 2001	Accumulation du gradient	clusters	Latin imprimé	Pixels	Debora/livres (16siècles)
Tseng and Lee, 1999	Stochastique (algorithme de Viterbi)	Tracé non linéaire	Manuscrit Chinois	Pixels	Documents manuscrits
Calabretto and Bozzi, 1998	Projection de Profile (image en niveau de gris)	Tracé linéaire	Ecriture cursive	Pixels	Bambi/manuscrits italiens (16siècle)
Likforman-Sulem *et al.*, 1995	Regroupement	Chaîne de caractère	Manuscrit Latin	Composantes connexes	Philectre
Likforman-Sulem and Faure, 1994	Regroupement	Chaîne de caractère	Manuscrit Latin	Composantes connexes	Philectre

Tableau 3-2 : Différentes techniques de segmentation en lignes

3.4.5 Segmentation en mots

Dans le cas de l'écriture cursive, le problème est encore plus complexe. Dans la communauté de la reconnaissance de l'écriture manuscrite, il est admis qu'il est impossible de segmenter directement un mot cursif en lettres. Une solution est de découper le mot en sous-parties de lettres. Pour la segmentation de l'écriture cursive. On distinguera deux approches [Lorigo 06], [Heutte 03]:

La segmentation en graphèmes, appelée également segmentation explicite, le mot est segmenté en sous-parties qui sont presque des lettres. L'analyse par fenêtres glissantes, appelée également segmentation implicite, le mot est découpé en bandes verticales.

Parmi les segmentations mots nous avons la segmentation à partir du squelette, du contour, des histogrammes, segmentation basée sur des réservoirs ou à partir de fenêtres glissantes [Likforman-Sulem 03], [Chatelain 06].

3.4.5.1 Segmentation à partir du squelette

A partir du squelette, on cherche à repérer certains motifs, pour en déduire les candidats de points de coupures. La détection de ces motifs introduit des calculs de courbures et d'angles, qui sont comparés à des seuils ajustés de manière à obtenir le résultat désiré. Cette approche est erronée dans environ 10% des cas. Les configurations difficiles à segmenter sont celles pour lesquelles les lettres sont souvent enchevêtrées, comme les "tt", ou les lettres à liaison haute ('b', 'o', 'v', 'w') avec leur successeur.

3.4.5.2 Segmentation à partir du contour

La segmentation appliquée aux contours consiste a déterminer les meilleurs points candidats de coupure entre graphèmes, en s'appuyant sur les extrema locaux du contour, qui sont associés selon un critère de proximité. La segmentation en graphèmes à partir du contour nécessite de nombreux ajustements avant de trouver les critères de décision. Cette mise au point par tâtonnements est le point commun de nombreux traitements d'images liés à la reconnaissance de l'écriture manuscrite. Faciles à ajuster lorsque la qualité de l'écriture est bonne, ces prétraitements peuvent avoir des comportements tout à fait irréguliers lorsque l'écriture est de mauvaise qualité.

3.4.5.3 Segmentation à partir des histogrammes

Cette méthode simple consiste à calculer des histogrammes de projection dans plusieurs directions proches de la verticale. Les droites choisies sont celles qui interceptent le moins de pixels noirs, avec une contrainte d'espacement régulier dans l'image. Cette méthode montre néanmoins ses limites lorsque les lettres sont très proches ou enchevêtrées (figure 3-9) [Maddouri 08].

Figure 3-9 : Histogramme de projection sur des mots en arabe

3.4.5.4 Segmentation basée sur des réservoirs

X.Dupré étend à l'écriture cursive la technique à base de réservoirs initialement appliquée à la segmentation de chiffres liés. Il souligne que les règles de décision sont plus difficiles à mettre en place dans le cas des lettres, car ces dernières sont de tailles variables.

3.4.5.5 Fenêtres glissantes

L'utilisation d'une fenêtre glissante revient à découper l'image en bandes verticales. Ce découpage peut être régulier ou non, éventuellement avec recouvrement partiel des bandes successives. Cette technique présente l'avantage d'être simple, robuste au bruit, et est indépendante de la connexité. Le défaut de cette méthode est que la séquence générée contient beaucoup de bruit (recouvrement de deux lettres successives). C'est également vrai dans le cas des lettres superposées verticalement, mais qui ne se touchent pas nécessairement.

3.5 Conclusion

De nombreux progrès ont été réalisés ces dernières années dans le domaine du traitement et de l'analyse d'images de documents, que ce soit pour la reconnaissance de documents imprimés ou la reconnaissance de l'écriture manuscrite.

Dans le traitement de l'image des documents numérisés, plusieurs techniques de numérisation et de prétraitements ont été présentées. Lorsque les documents sont bruités ou dégradés, l'analyse de document devient plus délicate. Il est donc essentiel de bien mener cette phase de traitement d'image pour mieux appréhender les difficultés rencontrées lors des processus de segmentation.

La difficulté de l'analyse d'images de documents est liée à la variabilité dans les contenus, à la variabilité dans les structures, et à la grande hétérogénéité dans les types de documents. En fonction de ces différents critères, les structures physiques et logiques sont plus ou moins identifiables, et plus ou moins dépendantes. Les stratégies d'analyse à mettre en place dépendent généralement de ces facteurs. Ainsi les problèmes rencontrés ne sont pas les mêmes pour les documents imprimés ou pour les documents manuscrits.

Après avoir exposé dans les chapitres précédents les structures des documents, le processus de reconnaissance de l'écriture manuscrite et finalement, le traitement et l'analyse des images, nous allons aborder la phase de catégorisation dans le chapitre qui suit.

CHAPITRE 4

CATEGORISATION PAR APPARIEMENT APPROXIMATIF DES CHAINES DE CARACTERES

4 Catégorisation par appariement approximatif des chaînes de caractères

4.1 Introduction

Depuis l'étendue explosion du nombre de documents dans le monde, un besoin s'est ressenti pour ordonner et organiser toute cette masse d'informations. Une solution pour répondre à cette problématique est la catégorisation automatique de textes. Considérée comme une dérivée de la recherche d'information (RI). Plusieurs recherches dans ce domaine ont été entreprises mais dans des secteurs restreints avec beaucoup de contraintes.

4.2 Système de catégorisation de texte classique

Un grand nombre de recherches sur la catégorisation de texte a été entrepris depuis 1960 sur les documents anglais, italiens, français, chinois et autres… ; Cependant pour l'arabe, le sujet a été très peu abordé. Avec les progrès technologiques (scanners, duplicateurs, supports de stockage,…), le volume de documents numériques n'a cessé de croître jusqu'à en rendre impossible une classification manuelle d'où le besoin d'une tâche de classification automatique. La catégorisation automatique de textes (Automated Text Categorization) [Aas 99], [Sebastian 02], consiste à classer de manière automatisée des documents suivant certains critères en thèmes prédéfinis, elle est reconnue comme un processus d'organisation supervisé dans le cadre duquel une ou plusieurs catégories sont affectées au document à catégoriser [Jaillet 03].

L'objectif d'un système de catégorisation est d'approximer la fonction de catégorisation exacte qui associe à chaque couple de (document/classe) une valeur (vraie/fausse), en fonction de l'appartenance ou non du document à la classe :

$$F : D \times C \rightarrow \{T, F\}$$

Avec D l'ensemble des documents et C l'ensemble des catégories.

On nommera F l'approximation d'une telle fonction dont l'objectif repose essentiellement sur une notion de similarité entre documents et classes [Sebastiani 02].

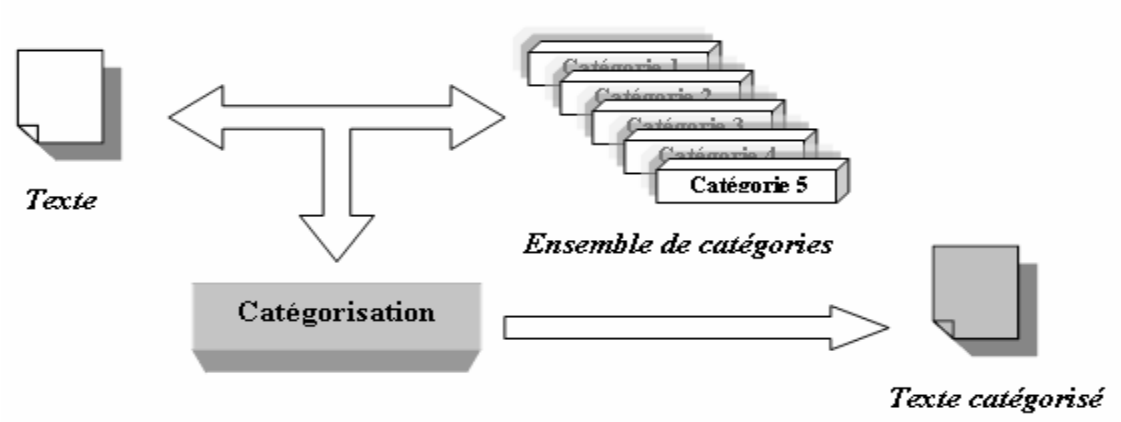


Figure 4-1 : Système de catégorisation de texte

Un autre aspect à mettre en évidence pour une meilleure catégorisation des textes selon [Joachims 00], c'est qu'on procède à la représentation du texte sous certains aspects de sa structure en cinq niveaux :

1- Morphological level: la structure des mots
2- Lexical level: les mots complets
3- Syntactic level: la structure des phrases
4- Semantic level: le sens du texte
5- Pragmatic level: le sens du texte selon le contexte et la situation

De nouveaux horizons sont apparus dans le domaine de la catégorisation automatique, qui reste un créneau compliqué, vu l'interaction des différents domaines scientifiques et la difficulté d'apprendre à une machine à distinguer les formes des écritures comme le fait si bien l'être humain [Carbonnel 05].

4.3 Appariement approximatif des chaînes de caractères

Dans la comparaison des chaînes de caractère, il existe deux approches : la première approche dite comparaison exacte de chaîne *Exact String Matching*, qui consiste à se prononcer si deux chaînes sont similaires ou non, tandis que la deuxième approche appelée appariement approximatif de chaînes *Approximate String Matching*, donne une mesure de similarité entre les chaînes [Parmentier 98], [Navarro 01].

Les applications de l'appariement approximatif de chaîne de caractères sont très vastes [Paleo 07] :

Bioinformatiques : pour l'alignement des séquences de l'ADN ou de l'Aminoacides dans les protéines. Ces séquences peuvent être vus comme des textes et peuvent contenir des mutations et des anomalies, l'appariement approximatif est nécessaire pour contrôler cet alignement.

Recherche documentaire : si les documents qui sont analysés contiennent des erreurs et du bruit, le document recherché pourrait ne pas être trouvé simplement parce que le mot approprié contient des erreurs et par conséquent il ne pourrait pas être découvert par l'appariement exact. Le bruit et les erreurs dans les documents peuvent venir de la numérisation par l'intermédiaire de l'OCR ou des techniques de reconnaissance de la parole et de l'écrit.

Classification des textes : l'appariement approximatif des chaînes de caractères dans la classification des textes basée sur les annotations est très utile pour parer aux erreurs incluses dans le texte.

Recherche Multi-langue : quelques langues sont parfois semblables dans leur vocabulaire, ayant des mots avec la même racine mais avec une morphologie dérivée légèrement différente (par exemple. « Algérie, Algrie », « Andorra, Andorre », « Bhoutan, Bhutan, Butao »). Avec l'appariement approximatif on peut rappeler tous ces mots, sans connaissance particulière de la forme de chaque langue.

Correction des textes : un système pourrait suggérer des corrections pour un mot donné qui est proche à un autre mot dans un lexique de mots. Ceci peut être utile et faisable que pour des applications a vocabulaire limité.

Filtrage de bruit dans le traitement du signal : par filtrage du signal contenant du bruit on peut restituer le signal d'origine selon un code admis.

Plusieurs algorithmes de comparaison de chaînes existent [Haou 08] : algorithmes de recherche de sous-chaîne, algorithmes d'alignement de chaînes et algorithmes de mesure de similarité.

4.3.1 Algorithmes de recherche de sous-chaîne :

Un algorithme de recherche de sous-chaine est un type d'algorithme de recherche qui a pour objectif de trouver une chaîne de caractères à l'intérieur d'une autre. Un tel algorithme fournit la position du premier caractère de la sous-chaîne recherchée dans la chaîne fournie en entrée.

Algorithme naïf

L'idée est de réaliser une comparaison caractère après caractère de la chaîne initiale et de la chaîne recherchée. On parcourt les caractères de la chaîne initiale tant qu'ils sont différents du premier caractère de la chaîne à trouver. Dès qu'on trouve un caractère identique, on parcourt les caractères suivants tant qu'ils correspondent. Si un caractère diffère alors qu'on n'a pas atteint la fin de la chaîne recherchée, alors on reprend la recherche du premier caractère identique, à partir du caractère suivant dans la chaîne initiale. Si tous les caractères correspondent, on retourne la position du premier caractère de la chaîne trouvée dans la chaîne initiale. Enfin, si aucune occurrence de la chaîne recherchée n'apparaît dans la chaîne initiale, l'algorithme se doit de le signaler, en retournant une valeur négative par exemple.

Algorithme d'Aho-Corasick

L'algorithme d'Aho-Corasick est un algorithme de recherche de chaîne de caractère (ou motif) dans un texte dû à Alfred Aho et Margaret Corasick et publié en 1975. L'algorithme consiste à avancer dans une structure de données abstraite appelée *dictionnaire* qui contient le ou les mots recherchés en lisant les lettres du texte T une par une. La structure de données est implantée de manière efficace, ce qui garantit que chaque lettre du texte n'est lue qu'une seule fois. Généralement le dictionnaire est implanté à l'aide d'un trie ou arbre digital auquel on rajoute des liens suffixes. Une fois le dictionnaire implanté, l'algorithme a une complexité linéaire en la taille du texte T et des chaînes recherchées.

L'algorithme extrait toutes les occurrences des motifs. Il est donc possible que le nombre d'occurrences soit quadratique, comme par exemple pour un dictionnaire a, aa, aaa, aaaa et un texte aaaa. Le motif a apparaît à quatre reprises, le motif aa à trois reprises, etc.

Algorithme de Boyer-Moore

L'algorithme de Boyer-Moore est un algorithme de recherche de sous-chaîne particulièrement efficace. Il a été développé par Bob Boyer et Strother Moore en 1977. L'algorithme de Boyer-Moore pré-traite la sous-chaîne (c'est-à-dire la chaîne recherchée), et non pas le texte (c'est-à-dire la chaîne dans laquelle la recherche est effectuée), à l'inverse de certains algorithmes, qui amortissent le coût du prétraitement du texte en effectuant de très nombreuses recherches répétitives. Le coût d'exécution de l'algorithme de Boyer-Moore peut être sub-linéaire, c'est-à-dire qu'il n'a pas besoin de vérifier chacun des caractères du texte, mais peut au contraire sauter certains d'entre eux. En général, l'algorithme devient plus rapide lorsque la longueur de la sous-chaîne s'allonge. Cette efficacité provient du fait que, pour chaque tentative infructueuse de correspondance entre les deux chaînes (texte et sous-chaîne), il utilise les informations déduites de cet échec pour éliminer le plus grand nombre possible de positions à vérifier. Car il effectue la vérification, c'est-à-dire qu'il tente d'établir la correspondance de la sous-chaîne à une certaine position, à l'envers. Par exemple, s'il commence la recherche de la sous-chaîne COMPARAISON

au début d'un texte, il vérifie d'abord la onzième position en regardant si elle contient un N. Ensuite, s'il a trouvé un N, il vérifie la dixième position pour regarder si elle contient le dernier O de la sous-chaîne, et ainsi de suite jusqu'à ce qu'il ait vérifié la première position du texte pour y trouver un C.

Algorithme de Knuth-Morris-Pratt

L'algorithme de Knuth-Morris-Pratt (*KMP*) est un algorithme de recherche de sous-chaîne, permettant de trouver les occurrences d'une chaîne *P* dans un texte *S*. Sa particularité réside en un pré-traitement de la chaîne, qui fournit une information suffisante pour déterminer où continuer la recherche en cas de non-correspondance. Cela permet à l'algorithme de ne pas ré-examiner les caractères qui ont été précédemment vérifiés, et donc de limiter le nombre de comparaisons nécessaires.

Algorithme de Rabin-Karp

L'algorithme de Rabin-Karp est un algorithme de recherche de chaînes de caractères crée par Michael O. Rabin et Richard M. Karp. Cette méthode recherche un motif donné dans un texte grâce à du hachage. L'algorithme n'est pas beaucoup employé pour les recherches d'une seule chaîne mais a une importance théorique et s'avère très efficace pour des recherches de sous-chaînes multiples.

4.3.2 Algorithmes d'alignement de chaînes

Algorithme de Needleman-Wunsch

L'algorithme de Needleman-Wunsch effectue un alignement global maximal de deux chaînes de caractères (appelées ici A et B). Il est couramment utilisé en bioinformatique pour aligner des séquences de protéines ou de nucléotides. L'algorithme a été présenté en 1970 par Saul Needleman et Christian Wunsch. L'algorithme est un exemple de programmation dynamique, tout comme l'algorithme de Levenshtein auquel il est apparenté. Il garantit de trouver

l'alignement de score maximal. Ce fut la première application de la programmation dynamique pour la comparaison de séquences biologiques.

4.3.3 Algorithmes de mesure de similarité

Distance de Jaro-Winkler

La distance de Jaro-Winkler mesure la similarité entre deux chaînes de caractères. Il s'agit d'une variante proposée en 1999 par William E. Winkler, découlant de la distance de Jaro (1989, Matthew A. Jaro) qui est principalement utilisée dans la détection de doublons. Plus la distance de Jaro-Winkler entre deux chaînes est élevée, plus elles sont similaires. Cette mesure est particulièrement adaptée au traitement de chaînes courtes comme des noms ou des mots de passe. Le résultat est normalisé de façon à avoir une mesure entre 0 et 1, le zéro représentant l'absence de similarité.

Distance de Hamming

La distance de Hamming doit son nom à Richard Hamming. Elle est utilisée en télécommunication pour compter le nombre de bits altérés dans la transmission d'un message d'une longueur donnée. Le poids de Hamming correspond au nombre de bits différents de zéro, il est utilisé dans plusieurs disciplines comme la théorie de l'information, la théorie des codes et la cryptographie.

4.3.4 Distance de Levenshtein

La distance de Levenshtein est considérée comme un exemple de la programmation dynamique (Richard Bellman). Son nom provient de Vladimir Levenshtein qui l’a définie en 1965. Elle est considérée comme une généralisation de la distance de Hamming [Paleo 07].

Reconnue aussi sous le nom de distance d'édition, elle est le nombre minimal de remplacements, ajouts et suppressions de lettres pour passer du mot A au mot B, d(A, B). Plus le

nombre de différences entre les deux chaînes de caractères est grand, autant la distance est plus grande. Lorsqu'on recherche le mot A dans un lexique L, si A se trouve dans L alors d(A,A)=0, ou bien A n'est pas inclut dans L, on peut rechercher les mots B de L les plus proches de A, tels que par exemple d(A, B) < k (k nombre petit a définir) [Coetzee03].

4.3.5 Algorithme de Levenshtein

Soit *s* la longueur de la chaîne *S* et *t* la longueur de la chaîne *T*. L'algorithme qui permet de calculer la distance d'édition entre les chaînes *S* et *T* est donné par :

1. Construire une matrice *M* de *s*+1 lignes et *t*+1 colonnes
2. Initialiser la première ligne à l'aide des coûts d'insertion
3. Initialiser la première colonne à l'aide des coûts de suppression
4. Pour i = 1 jusqu'à t faire
   Pour j = 1 jusqu'à s faire
   $M_{i,j} = min\ (M_{i-1,j} + C_{ins}(S_j), M_{i,j-1} + C_{suppr}(T_i), M_{i-1,j-1} + C_{subst}(S_j,T_i))$
   Fin
   Fin
5. Résultat = $M_{s,t}$

Avec :

$M_{i,j}$: l'élément de la matrice M situé ligne i et colonne j

S_j : la j-ème lettre du mot S

T_j : la i-ème lettre du mot T

4.3.6 Déroulement de l'algorithme

Pour mieux comprendre le fonctionnement de l'algorithme de Levenshtein, prenons un exemple simple : Soit s= « NICHE » et t= « CHIENS ». On peut transformer la chaîne s en t en 5 étapes:

- Suppression de N et I
- Ajout de I, N et S → CHIENS

La distance de Levenshtein entre "NICHE" et "CHIENS" est donc égale à 5.

Fonctionnement de l'algorithme

Soit n la longueur de la chaîne s (n=5)
Soit m la longueur de la chaîne t (m=6)

Si n=0 alors retourner d=m et quitter
Si m=0 alors retourner d=n et quitter
Avec d la distance recherchée.

Construire une matrice M de n+1 lignes et m+1 colonnes.
Initialiser de la première ligne par la matrice ligne [0,1,….., m-1, m] et la première colonne par la matrice colonne [0,1,….., n-1, n]

		C	**H**	**I**	**E**	**N**	**S**
	0	1	2	3	4	5	6
N	1	0	0	0	0	0	0
I	2	0	0	0	0	0	0
C	3	0	0	0	0	0	0
H	4	0	0	0	0	0	0
E	5	0	0	0	0	0	0

Soit Cout(i, j)=0 si A(i)=B(j) et Cout(i, j)=1 si A(i) <>B(j)

La matrice Coût résultante:

	C	**H**	**I**	**E**	**N**	**S**
N	1	1	1	1	0	1
I	1	1	0	1	1	1
C	0	1	1	1	1	1

H	1	0	1	1	1	1
E	1	1	1	0	1	1

La matrice M est remplit en utilisant la règle suivante M[i, j] est égale au minimum de:

- L'élément directement avant plus 1: M[i-1, j] + 1. (effacement)
- L'élément directement au dessus plus 1: M[i, j-1] + 1. (insertion)
- Le diagonal précédent plus le coût: M[i-1, j-1] + Cout(i, j). (substitution)

Le remplissage de la première ligne donne le résultat suivant:

		C	**H**	**I**	**E**	**N**	**S**
	0	1	2	3	4	5	6
N	1	1	2	3	4	4	5
I	2	0	0	0	0	0	0
C	3	0	0	0	0	0	0
H	4	0	0	0	0	0	0
E	5	0	0	0	0	0	0

Nous réitérons cette opération jusqu'à remplir la matrice :

		C	**H**	**I**	**E**	**N**	**S**
	0	1	2	3	4	5	6
N	1	**1**	2	3	4	4	5
I	2	**2**	2	2	3	4	5
C	3	**2**	3	3	3	4	5

H	4	3	**2**	**3**	4	4	5
E	5	4	3	3	**3**	**4**	**5**

La distance de Levenshtein entre les mots s et t se retrouve en M[n, m].
La dernière matrice fournit explicitement les opérations nécessaires pour passer d'une chaîne de caractères à l'autre [Coetzee 03].

4.4 Travaux dans le domaine de la catégorisation et la programmation dynamique

Les travaux sur la catégorisation de documents manuscrits se font rares ; Néanmoins on peut citer l'étude de [Vinciarelli 03]. A notre connaissance, Vinciarelli est le premier a présenté une étude sur la catégorisation automatique des textes manuscrits, le corpus utilisé pour l'apprentissage est celui de Reuters 21578 'news' au format électronique, annotées sur 120 thèmes. Ces news ont été retranscrites manuellement, puis soumises à un moteur de reconnaissance. Le taux d'erreur de reconnaissance mot varie entre 10% et 50%, néanmoins les performances de catégorisation ne sont détériorées que d'environ 10% par rapport aux performances obtenues sur la base de documents électroniques. L'effet de cette dégradation due aux transcriptions bruitées est négligeable même dans les deux cas importants: (IR) Information Retrieval et (TC) Text Categorization [Vinciarelli 03]. Notons que le classifieur utilisé par l'approche de Vinciarelli est le SVM, qui selon [Joachims 00] permet d'obtenir les meilleurs taux de pertinence car peu sensible aux phénomènes de sur-apprentissage. Une autre étude basée sur les travaux de Vinciarelli a été menée par Koh sur la catégorisation automatique des courriers entrants avec des résultats prometteurs.

Les études relatives aux documents imprimés dans la mouvance de l'apprentissage automatique qui nécessite un corpus d'entraînement libellé, on trouve le classifieur bayésien naïf (probalistic classifiers) Domingos et Pazzani en 1997, le classifieur arbre de décision (symbolic

classifiers) Chen et Ho en 2000, le K-PPV (exemple-based classifiers) Yavuz et Guvenir en 1998, les réseaux de neurones Schutze, Hull, Pederson et Wiener en 1995, …etc. Dernièrement, on parle de classifieurs hybrides et de comités de classifieurs [Sebastiani 02], [Feldman 07]. Ajoutons, le travail de Cavnar basé sur le modèle statistique des N-grammes employé dans la modélisation probabiliste du langage pour la prédiction des mots [Cavnar 94], on parle même de Chenks qui consiste à décomposer une phrase en syntagmes minimaux non récursifs, permettant ainsi de mieux capter des dépendances en contexte à plus longue distance que le modèle N-gramme [Schadle 04]. Pour palier au problème de construction d'une base d'apprentissage volumineuse, (fastidieuse en temps et en argent) l'approche de [Rehel 05] classifie les textes sur la cooccurrence de mots provenant de documents non étiquetés en s'appuyant sur un lexique d'une ontologie comme WordNet. Notons le modèle de catégorisation de [Spitz 00] de documents électroniques par codage de caractéristiques des caractères en se basant sur les travaux de Ittner en 1995 de l'Université de Washington sur 605 images de documents avec un taux de 72% (Recall) et 82% (precision).

Concernant les travaux de recherche basés sur la mesure de similarité, le premier travail théorique, qui constitue un modèle formel pour manipuler des données bruités, est basé sur la théorie développée par Shanon en 1948. Dans les laboratoires AT&T Bell ont a pu élaboré une voie de transmission bruité telle qu'une ligne téléphonique. Le groupe de recherche de Yorktown Heights à New York a appliqué pour la première fois ce modèle de canal bruité (Noisy Channel Model -NCM-), connu aussi, sous le nom de modèle de canal source, dans un système continu de reconnaissance de la parole.

Depuis, le modèle a été porté dans les secteurs de la traduction automatique, de la correction des textes et dans d'autres applications comme l'étiquetage des discours, l'OCR, la reconnaissance d'écriture et la recherche documentaire. Par exemple, dans la correction de l'orthographe, le texte bruité correspond au texte, contenant des erreurs, résultant d'un dactylographe ou d'un système de reconnaissance d'écrit/discours. Dans la traduction automatique, il correspond à un texte dans une autre langue. Ainsi, la solution au problème est de

trouver ou récupérer le texte original du texte produit par des opérations de base d'édition transformant les données résultantes depuis les originaux [Sari 09].

Dans la littérature, le travail de Spitz semble être le plus proche de la recherche actuelle. Il a proposé l'utilisation de codes sur les caractéristiques (les formes) des caractères [Spitz 00]. D'abord, les mots sont identifiés. Ensuite, chaque caractère dans un mot, est attribué à un code selon la forme du caractère. Par exemple, des lettres avec des ascendantes peuvent être désignées par un code ‘A’, celles avec des jambages par le code ‘g’ (figure 4-2) [Spitz 95]. L'ordre des codes obtenus se note WST (Word Shape Tokens). Notez que cette méthode détermine seulement la forme générale d'un caractère plutôt que d'essayer d'identifier différents caractères. L'attente ici est que l'information sur la forme peut être obtenue plus exactement avec un coût inférieur que celle proposée par l’OCR. A juste titre, Smeaton a démontré que ce dispositif est très utile dans le cas ou la qualité des systèmes d’OCR est très basse [Smeaton 97], [Spitz 00] ou même sans l’utilisation de ces OCR, qui provoquent des erreurs lors de segmentation de caractères, surtout quand ces derniers sont adjacents et se touchent, ainsi le“m”, “lc”, “vv” sont reconnus à tort respectivement comme “m”, “k”, “w” et vice versa [Bai 09]. Chen et bloomberg en 1998 ont proposé une approche libre de segmentation en utilisant l'information de la forme du mot. Ils identifient d'abord, des contours supérieurs et inférieurs de chaque mot en utilisant la morphologie et extraient ensuite, l'information de la structure basée sur l'emplacement des pixels parmi ces découpes. Après, le mot codé est décodé par l’algorithme de Viterbi par correspondance de l'image de mot avec le mot-clé donné [Sari 09]. L’article de [Khurshid 08] présente une approche pour la recherche documentaire reposant sur la l'appariement de mots dans les documents imprimés anciens. Il a proposé une méthode pour rechercher un mot dans une image de document basée sur la mise en correspondance des caractéristiques des images de caractères contenus dans les mots. [Yadav 09] a utilisé la programmation dynamique pour évaluer la similarité des images de mot partielle pour la recherche d’information dans une base d’image de documents imprimés.

Raw text	The Economy in 1988 The global economic development observed in 1988 was far more rapid than had been predicted. At around 4%, Western industrial nations real growth reached its highest level since 1984. The fact that international economic growth was being stimulated more and more by investment activity was to the advantage of Swiss exporters, as was the contraction in the external value of the Swiss franc witnessed during the course of the year.
Original mapping	AAx Axxxxxg ix AAAA AAx gAxAxA xxxxxxix AxxxAxgxxxA xAxxxxxA ix AAAA xxx Axx xxxx xxgiA AAxx AxA Axxx gxxAixAxA. AA xxxxxA AA, AxxAxxx ixAxxAxixA xxAixxx xxxA gxxxAA xxxxAxA iAx AigAxxA AxxxA xixxx AAAA. AAx AxxA AAxA ixAxxxxAixxxA xxxxxxix gxxxAA xxx Axixg xAixxAxAxA xxxx xxA xxxx Ag ixxxxAxxxA xxAixiAg xxx Ax AAx xAxxxAxgx xA Axixx xxgxxAxxx, xx xxx AAx xxxAxxxAixx ix AAx xxAxxxxA xxAxx xA AAx Axixx Axxxx xiAxxxxxA Axxixg AAx xxxxxx xA AAx gxxx.
Enhanced mapping	AAe Aexxxxg ix AAAA AAe gAxAxA eexxxxie AexeAxgxexA xAxexxeA ix AAAA xxx Axx xxxe xxgiA AAxx AxA Aeex gxeAieAeA. AA xxxxxA AA, AexAexx ixAxxAxixA xxAixxx xexA gxxxAA xexeAeA iAx AigAexA AexeA xixee AAAA. AAe AxeA AAxA ixAexxxAixxxA eexxxxie gxxxAA xxx Aeixg xAixxAxAeA xxxe xxA xxxe Ag ixxexAxexA xeAixiAg xxx Ax AAe xAxxxAxge xA Axixx exgxxAexx, xx xxx AAe exxAxxeAixx ix AAe exAexxxA xxAxe xA AAe Axixx Axxxe xiAxexxeA Axxixg AAe exxxxe xA AAe gexx.

Figure 4-2 : Codage des mots d'un texte par Spitz

Contrairement aux travaux sur le latin, peu de recherches ont été réalisées sur la langue arabe, citons l'article de [Khreisat 06] qui démontre une étude sur la classification des textes arabes utilisant les N-grammes en comparant les distance de similarité 'Manhattan distance' et 'Dice'. Autre exemple, l'utilisation du classifieur bayézien par [Joher 06]. Rappelons que tous ces chercheurs ont basé leurs travaux uniquement sur des documents électroniques.

Pour les manuscrits arabes, notons l'exemple d'utilisation des mesure de similarité dans la reconnaissance de formes par [Djeddi 09] qui a proposé une approche locale d'identification hors-ligne de scripteurs pour l'écriture arabe, le scripteur inconnu d'un document et selon son écriture spécifique sera identifié par un calcul de similarité entre le document considéré et l'ensemble des documents de la base de référence. La prise de décision est basée sur l'algorithme de Wagner-Fisher pour le calcul de la distance d'édition. [Sari 09] a développé un système de recherche de mot dans des documents manuscrits arabes basé sur la distance de levenshtein. Benmohamed et Sari dans [Benmohamed 09] ont suggéré une approche d'indexation et de recherche de documents anciens arabes par une modélisation de la structure du document via une représentation du contenu en utilisant les descripteurs de contour.

4.5 Conclusion

Dans la recherche d'information (RI) ainsi que dans la fouille de données, les algorithmes de recherches prennent une place de plus en plus importante. La recherche d'une chaîne de caractère dans un flot de données est un domaine de recherche algorithmique des plus anciens. En reconnaissance de l'écriture, la distance de Levenshtein est la méthode la plus couramment utilisée pour le calcul de distance d'édition entre deux chaînes. C'est pour cette raison, que nous avons porté une attention particulière sur la technique d'appariement approximatif des chaînes de caractères pour pouvoir l'utiliser ultérieurement dans la phase de catégorisation. Le chapitre suivant décrit l'implémentation de notre modèle basé sur la technique expliquée auparavant.

CHAPITRE 5

EXPERIMENTATION ET RESULTATS

5 Expérimentation et résultats

5.1 Introduction

Notre modèle proposé pour la catégorisation automatique des manuscrits arabes se compose d'une série de processus (figure 5-1) et (Annexe).

Nous avons opté pour une politique qui consiste à n'extraire des documents manuscrits que l'information utile pour la catégorisation. Ainsi, la reconnaissance de mots manuscrits isolés ne se fera que sur des mots pertinents issus d'un dictionnaire préétabli pour la tache de classification.

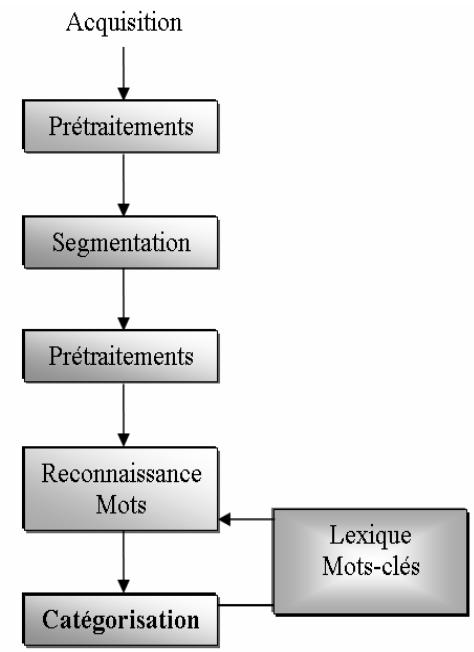


Figure 5-1 : Structure du modèle proposé

5.2 Les prétraitements

Après une tache de numérisation des manuscrits Arabes, une série de prétraitements est appliquée (figure 5-2)

Figure 5-2 : Résultats de prétraitement

5.2.1 Notion sur l'image numérique

Une image numérique est définie comme un signal fini bidimensionnel échantillonné à valeurs quantifiées dans un certain espace de couleurs. Elle est constituée de points (pixels).

- Signal fini : une image possède des dimensions finies : 640x480, 800x600 points…
- Signal bidimensionnel : une image possède deux dimensions : largeur, hauteur.
- Signal échantillonné : les pixels d'une image sont régulièrement espacés sur une grille carrée.
- Valeurs quantifiées : les valeurs des pixels appartiennent à un intervalle borné connu.
- Espace de couleur : il existe de nombreuses façon de percevoir les couleurs d'une image, l'espace de représentation le plus connu est l'espace RGB (rouge-vert-bleu).

Autrement dit, une image est une matrice MxN de valeurs entières prises sur un intervalle borné [0,Ng] où Ng est la valeur maximale du niveau de gris. (figure 5-3)

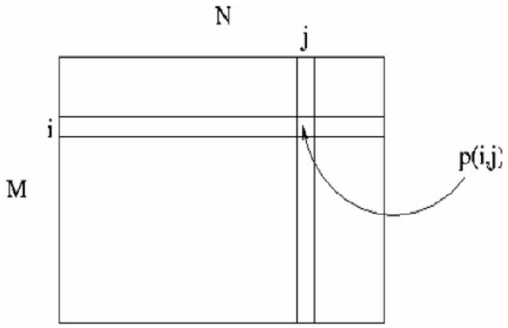


Figure 5-3 : Image numérique

p(i,j) est le niveau de gris du pixel de coordonnées ligne i et colonne j dans l'image.
p(i,j) est contenu dans [0, Ng]. Les valeurs des niveaux de gris sont des entiers.

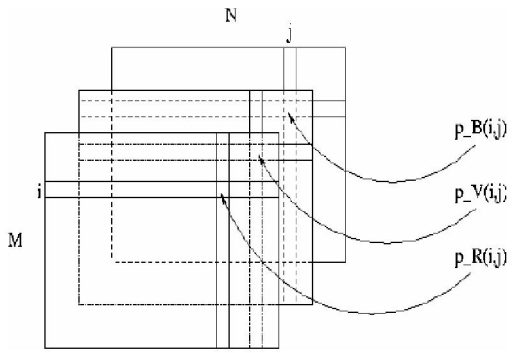


Figure 5-4 : Image couleur

Une image couleur est la composition de trois (ou plus) images en niveaux de gris sur trois (ou plus) composantes. On définit donc trois plans de niveaux de gris, un rouge, un vert et un bleu. La couleur finale est obtenue par synthèse additive des ces trois (ou plus) composantes. (figure 5-4)

5.2.2 Le niveau de gris

L'image traitée est en niveau de gris (c-à-d les intensités vont du blanc jusqu'au noir) où chaque pixel est représenté par une association de trois couleurs rouge, vert et bleu (RGB). (figure 5-5)

Par exemple, le niveau (X=255,Y=255,Z=255) correspond au blanc où :

X est la position du rouge

Y est la position du vert

Z est la position du bleu.

Pour transformer une image couleur en niveaux de gris, on remplace les composantes de chaque pixel par sa valeur de luminosité**,** (r, g, b) devient (L, L, L) où L = (0,3*r+0,59*g+0,11*b).

Une image en niveaux de gris autorise un dégradé de gris entre le noir et le blanc. En général, on code le niveau de gris sur un octet (8 bits) soit 256 nuances de dégradé. L'expression de la valeur du niveau de gris avec Ng = 256 devient: p(i,j) est contenu dans [0, 255].

Pour chaque pixel (i, j) de l'image **faire**

I (i,j) =0 ,3***r** + 0,59***g**+ 0,11***b**

Fin pour ;

Où **I** est l'intensité de chaque pixel.

r : la luminosité de la couleur rouge.

g : la luminosité de la couleur verte.

b : la luminosité de la couleur bleue.

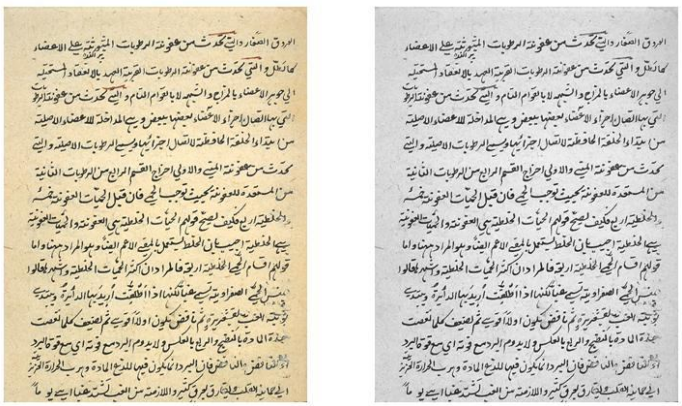


Figure 5-5 : Niveau de gris

5.2.3 La binarisation (seuillage)

Selon la distribution des niveau de gris des images, certaines images se prêtent relativement bien à une opération de binarisation (figure 5-6). L'objectif est de convertir des images en niveau de gris en images binaires. De cette façon, la taille des images est réduite simplifiant leur traitement. L'algorithme de binarisation d'une image est le suivant (« image » représente l'image initiale et « imageb » l'image résultat de la binarisation) :

pour i **variant de** 1 **à** hauteur_image **faire**
  pour j **variant de** 1 **à** largeur_image **faire**
    si image(i,j) < seuil **alors** imageb(i,j) ← 0
    sinon imageb(i,j) ← 1
  fin pour
fin pour

Une image binaire est une image MxN où chaque point peut prendre uniquement la valeur 0 ou 1. Les pixels sont noirs (0) ou blancs (1). Le niveau de gris est codé sur un bit (Binary digIT). Dans ce cas, on revient au cas donné en définition.
avec Ng = 2 et la relation sur les niveaux de gris devient: p(i,j) = 0 ou p(i,j) = 1.

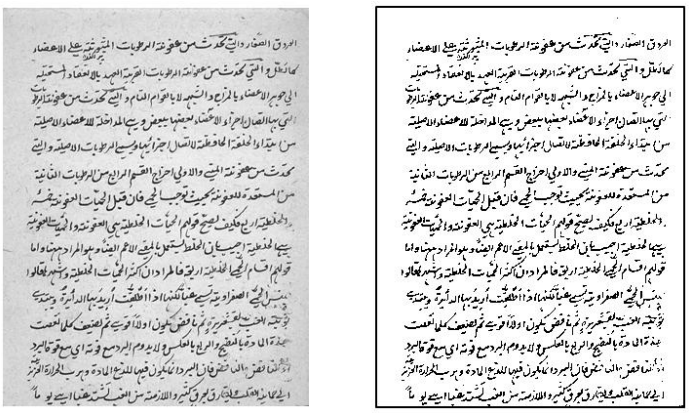


Figure 5-6 : Binarisation

5.2.4 Le lissage

Le filtrage est un traitement local qui agit sur les pixels en examinant leur voisinage (figure 5-8). Il consiste à remplacer l'intensité d'un pixel ou bien son niveau de gris (qu'on notera par la suite I(i,j)) par la moyenne(M) des intensités des pixels entourant. La figure 5-7 ci-dessous représente deux voisinages du pixel p :

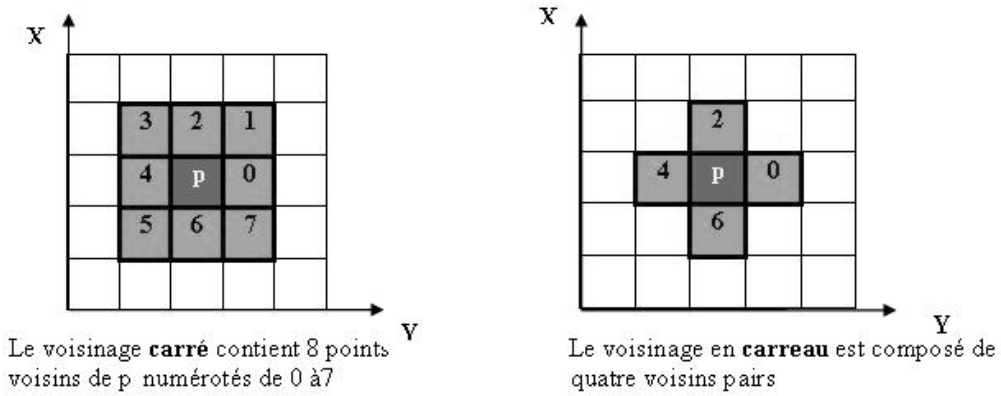


Figure 5-7 : Voisinage carreau et carré

La moyenne est calculée par la formule suivante

$M=1/8 \sum I(i,j)$ où $|i-x| \leq 1 ; |i-y| \leq 1 ; (i , j) \neq (x , y)$

$M=1/4 \sum I(i,j)$ où $|i-x| \leq 1 ; |i-y| \leq 1 ; (i , j) \neq (x , y)$

Le filtrage effectué dans notre application est basé sur le voisinage carré. L'algorithme suivant illustre cette opération :

Pour chaque pixel (**i,j**) de l'image **faire**

I(**i,j**)=1/8 **(**I(**i-1, j-1**) + I(**i-1, j**) + I(**i-1,j+1**) + I(**i,j+1**)+

I(**i+1,j+1**)+ I(**i+1,j**) + I(**i+1,j-1**) + I(**i,j-1**) +

I(**i-1,j-1**)**) ;**

Fin pour ;

où I est l'intensité de chaque pixel.

Figure 5-8 : Lissage

5.3 La segmentation

La segmentation consiste à créer des partitions dans l'image en identifiant des pixels similaires. Le critère de similarité permet de contrôler l'aspect final de la segmentation (regroupement par couleur, par homogénéité, par taille). La segmentation a pour objectif de différencier des zones d'intérêt (par exemple objets/fond). C'est généralement une première étape d'un traitement plus complexe, comme le filtrage adaptatif ou la reconnaissance de formes. Les méthodes de segmentation étant sensibles au bruit, il est nécessaire de commencer par nettoyer l'image en appliquant les filtres usuels d'atténuation de bruit.

La segmentation par régions est une approche spécifique dans laquelle, on cherche à construire des surfaces en regroupant des pixels voisins suivant un critère d'homogénéité. Au final, la segmentation par région crée un ensemble de régions qui ont les propriétés suivantes :

- la réunion de toutes les régions donne l'image entière.
- les régions sont connexes (c'est à dire que tous les pixels d'une même région sont jointifs).

- tous les pixels d'une même région sont homogènes entre eux.
- les pixels de deux régions adjacentes ne sont pas homogènes entre eux.

Cette approche se distingue, par exemple, des segmentations par contours ou par seuillage dans lesquelles les régions créées ne possèdent pas toutes ces propriétés.

5.3.1 Séparation texte/graphique

Les manuscrits arabes sont la cible de notre recherche, ils combinent des parties manuscrites et graphiques, un prétraitement de plus est appliqué : celui de « séparation texte/graphique » (figure 5-9).

L'idée de la méthode par croissance de régions consiste à regrouper des petites régions uniformes et adjacentes d'une taille de quelques pixels dans l'image jusqu'aucun regroupement ne sera réalisable.

L'idée de la méthode utilisée consiste à explorer l'image à partir de petites régions et à faire croître celles-ci en utilisant les techniques de recherche dans un arbre. L'algorithme qui permet de faire croître au maximum une région avant de s'intéresser à la suivante est comme suit:

```
Debut
    pour chaque pixel (i,j) de l'image faire
        si (I(i,j) < > 0) alors
            sauvegarder le point initial I(i,j) ;
            CROISSANCE ( pixel (i,j) ) ;
            incrementer numero region ;
        fin si ;
    fin pour ;
fin.
```

La procédure CROISSANCE permet d'avoir les pixels de chaque région ainsi que leurs pixels voisins.

Procedure CROISSANCE (pixel (i,j))

Debut

on met I (i,j) à 0 afin de ne pas réexaminer ce point (marquage)

pour tout pixel adjacent à I (i,j) faire

si I (i,j) < > 0 alors CROISSANCE (pixel (i,j)) ;

fin si ;

fin pour ;

fin.

pour séparer les zones de textes des zones graphiques, on choisi comme critère le nombre de pixel présents dans chaque région, si ce nombre est supérieur à un seuil donné alors cette région est une image sinon c'est un texte.

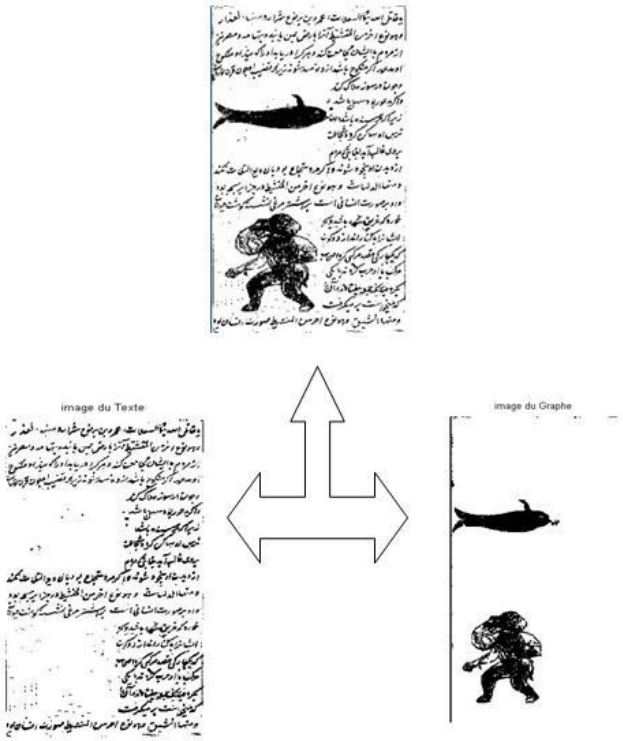


Figure 5-9 : Séparation texte/graphique

5.3.2 Segmentation du texte en lignes

Nous arrivons à l'étape de la segmentation du document en lignes de texte en utilisant la méthode de histogramme, qui suppose que les lignes de textes soient : droites ou très peu inclinées, relativement espacées de manière à ce que les hampes et les jambages des lettres ne se chevauchent pas. Alors, il sera possible d'opérer une projection horizontale de l'ensemble du document (figure 5-10).

Figure 5-10 : Segmentation en lignes

5.3.3 Segmentation des lignes en mots

Ensuite, on procède à une segmentation de la ligne en mots selon un critère de distance inter-mots (figure 5-11). La règle étant que l'espace entre les mots est plus grand que l'espace entre les lettres des mots.

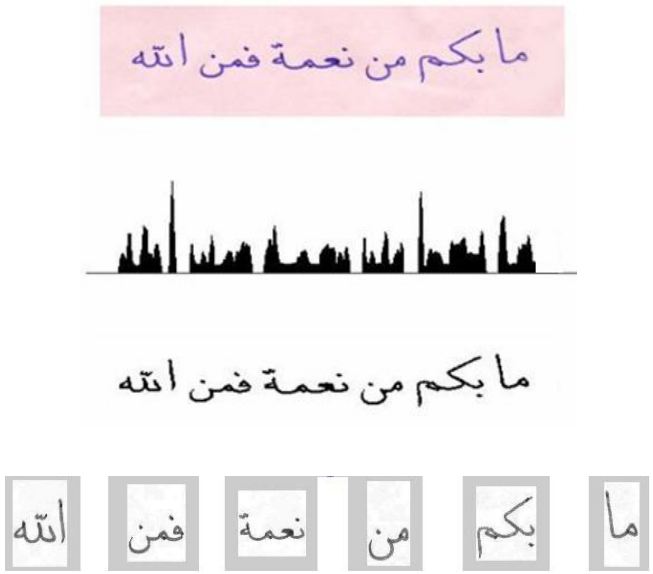


Figure 5-11 : Segmentation en mots

5.4 Prétraitements mots

Une autre suite de prétraitements suivra pour nettoyer le mot de bruit, subit lors de la phase de segmentation (lissage, détection des contours,…).

5.4.1 Morphologies mathématiques

Le mot segmenté subit une série de transformations par morphologie mathématique (figure 5-12) pour atténuer le bruit.

La composition d'une dilatation morphologique avec l'érosion par le même élément structurant produit deux autres opérateurs morphologiques, l'ouverture et la fermeture morphologique. Cette dernière est le dual de la première. L'élément structurant joue le rôle de modèle local, ou de sonde. Il est promené partout sur l'image à traiter, et à chaque position on étudie sa relation avec l'image binaire, considérée comme un ensemble. Ces relations peuvent être du type « est inclus dans l'ensemble », ou « touche l'ensemble ».

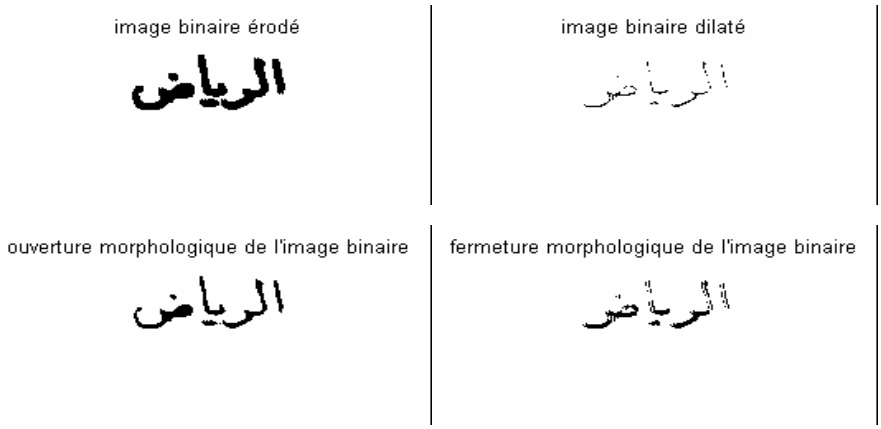


Figure 5-12 : Morphologies mathématiques sur le mot

5.4.2 Détection du contour

L’extraction du contour est une fonction économique qui permet d’alléger le processus de reconnaissance d’objet et de localisation tridimensionnelle (figure 5-13).

Figure 5-13 : Contour du mot

Le contour de la forme est extrait et codé selon les huit directions de Freeman (figure 5-14).

Le code de Freeman se base sur 3 éléments pour caractériser une forme: Les coordonnées (x,y) d'un point appartenant au contour de la forme, un sens de rotation et la chaîne de caractère qui code le contour.

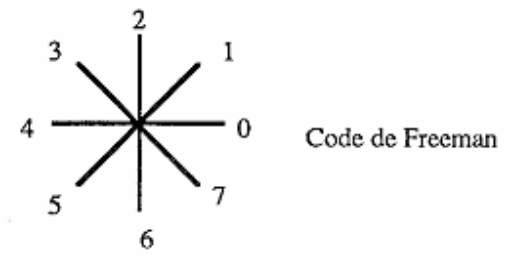


Figure 5-14 : Les 8 directions de Freeman

5.5 Reconnaissance mots isolés

Etant donné qu'il n'était pas possible de faire une reconnaissance intégrale du contenu des documents manuscrits, nous avons choisi une approche directe : celle qui consiste à retenir l'information utile des documents manuscrits pour la catégorisation. Ainsi, la reconnaissance de mots manuscrits isolés ne se fera que sur des mots pertinents issus d'un dictionnaire préétabli pour la tâche de classification.

5.5.1 Extraction des caractéristiques

Nous avons utilisé des caractéristiques de haut niveau (structurelles) (figure 5-15) qui sont indépendantes des styles d'écritures évitant ainsi le problème de la variation des formes et qui peut être bénéfique dans un contexte omni-scripteurs. On retrouve par rapport à la ligne de base et la zone médiane de chaque mot les Hampes, Jambages, Boucles, Points diacritiques haut et bas.

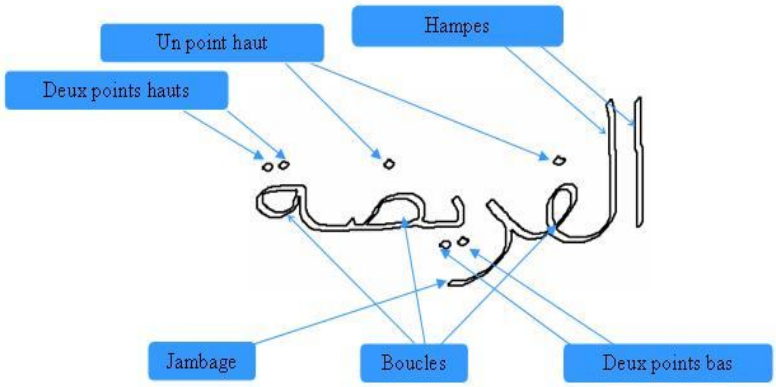


Figure 5-15 : Caractéristiques utilisées du mot

5.5.1.1 Localisation de la ligne de base

Les lignes d'appui délimitent les zones contenant les ascendants et descendants. Ces lignes sont importantes en reconnaissance de l'écriture. Cette information est utilisée pour détecter ascendants et descendants, et également pour normaliser les primitives, les rendant ainsi moins dépendantes de la hauteur de l'écriture. C'est par la méthode de projection horizontale de l'image, que la ligne de base est reconnue. Elle correspond à celle dont la projection contient le plus grand nombre de pixels noirs (figure 5-16).

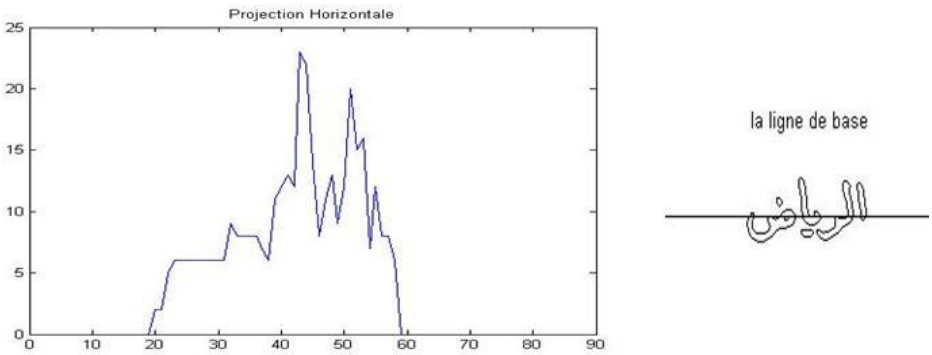


Figure 5-16 : Ligne de base

5.5.1.2 Détection des points diacritiques

Les points diacritiques sont les points simples ou multiples qui peuvent être 1 point, 2 points ou 3 points positionnés au dessus ou au dessous du corps du caractère. Leur position est déterminée par rapport à la zone supérieure ou inférieure du mot respectivement les points diacritiques hauts et bas.

La détection des points diacritiques se fait par un calcul de la surface de la composante connexe par rapport à un seuil donné (figure 5-17).

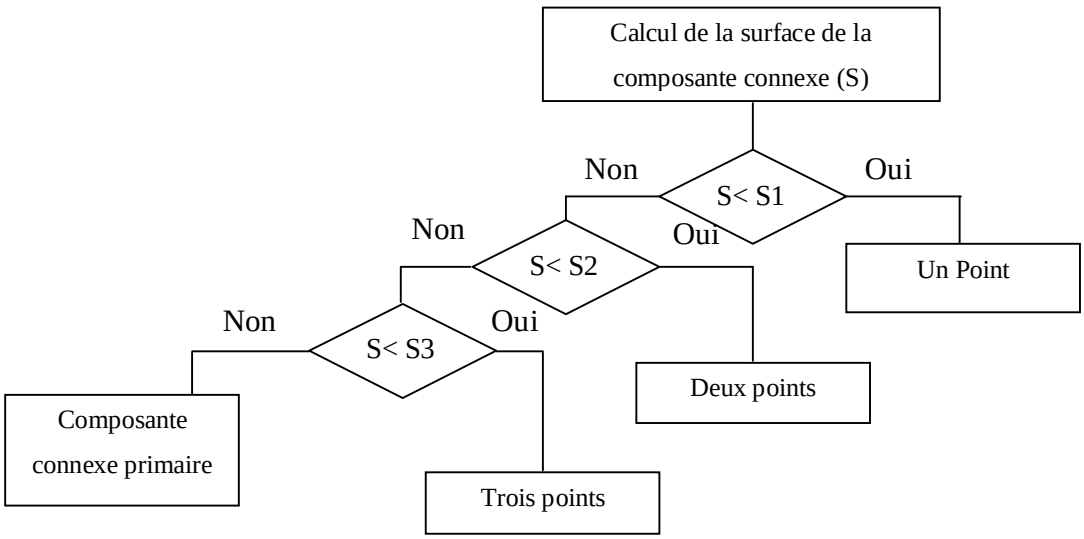


Figure 5-17 : Détection des points diacritiques

Un point diacritique est considéré au-dessus ou au-dessous, respectivement si sa position vis-à-vis la ligne de base est au-dessus ou au-dessous.

5.5.1.3 Détection des jambages

Les jambages ou les descendants sont des primitives utilisées souvent dans la reconnaissance de l'écriture manuscrite. Elles sont représentées par une descente qui se termine

en dehors de la zone inférieure. Elles correspondent aux pixels ayant la valeur 5, 6 et 7 selon le code de Freeman.

5.5.1.4 Détection des hampes

Contrairement aux jambages, les hampes sont représentées par une montée qui dépasse la zone supérieure. Elles correspondent aux pixels ayant la valeur 1, 2 et 3 selon le code de Freeman.

5.5.1.5 Détection des boucles

Les boucles ou les cercles se trouvent généralement dans la zone médiane dans le cas de l'écriture arabe. Elles sont définies comme étant des contours internes dans d'autres.

5.5.1.6 Détection des composantes connexes primaires

Les composantes connexes d'une image représentent un regroupement de tous les pixels voisins dans un ensemble commun. A ce niveau les points diacritiques sont considérés comme composantes connexes (secondaires). Les pseudo-mots appelés composantes connexes primaires sont des ensembles de caractère liés ou attachés (figure 5-18).

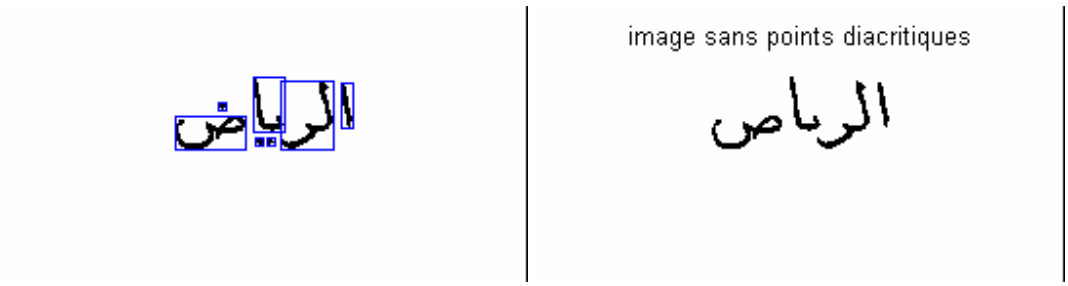


Figure 5-18 : Les composantes connexes d'un mot

5.5.2 Codage

Une opération de codification pour chaque mot est réalisée par la suite du processus d'extraction des caractéristiques. Ces codes seront comparés ultérieurement par rapport aux codes de mots clés des catégories sélectionnées auparavant.

La codification est réalisée par une représentation en séquence des caractéristiques suivantes (tableau 5-1) :

Caractéristiques	***Initiales***
Les boucles	b
Les hampes	h
Les jambages	j
Un point haut	r
Deux points haut	s
Trois points haut	t
Un point bas	u
Deux points bas	v

Tableau 5-1 : Correspondances des caractéristiques mots utilisés

Algorithme de codage :

Entrée : images des mots après extraction des caractéristiques

Sortie : chaîne de code

Début

Pour chaque image de mots faire :

```
Début
    Pour chaque composante connexe primaire (pseudo-mot) faire :
    (On commence par un balayage de haut en bas et de droite à gauche)
        Pour chaque pixel de la composante faire :
        Si {un point haut} alors
            Début
                Ajouter « r » à la chaîne de code
                Marquer ce pixel et tous ses voisins
            Fin
        Sinon si {deux points haut} alors
            Début
                Ajouter « s » à la chaîne de code
                Marquer ce pixel et tous ses voisins
            Fin
        Sinon si {trois points haut} alors
            Début
                Ajouter « t » à la chaîne de code
                Marquer ce pixel et tous ses voisins
            Fin
        Sinon si {un point bas} alors
            Début
                Ajouter « u » à la chaîne de code
                Marquer ce pixel et tous ses voisins
            Fin
        Sinon si {deux points bas} alors
            Début
                Ajouter « v » à la chaîne de code
                Marquer ce pixel et tous ses voisins
```

```
            Fin

        Sinon si {hampe}
            Début
                Ajouter « h » à la chaîne de code
                Marquer ce pixel et tous ses voisins
            Fin
        Sinon si {jambage}
            Début
                Ajouter « j » à la chaîne de code
                Marquer ce pixel et tous ses voisins
            Fin
        Sinon si {boucle}
            Début
                Ajouter « b » à la chaîne de code
                Marquer ce pixel et tous ses voisins
            Fin
        Ajouter « § » à la chaîne de code
Fin
Fin.
```

5.6 Catégorisation

Pour la phase de catégorisation, nous avons adopté une autre voie, différente de celle d'un système classique de catégorisation comme dans [Sebastiani 02], [Vinciarelli 03], [Aas 99]. Ce choix est dicté par le fait que la constitution d'une base de documents manuscrits annotés au niveau mot s'avère fastidieuse et longue. En effet, nous avons choisi de calculer l'appartenance d'un document à une catégorie par un calcul de similarité entre les mots a reconnaître du

manuscrit avec ceux du lexique préétabli, ou chaque catégorie est définie par ces mots clés. On parle alors, de la distance de Levenshtein.

5.6.1 Codage mots-clés du lexique

Vu la difficulté observée dans les manuscrits arabes anciens, nous avons construit une base de textes manuscrits de différentes catégories (76 mots-clés).

Les catégories retenues pour le test sont représentées dans le tableau 5-2.

RELIGION	الدين, التفسير, الفقه , الفتوى, الإسلام , الله, الشريعة, القرآن, الرسول , الرحمان.	**الدين**
HISTOIRE	التاريخ , الحرب , السياسة , الاستعمار , الجيش , الجنود , المؤرخ , الملك , الباي , الداي.	**التاريخ**
LITTERATURE	الأدب , النثر, القصيدة , الرواية, المسرحية , المقالة , الحوار, الترجمة, النقد .	**الادب**
CHIMIE	الكيمياء , سائل , محلول , مختبر, ذرة , الكتلة , كربون , هيدروجين, تجربة .	**الكيمياء**
MEDECINE	الطب, العظام , الدم , الخلية , العصب , التعفن, المرض, الغثيان.	**الطب**
ASTRONOMIE	الفلك , النجم , الكوكب , المجرى , الشمس , القمر, مدار, نيزك , شهاب , الزهرة .	**الفلك**

الرياضة	الرياضة , كرة , جري , مباراة , مبارزة , مضمار, حلبة , ملاكمة , ملعب, دراجة.	SPORT
الموسيقى	الموسيقى , عازف , فرقة , عود , ناي , بيانو , جوق نوطة , أغنية , فنان.	MUSIQUE

Tableau 5-2 : Catégories retenues

Un exemple de mots clés codés est représenté dans le tableau 5-3

Mots	***Codes***
الرحمان	§jr§hb§jh§h
الرياضة	§sbrb§vh§jh§h

Tableau 5-3 : Exemple de mots codifiés

5.6.2 Calcul de la distance d'édition

Le mot est affecté à la catégorie dés que la distance d'édition est minimale, ce seuil sera fixé par la suite, après comparaison des résultats sur plusieurs valeurs. Le pourcentage d'appartenance du document à une classe est calculé par rapport aux mots trouvés par catégories. De ce fait, le document appartient à cette classe quand le pourcentage est le plus élevé (figure 5-19).

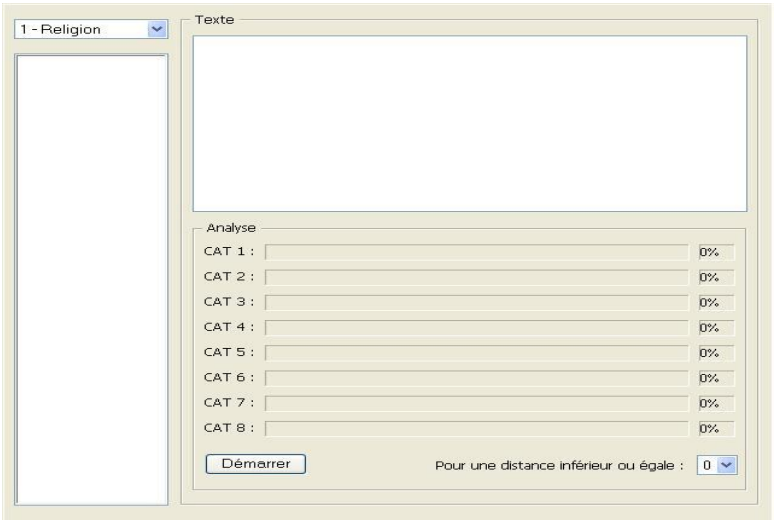


Figure 5-19 : Module de catégorisation

5.6.2.1 Exemple de distance entre deux mots

Voir un exemple de codage entre les mots الرحمان et الرحيمان dans le tableau 5-4.

الرحمان	§jr§hb§jh§h
الرحيمان	§jr§hbv§jh§h

Tableau 5-4 : Codification de deux mots

5.6.2.2 Calcul de distance entre deux mots

Le calcul de distance entre les deux mots donnés comme exemple est le suivant :

§jr§hb§jh§h 11 lettres sont restées inchangées.

§jr§hbv§jh§h Une seule lettre a été ajoutée.

		§	j	r	§	h	b	v	§	j	h	§	h
	0	1	2	3	4	5	6	7	8	9	10	11	12
§	1	0	1	2	3	4	5	6	7	8	9	10	11

j	2	1	0	1	2	3	4	5	6	7	8	9	10
r	3	2	1	0	1	2	3	4	5	6	7	8	9
§	4	3	2	1	0	1	2	3	4	5	6	7	8
h	5	4	3	2	1	0	1	2	3	4	5	6	7
b	6	5	4	3	2	1	0	1	2	3	4	5	6
§	7	6	5	4	3	2	1	1	1	2	3	4	5
j	8	7	6	5	4	3	2	2	2	1	2	3	4
h	9	8	7	6	5	4	3	3	3	2	1	2	3
§	10	9	8	7	6	5	4	4	3	3	2	1	2
h	11	10	9	8	7	6	5	5	4	4	3	2	1

d(a,b)=1

Tableau 5-5 : Distance entre deux mots

5.7 Résultats des testes

Nous avons testé la catégorisation sur 47 textes codifiés manuellement, pour une simple raison : c'est que le module d'extraction de caractéristiques affichait des résultats incorrects dans la plupart du temps. Quelques résultats obtenus sont résumés dans le tableau 5-6 suivant :

Documents	**Catégories**								**Résultat catégorisation**
	R%	**H%**	**L%**	**C%**	**M%**	**A%**	**S%**	**Mu%**	
CAT62	21	7	28	14	42	71	7	0	Astronomie
CAT31	16	16	48	32	36	32	16	12	Littérature
CAT14	57	28	28	28	28	14	28	0	Religion
CAT24	14	42	28	14	14	0	0	14	Histoire

CAT71	0	0	0	25	16	16	58	29	Sport
CAT13	80	60	40	40	40	60	0	0	Religion
CAT23	16	33	0	0	0	0	16	16	Histoire
CAT81	0	0	5	21	15	5	5	52	Musique
CAT72	0	0	0	7	0	0	92	21	Sport
CAT82	0	0	16	0	8	0	0	83	Musique

Tableau 5-6 : Exemple des résultats de la catégorisation

Avec

R : Religion

H : Histoire

L : Littérature

C : Chimie

M : Médecine

A : Astronomie

S : Sport

Mu : Musique

Discussions

Les résultats obtenus par le module de catégorisation sont corrects pour des distances inférieures ou égales à 2, ce qui prouve que le module de catégorisation est fonctionnel. Néanmoins, il faut faire attention aux difficultés qu'on peut rencontrer ; certains mots distincts de catégories différentes peuvent faire référence à un même code, c'est la synonymie. Ou encore un même mot unique peut véhiculer des codes différents, on parle de polysémie.

L'algorithme de la distance de Levenshtein ne s'occupe que de trois opérations, il ne sait détecter que la suppression ou l'insertion d'une lettre, ainsi que le remplacement d'une lettre par

une autre. Il serait intéressant d'utiliser une version améliorée de cet algorithme pour mieux capter la dynamique des mots en arabes.

Une collaboration avec un expert en linguistique de la langue arabe sera bénéfique pour mieux cerner le problème de la reconnaissance et la réussite des travaux de recherche dans ce domaine, ainsi qu'une meilleure compréhension de tous les aspects de cette dernière.

5.8 Conclusion

Nous avons présenté dans ce chapitre une approche permettant de réaliser une catégorisation thématique des documents manuscrits arabes. Elle est réalisée à partir des mots clés proposés à priori. Evitant ainsi, la constitution d'une base de documents manuscrits annotés au niveau mot.

Dans la littérature, il existe peu d'approches utilisées pour résoudre le problème de la catégorisation des documents écrits, encore moins pour les manuscrits arabes.

Les résultats obtenus par le module de catégorisation, sont concluants et nous encouragent à suivre cette voie, pour améliorer le modèle proposé au niveau de la reconnaissance des mots, précisément dans le module d'extraction des caractéristiques, qui reste faillible et dresse actuellement un obstacle majeur dans la reconnaissance de l'écrit et rend la phase de classification encore plus complexe et difficile à mettre en oeuvre.

CHAPITRE 6

CONCLUSION ET PERSPECTIVES

6 Conclusion et perspectives

6.1 Conclusion

A travers l'écriture de ce mémoire nous avons constaté que la catégorisation automatique de document est un processus long et compliqué à mettre en place, surtout celle des textes manuscrits arabes.

Cette première conclusion n'est que logique, quand on voit que le processus de catégorisation est la combinaison d'une variétés de sous processus inspirés des techniques développées dans plusieurs domaines scientifiques : la numérisation, les prétraitements, l'analyse de l'image par les procèdes de segmentation, la reconnaissance de l'écrit et enfin, la catégorisation des textes.

La majorité des méthodes utilisées pour la phase de catégorisation automatique des textes sont basées sur l'apprentissage. C'est pourquoi nous avons adopté une autre démarche basée sur le calcul de distance d'édition, l'algorithme utilisé est celui de Levenshtein. L'approche repose sur un dictionnaire limité (approche à vocabulaires fermés). Dans lequel, cette problématique a été très peu abordée, sauf dans des travaux traitants des documents textuels imprimés.

La spécificité des manuscrits anciens et la complexité du caractère cursif arabe posent d'énormes obstacles pour le processus de classification. On note aussi, l'absence de modèles, dans ce domaine précis, qui nous permettent de comparer les résultats. Ajouter à cela, le manque d'une base d'images de référence pour mener à bien les expérimentations.

Toutes ces démarches entreprises dans le domaine comportent quelques lacunes qui doivent être résolues, cela nous amène à constater que ce domaine est très prometteur et très porteur pour le futur proche.

6.2 Perspectives

Les perspectives de notre travail portent sur :

- La conception d'une base d'image de document manuscrit de test plus grande.
- La construction du lexique mots-clés automatiquement.
- Les catégories hiérarchisées.
- La reconnaissance mots globale et analytique (combinées).
- L'utilisation d'autres opérateurs pour le prétraitement.
- L'utilisation d'autres types d'approches de séparation texte/graphique.
- L'utilisation d'autres méthodes de segmentation en lignes
- L'utilisation d'autres méthodes de segmentation en mots.
- L'ajout d'autres types de caractéristiques du mot.
- Les comités de classifieurs.
- La collecte de documents anciens afin de préparer une base de données qui servira pour les traitements futurs.

BIBLIOGRAPHIE

Bibliographie

[Aas99] K.Aas, L.Eikvil. Text categorisation: A survey. Research Report, Norwegian Computing Center, 1999.

[Agfa-Gevaert94] Agfa-Gevaert, An Introduction to Digital Scanning (Digital Colour pepress,4), Mortser (Belgique), Agfa-Gevaert, 1994.

[Al-Shatnawi08] A.AL-Shatnawi, O.Khairuddin. Methods of Arabic Language Baseline Detection – The State of Art IJCSNS International Journal of Computer Science and Network Security, Vol.8 No.10, October 2008.

[Amin98] A.Amin. Off-line arabic character-recognition : The state of the art. Pattern Recognition, 31(5):517–530, 1998.

[Arrivault02] D.Arrivault. Apport des Graphes dans la Reconnaissance Non-Contrainte de Caractères Manuscrits Anciens. Thèse de doctorat, Université de Poitiers, 2002.

[Atanasiu03] V.Atanasiu. Le phénomène calligraphique à l'époque du sultanat mamluk. Thèse de Doctorat, Ecole pratique des Hautes Etudes, Section des Sciences historiques et philologiques, Paris, 2003.

[Augustin01] E.Augustin. Reconnaissance de mots manuscrits par systèmes hybrides : Réseaux de Neurones et Modèles de Markov Cachés. Thèse de Doctorat, Université René Descartes, Paris v, 2001.

[Bai09] S.Bai, L.Li, C.L.Tan. Keyword Spotting in Document Images through Word Shape Coding. 10th International Conference on Document Analysis and Recognition, 2009.

[Baillie03] J.C.Baillie. Traitement d'Image et Vision Artificielle. Cours ES322, module D9. Ecole Nationale Supérieure de Techniques Avancées, Paris, 2003.

[Belaid 06] A. Belaid. Ecrit et document : Concepts, techniques et expérimentations. Cours sca-m2-tmn. Loria, 2006.

[BenAmor06] N.Ben Amor. Multifont Arabic Characters Recognition Using HoughTransform and HMM/ANN Classification. National Engineering School of Tunis, Tunisia journal of multimedia, vol. 1, no. 2, 2006.

[Benmohamed 09] A. Benmohamed, T. Sari, M. Sellami. Une approche semi automatique pour la recherche de documents anciens. Journées sur la Gestion Electronique de Documents & Réseaux de Recherche en Informatique GED'09 , p157-164, les 20-21 Mai, Annaba, 2009.

[Bensefia03] A.Bensefia, T.Paquet, L.Heutte. Documents Manuscrits et Recherche d'Information. Article dans Document Numérique, Hermès, vol. 7, no. 3-4, pp. 47-60, 2003.

[Bloch06] I.Bloch, F. Tupin, A.Manzanera. TERI : Traitement et reconnaissance d'images. Cours Master 2 IAD. Ecole Nationale Supérieure de Techniques Avancées, Paris, 2006.

[BnF09a] BnF :Bibliothèque nationale de France. Mémento sur les formats d'images. Février, 2009.

[BnF09b] BnF :Bibliothèque nationale de France. Charte de numérisation. Référence : BnF-ADM-2008-068449-01, février 2009.

[Boulehmi08] H. Boulehmi – B. Seddik – A. Kricha – N. Essoukri Ben Amara. Prétraitement de documents anciens. Actes du dixième Colloque International Francophone sur l'Ecrit et le Document CIFED, 2008.

[Boussellaa06] W.Boussellaa, A.Zahour, B.Taconet, A.Benabdelhafid, A.Alimi. Segmentation texte /graphique : Application au manuscrits Arabes Anciens. Neuvième Colloque International Francophone sur l'Ecrit et le Document, Fribourg, Suisse. 18-21 Septembre, 2006.

[Burrow04] P.Burrow. Arabic Handwriting Recognition. School of Informatics, University of Edinburgh, 2004.

[Caillault05] E.P.Caillault. Architecture et Apprentissage d'un Système Hybride Neuro-Markovien pour la Reconnaissance de l'Ecriture Manuscrite En-Ligne.these de doctorat. Université de Nantes 2005.

[Camillerapp02] J.Camillerapp, L.Pasquer, B.Coüasnon. Indexing old printed forms with handwritten last name. Rapport IRISA/Insa Rennes, 2002.

[Carbonnel05] S.Carbonnel. Intégration et modélisation de connaissances linguistiques pour la reconnaissance d'écriture manuscrite en-ligne. Thèse de doctorat. Université de Rennes, 2005.

[Cavnar94] W.B.Cavnar, J.M.Trenkle.: "N-gram-based text categorization". In Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval, 1994.

[Chatelain06] C.Chatelain. Extraction de séquences numériques dans des documents manuscrits quelconques. Thèse de doctorat Université de Rouen 2006.

[Coetzee03] D.Coetzee. Levenshtein distance. From the english wikipedia, United States. Version française (Bilbo 2004). 18 December, 2003.

[DEBORA99] Le projet DEBORA (Digital access to Books of the renaissance) a été présenté dans le cadre d'un appel d'offre de l'union européenne Télématique et Bibliothèque. Objectif : développer des outils permettant l'accès à des collections numérisées du XVIeme siècle.

[Djeddi09] C. Djeddi. Identification et Vérification de Scripteurs pour l'Ecriture Arabe. Mémoire de Magister en Informatique. Université de Tébessa, 2009.

[El-Hajj05] R. El-Hajj, L. Likforman-Sulem, C. Mokbel. Arabic handwriting recognition using baseline dependent features and Hidden Markov Modeling, ICDAR 05, Seoul, Corée du Sud, Août 2005.

[Falou98] W.Falou. Reconnaissance de caractères manuscrits par réseau de neurones. Thèse master en modélisation et ingénierie du logiciel scientifique. Université de Rennes, 1998.

[Farah05] N.Farah, A.Ennaji, T.Khadir, M.Sellami. Benefit of multiclassifier systems for Arabic handwritten words recognition. Proceedings of the 2005 Eight International Conference on Document Analysis and Recognition (ICDAR'05), 2005.

[Feldman07] R.Feldman. Categorization, Text Mining Handbook : Advanced Approaches in Analyzing Unstructured Data James Sanger Cambridge University Press, UK, 2007.

[Gumah08] M.E. Gumah, E. Schneider. Arabic Handwriting Recognition: Challenges and Solutions. Knowledge Management International Conference and Exhibition, 2008.

[Haou08] H. Haou, Z. Makhlouf, Catégorisation automatique de documents arabes manuscrits. Mémoire d'ingénieur, Université de Annaba, 2008.

[Heutte03] L.Heutte. Analyse et Reconnaissance de l'Ecriture: de Nouvelles Perspectives en Traitement Automatique de Documents Manuscrits. HDR Université de Rouen, 2003.

[Humbert02] G.Humbert. «Introduction», Revue des mondes musulmans et de la Méditerranée, n°99-100 - La tradition manuscrite en écriture arabe, p7-17, novembre 2002.

[Jaillet03] S.Jaillet, M.Teisseire, J.Chauche, V.Prince. Classification automatique de documents. In INFORSID, Nancy, 03/06/2003 - 06/06/2003. p.87-102, 2003.

[Joachims00] T.Joachims. The maximum-margin approach to learning text classifiers methods, theory and algorithms university of Dortmund. Thesis, 2000.

[Joher06] A.Z.Joher, F.Al-hajar. Kassem. Automatic Arabic Text Categorization With Bayesian Learning. Damascus University – Department of Artificial Intelligence, 2006.

[Jumari02] K.Jumari, M.A Ali. A survey and comparative evaluation of selected off-line arabic handwritten character recognition systems. Jurnal Teknologi, 36(E), Universiti Teknologi Malaysia, Jun 2002.

[Khreisat06] L.Khreisat. Arabic Text Classification Using N-Gram Frequency Statistics A Comparative Study. In proceedings of the international conference on data mining, Nevada, USA, pp. 78-82, 2006.

[Khurshid08] K.Khurshid, C.Faure, N.Vincent. Recherche de mots dans des images de documents par appariement de caractères. Actes du dixième Colloque International Francophone sur l'Ecrit et le Document CIFED, 2008.

[Kirtas01] Kirtas Technologies a été fondé en 2001 par Lotfi Belkhir autour de la volonté de transformer en format numérique l'immense connaissance contenue dans les rayons des bibliothèques, dans les archives publiques et d'entreprises, 2001.

[Lesk97] M.Lesk, M.Kaufmann. Practical Digital Libraries, Books, Bytes and Bucks, San Francisco, 1997.

[Likforman-Sulem03] L.Likforman-Sulem. Apport du traitement des images à la numérisation des documents manuscrits anciens. Rapport 2003.

[Likforman-Sulem06] L.Likforman-Sulem, A.Zahour, B.Taconet. Text Line Segmentation of Historical Documents: a Survey. Submitted to Special Issue on Analysis of Historical Documents, International Journal on Document Analysis and Recognition, Springer, 2006.

[Lorigo06] LM.Lorigo, V.Govindaraju. Offline Arabic Handwriting Recognition: A Survey. IEEE transactions on pattern analysis and machine intelligence, vol. 28, no. 5, 2006.

[MADONNE06] Consortium. MADONNE : MAsse de DOnnées issues de la Numérisation du patrimoiNE. In Atelier ANAGRAM'06, Fribourg, Suisse, 2006.

[MARA96] MARA, département de manuscrits arabes et ajamis (mara) de l'institut de recherches en sciences humaines, Université Abdou Moumouni, Niamey (NIGER), document publié par l'I.R.S.H., Sudanic Africa, 7, 165-169, 1996.

[Mermet08] J.M.Mermet. Techniques de Numérisation, Cours licence BDAN,IUT2 DocumentationDSI, Université de Grenoble, 2008-2009.

[Miled97] H. Miled, C. Olivier, M. Cheriet, K. Romeo-Pakker. Une Méthode Rapide de Reconnaissance de l'Ecriture Arabe Manuscrite. Seizième colloque gretsi, 15-19 septembre, Grenoble, 1997.

[Navarro01] G.Navarro. A Guided Tour to Approximate String Matching. ACM Computing Surveys, Vol. 33, No. pp. 31–88, March 2001.

[Nicolas06] S.Nicolas. Segmentation par champs aléatoires pour l'indexation d'images de documents. Thèse de doctorat Université de Rouen, 2006.

[Nosary02] A.Nosary, L.Heutte, T.Paquet. Reconnaissance de mots manuscrits par segmentation-reconnaissance : apports d'une reconnaissance lettres par niveau avec rejet. in Colloque International Francophone sur l'Ecrit et le Document, CIFED'2002, Hammamet, Tunisie, pp. 355-364, 2002.

[Paleo07] B.W.Paleo. An approximate gazetteer for GATE based on levenshtein distance. Proceedings of the Twelfth ESSLLI Student Session of the European Summer School in Logic, Language and Information (ESSLLI), p197-208, 2007.

[Parmentier 98] F. Parmentier. Spécification d'une architecture émergente fondée sur le raisonnement par analogie : Application aux références bibliographiques. Thèse de doctorat, Université Henri Poincaré Nancy1. 1998.

[Ramel06] J.Y.Ramel. Propositions pour la représentation et l'analyse de documents numériques. HDR Université François Rabelais de Tours, 2006.

[Rehel05] S.Rehel. Catégorisation Automatique de Textes et Cooccurrence de Mots provenant de Documents Non Etiquetés. Mémoire pour l'obtention du grade de maître sciences (M.Sc.), Faculté des études supérieures de l'Université Laval, 2005.

[SAPCCA04] SAPCCA : Sauvegarde du Patrimoine Culturel de Civilisation Ancienne, projet proposé et retenu lors de la troisième assemblée générale ordinaire de l'association CEMUR (Coopération Europe-Maghreb des Universités en Réseau), Sfax, Tunisie, 2004.

[Sari09] T.Sari, A.Kefali. A search engine for Arabic documents. Journées sur la Gestion Electronique de Documents & Réseaux de Recherche en Informatique GED'09 - Dicav, p165-174, les 20-21 Mai, Annaba, 2009.

[Sayre73] K. M. Sayre. Machine recognition of Handwritten words : A project report. Pattern Recognition, vol. 5, pages 213–228, 1973.

[Schadle04] I.Schadle, J.Y.Antoine, B.Le Pévédic, F.Poirier. SibyMot: Modélisation stochastique du langage intégrant la notion de chunks. In TALN, Laboratoire VALORIA, Université de Bretagne Sud (EA 2593), 2004.

[Sebastiani02] F. Sebastiani. Machine learning in automated text categorization. ACM Computing Surveys, 34(1)-47, 2002.

[Smeaton97] A.Smeaton, A.L. Spitz. Using character shape coding for information retrieval, Proc. 4th Intern. Conf. on Doc. Anal. & Recogn., IEEE Computer Society Press, pp.974–978, 1997.

[Spitz00] A. Lawrence Spitz, A.Maghbouleh. Text Categorization using Character Shape Codes. In Symposium on Electronic Imaging Science and Technology SPIE, pp174-181, 2000.

[Spitz95] A.L. Spitz. Using character shape codes for word spotting in document images", Dori D. and Bruckstein A. (Eds.), Shape, Structure and Pattern Recognition, World Scientific, Singapore, pp.382–389, 1995.

[Tan01] C.L. Tan, B. Yuan. Document text segmentation using multi-band disc model, Proc. of the SPIE Document Recognition and Retrieval VIII, pp.212-221, January 2001.

[Tappert90] C.C. Tappert, C.Y.Suen, T.Wakahara. The State of the Art in On-Line Handwriting Recognition. IEEE transitions on pattern analysis and machine lntelligence, Vol. 12, No. 8, August 1990.

[Vinciarelli01] A.Vinciarelli. A survey on off-line Cursive Word Recognition. Report in Pattern Recognition 35 1433–1446, 2001.

[Vinciarelli03] A.Vinciarelli. Noisy Text Categorization. Research Report, IDIAP 04-03, 2003.

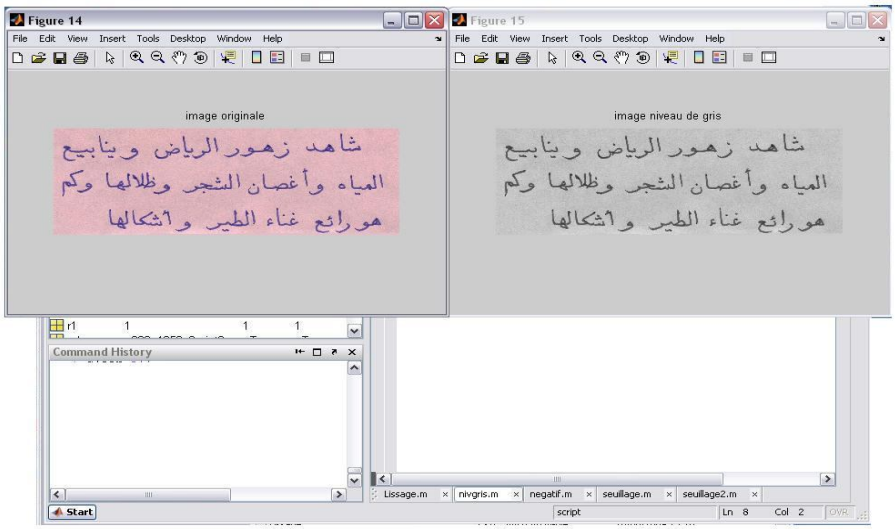
[Yadav09] S.Yadav , S.Sawarkar. Retrieval Of Information In Document Image Databases Using Partial Word Image Matching Technique. Proceedings of the international MultiConference of Engineers and Computer Scienntists, IMECS Vol1, Hong Kong, 2009.

[Yen04] S. Yen, Y. Chen, H. Lin, C. Wang. The Extraction of Text/Graphs from Degraded Documents. 10th International Multimedia Modelling Conference, pp.181, 2004.

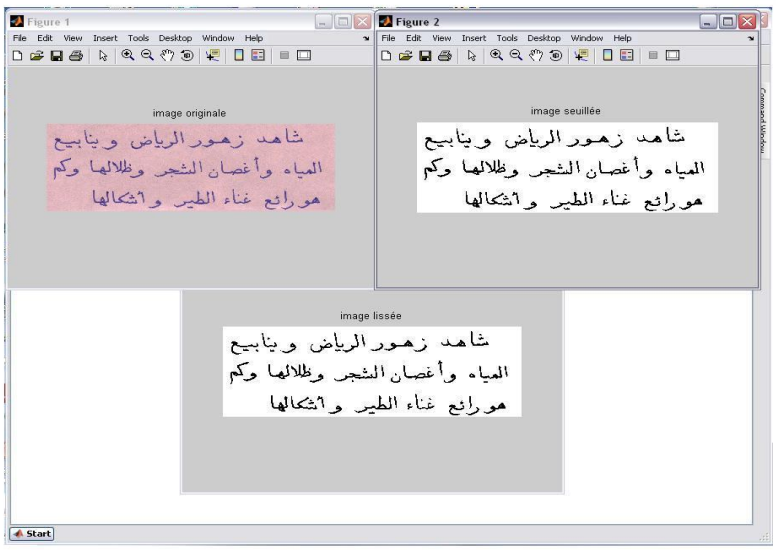
[Young98] I.T. Young, J.J. Gerbrands, L.J. van Vliet. Fundamentals of Image Processing. ISBN 90–75691–01–7, Printed in The Netherlands at the Delft University of Technology, 1998.

[Zaidi08] R. Zaidi, K. Zulkiflee, M. Idris, E. Tamil, M. Noorzaily, M. Noor, R. Salleh, M. Yaakob, M. Yaacob. Off-line Handwriting Text Line Segmentation: A Review. In IJCSNS, International Journal of Computer Science and Network Security, VOL.8 No.7, July 2008.

ANNEXE

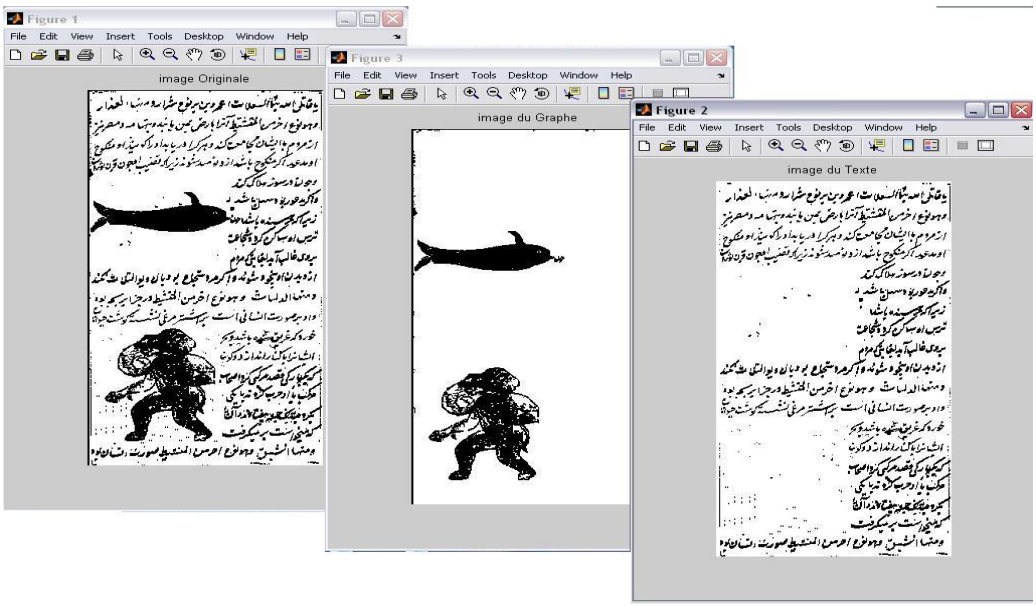


Prétraitements : Niveau de gris



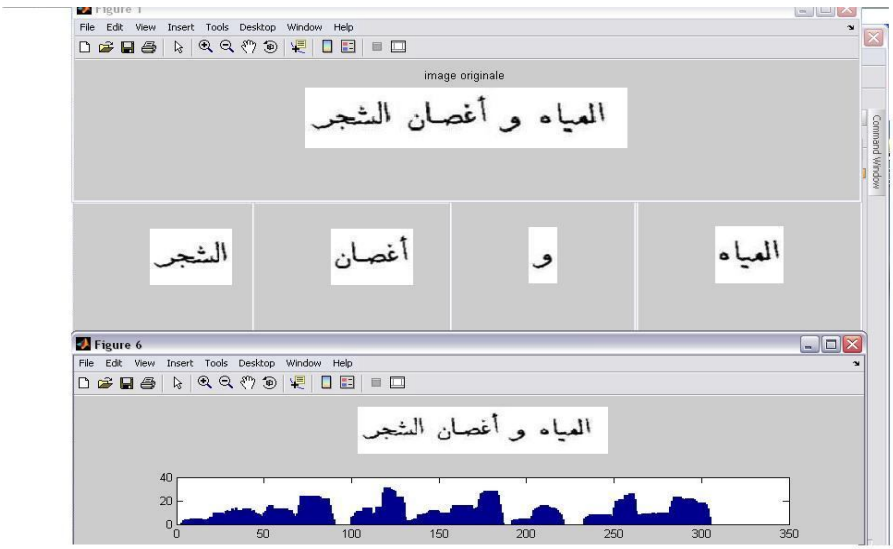
Prétraitements : Seuillage et lissage

Prétraitements : Seuillage adaptatif



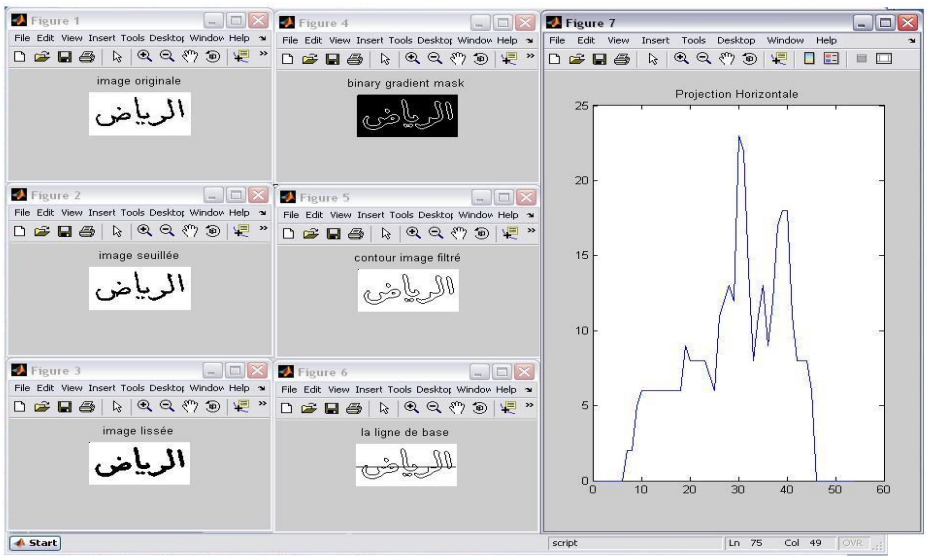
Segmentation : Séparation texte/graphique (seuil=400px)

Segmentation en lignes

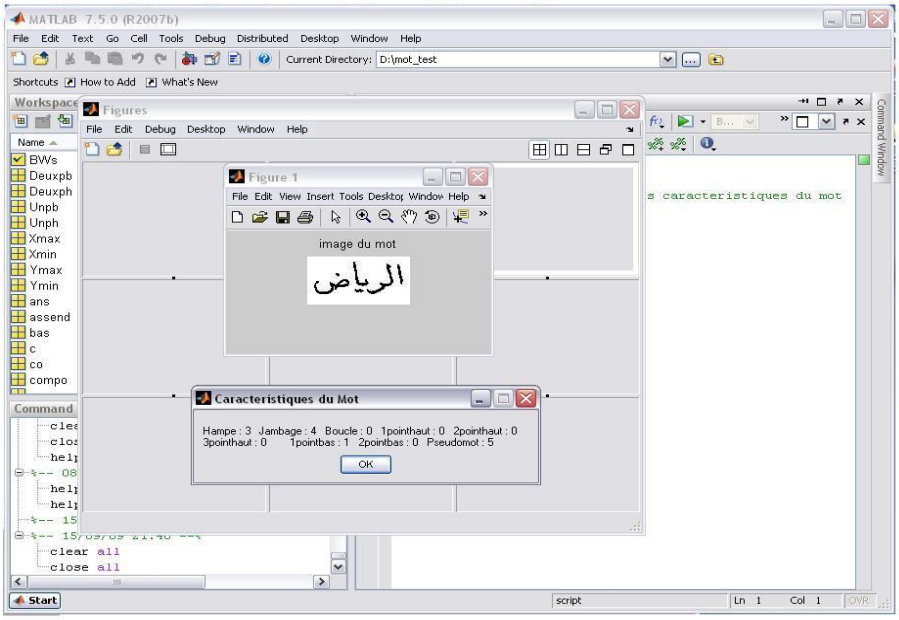


Segmentation en mots

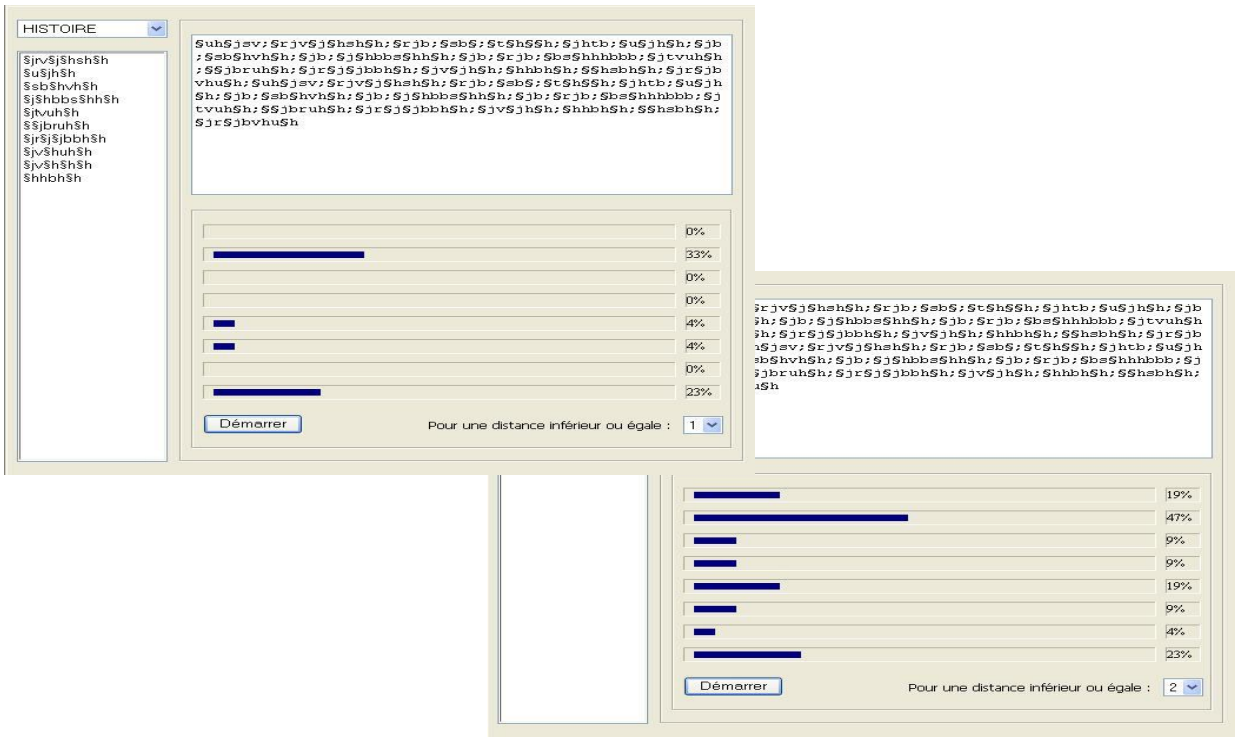
Opérations de morphologie mathématique



Traitements mot



Caractéristiques mot



Catégorisation d'un texte pour différentes distances