République Algérienne Démocratique et Populaire

Ministère de l'enseignement supérieur et de la recherche scientifique

Université de 8 Mai 1945 - Guelma -

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

Département d'Informatique

Mémoire de fin d'études Master

Filière : Informatique

Option : Ingénierie de médias

Thème :

Localisation du matricule dans les relevés de notes du BAC

Encadré Par :

Dr. Abderrahmane KEFALI

Présenté par :

Imane LATRECHE

Amel MESSAADI

Juin 2017

Abstract

The production of documents does not stop increasing, more than five thousand e-mails are sent every day. The company wonders how to keep its digital data. A solution adopted by the big companies and the enterprises is to replace the paper archiving by the digital archiving.

The work sent in this report joins in made in this context. We are interested in this work in the electronic archiving of the grade sheets of the high school diploma. In this report we propose how we can locate a very important information constituting this document which is **registration number of the student** to facilitate the digitalization of the archive and the information retrieval (example: research of a grade sheet).

We started with a physical study and visual analysis of various versions of these grade sheets of bac (from bac 1997 to bac 2016). This physical study has allowed us to understand the structure of these documents and afterward to propose a generic method for all formats. The proposed system bases on the extraction of registration number after a segmentation of the document by taking into account its physical structure. However, several stages of processing derived mainly from the document analysis domain, have been joined in the proposed system in order to reach the goal: preprocessing (gray levels transformation, Binarization), physical structure extraction (segmentation), logical labeling of registration number. Noting that the registration number is extracted in a sub-image form. Furthermore, and to test the performances of our system on a real application, we added a recognition module which receives the image of the registration number and produces a registration number in characters string form. This registration number serves as key of indexation and retrieval for the grade sheets in a database.

We also showed the effort which we realized to solve the problems involved in order to reach the aimed goals. Finally, we presented a complete and global development on the way of use of the proposed system. The proposed system is tested on a collection of images and the obtained results are encouraging.

Keywords: digital archiving, segmentation, documents analysis, structures recognition, physical and logical structure.

Résumé

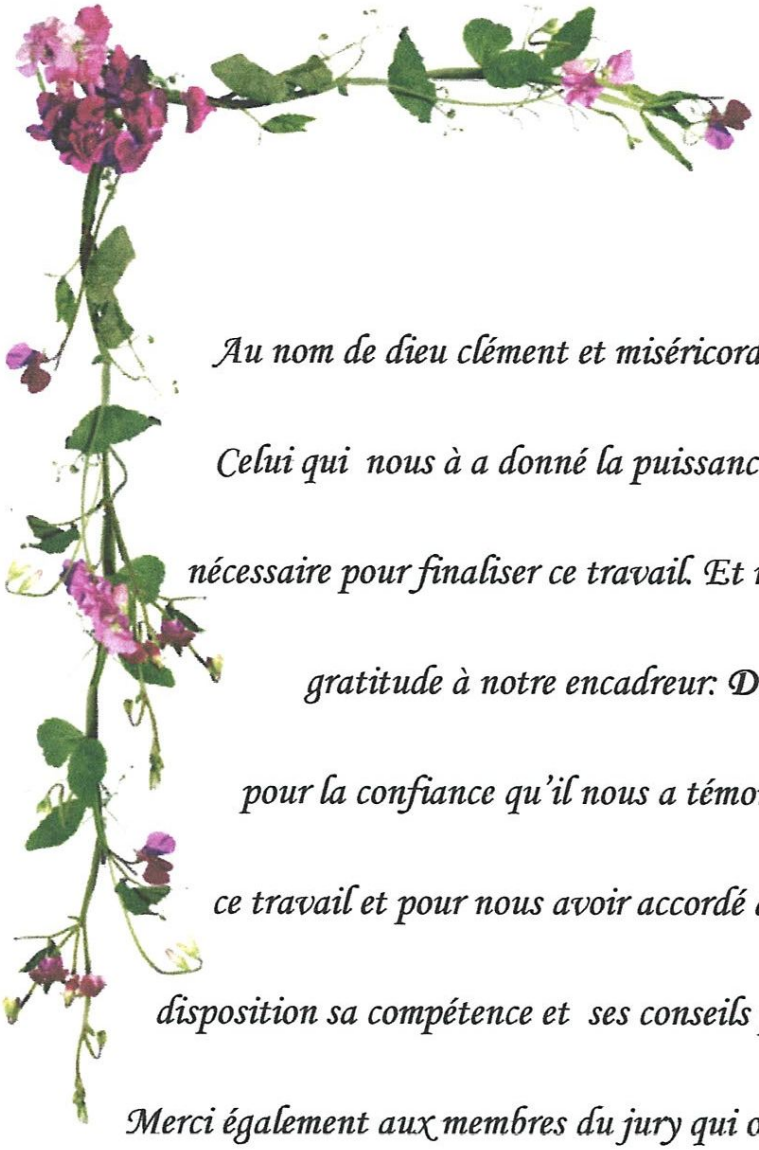
La production de documents ne cesse d'augmenter, plus de cinq mille mails sont envoyés chaque jour. La société se demande comment conserver ses données numériques. Une solution adoptée par les grandes sociétés et entreprises est de remplacer l'archivage papier par l'archivage numérique.

Le travail adressé dans ce mémoire s'inscrit dans ce contexte. Nous nous intéressons dans ce travail à l'archivage électronique des relevés de notes du baccalauréat. Dans ce mémoire nous proposons comment on peut localiser une information très importante constituant ce document qui est **le matricule de l'étudiant** pour faciliter la numérisation de l'archive et la recherche de l'information (exemple : la recherche d'un relevée de note).

Nous avons commencé par une étude physique et analyse visuelle des déférents versions de ces relevés de bac (de 1997 jusqu'à bac 2016). Cette étude physique nous a permet de comprendre la structure de ces documents et par la suite de pouvoir proposer une méthode générique pour toutes les formats. Le système proposé repose sur l'extraction du matricule après une segmentation du document en tenant compte de sa structure physique. Cependant, plusieurs étapes de traitement tirées principalement du domaine de l'analyse de documents ont été réunies dans le système proposé afin d'atteindre l'objectif visé : prétraitement (transformation en niveaux de gris, binarisation), extraction de la structure physique (segmentation), étiquetage logique du matricule. Notons que le matricule est extrait sous forme d'une sous-image. De plus, et en vue de tester les performances de notre système sur une application réelle, nous avons ajouté un module de reconnaissance qui reçoit l'image du matricule et résulte le matricule sous forme de chaine de caractères. Ce matricule serve de clé d'indexation et de recherche des relevées de notes dans une base de données.

Nous avons également montré l'effort que nous avons réalisé pour résoudre les problèmes rencontrés dans le but d'atteindre les objectifs visés. Enfin, nous avons présenté un développement complet, global sur la manière d'utilisation du système que nous avons proposé. Le système proposé est testée sur une collection d'images et les résultats obtenus sont encourageants.

Mots-clés : archivage numérique, segmentation, analyse de documents, reconnaissance de structures, structure physique et logique.



Remerciement

Au nom de dieu clément et miséricordieux, le grand merci lui revient,

Celui qui nous à a donné la puissance, le courage, et la détermination nécessaire pour finaliser ce travail. Et nous voudrons exprimer toute notre gratitude à notre encadreur: ***Dr. Abderrahmane KEFALI***

pour la confiance qu'il nous a témoigné en acceptant de diriger ce travail et pour nous avoir accordé de son temps et avoir mis à notre disposition sa compétence et ses conseils pour un meilleur maitrise du sujet.

Merci également aux membres du jury qui ont accepté d'évaluer notre travail.

Tous les enseignants du département d'informatique qui ont contribué à notre formation, et pour leur précieux conseils.

Nos collègues de notre promotion, nos familles et toutes les personnes qui nous ont aidé et soutenu de prés ou de loin tout le long de ce travail.

Imane & Amel.

Dédicace

Je commence mes dédicaces au nom de Dieu et puis de son prophète Mohamed.

Je tiens à dédier ce modeste travail

Aux êtres les plus chers à mes yeux qui m'ont soutenu durant toutes mes études à savoir :

mes très chers parents. Je pris le bon dieu de les garder près de moi pour fleurir le chemin de ma vie.

Mon cher époux Med.Amine, l'âme de ma vie, ce lui qui a été toujours à coté de moi toute cette année.

Mon beau père, ma belle mère, et mes belles sœurs : Amel, Hadjer. Et mon frère : Brahim.

Je dédie ce travail à ma chère grande sœur Selma et son mari Adel et leurs petits fils, mes chers : Wassim et Nazim.

A mes chers frères Hlchem, Mohamed Amine.

A tous mes amis : Amel, Manel, Aicha, Chaima.

Mes chères cousines : Zyneh, Rokia, Anfel, Assia, Houda, Noura, Boutaina.

A tous ceux qui me sont chers je dédie ce modeste travail.

Imane.

Dédicace

Je commence par rendre grenade à DIEU et à sa bonté pour la patience, la compétence et le courage qu'il m'a donné pour arrive à ce stade.

Je dédie ce mémoire :

À ma mère Meriem qui ne cesse jamais de m'encourager, à mon père Salah qui était toujours à mes côtés à tout moment, Que dieu tout puissant les garde pour moi,

À mes sœurs Nadjiba, Wided, Zahra, Karima, Imen, Khadidja, et Maroua pour leur amour et leur Soutien inconditionnel.

À toute mes deux grandes familles : la famille Messaadi et Kacha à tous leurs enfants : Mohamed, Akram, Islam, Wefia, Sewsen, Mayseme, Rimasse, Ziyade, Mouaad, Louaay, joumana, mohamed amine, israa.

Mes chères cousines :Youssra, Mayssoune, Bouchra, Rima, Lilya, Ranya, Wafa, Ilhaame.

À tous mes amies : ma binôme Imane, et radja, manel, Ranya, Khadidja, Radia, aicha, Majda, Khawla, asma et la chère qui est toujours dans mon esprit et ma mémoire Dr.Khawla Bouaffia.

Et bien sur sans oublier mon marie HAMZA qui toujours m'a aidé soit moralement ou bien physiquement.

A mes collèges de ma promotion et à toute personne qui m'a aidé et soutenue de près ou de loin dans ce travail.

Amel.

Table de matière

Table de figures

Chapitre I

Chapitre II

Chapitre III

Chapitre IV

Liste des tableaux

Chapitre I

Chapitre IV

Introduction générale

Les nouvelles technologies de transmission et de conservation de l'information sont la grande innovation des temps modernes. En effet, après plusieurs années de tâtonnements, la dématérialisation[1] fait aujourd'hui partie du quotidien du citoyen contemporain. Elle se généralise pour tous les domaines de la vie des entreprises, des autorités administratives, et même de celle des particuliers.

Pour cela l'archivage électronique est incontournable pour sauvegarder les données importantes d'une entreprise. En effet, il est difficile de garder la forme originale de document à cause de son état de stockage, comme les taches souvent dues à l'humidité, mais peuvent aussi être la cause de mauvais usages ; Les pliures et ondulations peuvent avoir plusieurs origines : par exemple, le coin d'un document peut être plié ; la reliure peut aussi créer une ondulation locale sur une partie du document.

Grâce à l'archivage électronique, la sécurisation des données devient moins difficile et plus fiable.

Dans la plupart de temps les documents à archiver comportent plusieurs informations, non seulement le texte ou les caractères écrits, mais également le type et la taille de police utilisée, la couleur de l'écriture, en plus d'autres informations supplémentaires décrivant l'organisation et la structuration des différents éléments des documents. Sans ces informations supplémentaires apportées par la structure contenue dans un document, la lecture ou la localisation correcte de ce dernier serait impossible. La structure du document a plusieurs rôles importants autres que la lecture, car elle porte aussi une information sur le contenu du document. Ce dernier est responsable de la traduction de la fonction du texte et de l'intention de son auteur en même temps.

De ce fait, la compréhension d'un document nécessite la reconnaissance de sa structure en plus de son contenu textuel puis on peut localiser n'importe qu'elle information constituant ce document.

L'objectif du présent travail s'inscrit dans cette démarche. Nous nous intéressons à analyser la structure d'un type particulier de documents (relevée de notes du BAC), afin de localier une information importante constituant ce document, à savoir la matricule de l'étudiant. L'objectif de notre travail à long terme est la construction d'un système grand système d'archivage

[1] « Dématérialiser » un document consiste à transférer, souvent grâce à la numérisation, un document d'un support papier à un support informatique. La numérisation peut être suivie de reconnaissance de documents et de caractères.

électronique des relevés de BAC qui intègre plusieurs fonctionnalités, acquisition, compression, prétraitements, analyse et reconnaissance, recherche, etc. Ce système permettra sans doute de faciliter le travail des agents dans les services de scolarité et d'archive de l'université.

Le présent mémoire est organisé en quatre chapitres comme suit :

Chapitre I: Etat de l'art sur l'archivage électronique

Ce chapitre est consacré à la présentation de l'archivage électronique, son rôle et ses avantages. Il expose les différentes étapes de cycle de vie d'un archive et la différence entre ce dernier et la numérisation.

Chapitre II: analyse de document

Le deuxième chapitre introduit la reconnaissance de la structure physique de document. Il survole les différentes approches de reconnaissance de la structure physique.

Chapitre III: Conception du système

Le troisième chapitre détaille les différentes étapes suivies afin d'acquérir un système permettant l'analyse et la localisation du matricule de l'étudiant.

Chapitre IV: Implémentation et résultats

Ce chapitre illustre l'implémentation de notre application, la collection de test et les résultats obtenus.

Chapitre I

Etat de l'art sur

l'archivage électronique

I.1. Introduction

Le développement des technologies de l'information et de la communication a profondément modifié les méthodes de travail en facilitant et en accélérant considérablement la production, le partage et le stockage d'informations numériques. En parallèle, la reconnaissance de l'écrit électronique a ouvert la voie à « l'administration électronique », à la dématérialisation des processus métier et à la production d'originaux numériques.

C'est pourquoi l'archivage numérique/électronique est devenu un véritable enjeu pour les directions des systèmes d'information.

Ce chapitre est consacré à l'archivage électronique de document. Nous présentons tout d'abord c'est quoi l'archivage et pourquoi archiver, Puis nous expliquons les Processus traditionnels d'archivage d'un document, et nous allons détailler les différentes étapes de cycle de vie d'un archive. Par la suite on va définir l'archivage électronique et la différence entre ce dernier et la numérisation. Ensuite, nous allons donner une définition sur les systèmes d'archivage électronique avec quelques objectifs, avant de conclure.

I.2. Notion de l'archivage

I.2.1. Définition de l'archivage

Dans les dictionnaires de l'information, l'archivage est défini comme « l'ensemble des méthodes, processus et outils mis en œuvre pour gérer et conserver les documents qui ont cessé d'être d'utilité courante » [COL 04].

Une autre définition intéressante est la suivante : « L'archivage est l'action de mettre en archive, d'archiver .Employé surtout à l'origine pour les seuls documents électroniques, comme un synonyme de stockage ou de sauvegarde, il tend de plus en plus à être utilisé pour tous les documents, quels qu'en soient la nature et le support, et à remplacer « conservation » » [WEB 1].

I.2.2. Pourquoi archiver?

Tout d'abord, l'archivage permet aux entreprises, de garder une trace des activités, les informations archivées pourront ou pourraient être réutilisées pour expliquer ou justifier quelque chose. Comme l'explique Marie-Anne Chabin [CHA 10], l'archivage n'est pas utilisé pour une unique raison, mais pour plusieurs telles que :

la maintenance des documents pendant la période souhaitée (conservation sécurisée) ; et la sortie du système (destruction). S'y ajoute le sous-processus de mise à disposition tout au long du cycle de vie [CHA 10].

Les principales étapes de ce processus sont :

- **Identifier les documents d'archives :** La pertinence et l'authenticité de chaque document d'archives doivent pouvoir être établies à chaque étape de sa vie, en s'assurant qu'il n'a pas subi d'altération volontaire ou accidentelle de son contenu.
- **Stocker les documents d'archives :** Les documents d'archives doivent être stockés de telle façon qu'ils soient à la fois suffisamment accessibles et protégés des dégâts environnementaux. Un document classique, sur papier ordinaire, peut être rangé dans un bureau, dans un placard à archives. Mais un local d'archives digne de ce nom doit être conforme à des normes environnementales spécifiques, notamment en matière de température et d'hygrométrie.
- **Faire circuler les documents d'archives :** Les documents d'archives sont conservés pour que l'on puisse les retrouver au moment voulu. Retrouver le document, le suivre pendant qu'il est hors du local d'archives puis le récupérer : ce sont les étapes de la circulation.
- **Le sort final de documents d'archives :** Le sort final de documents d'archives n'est pas toujours la destruction. Il peut aussi prendre la forme d'un versement dans un service d'archives historiques, un musée ou une bibliothèque. [WEB 2].

I.2.4. Cycle de vie d'un archive

Le cycle de vie d'un archive est illustré par la figure suivante :

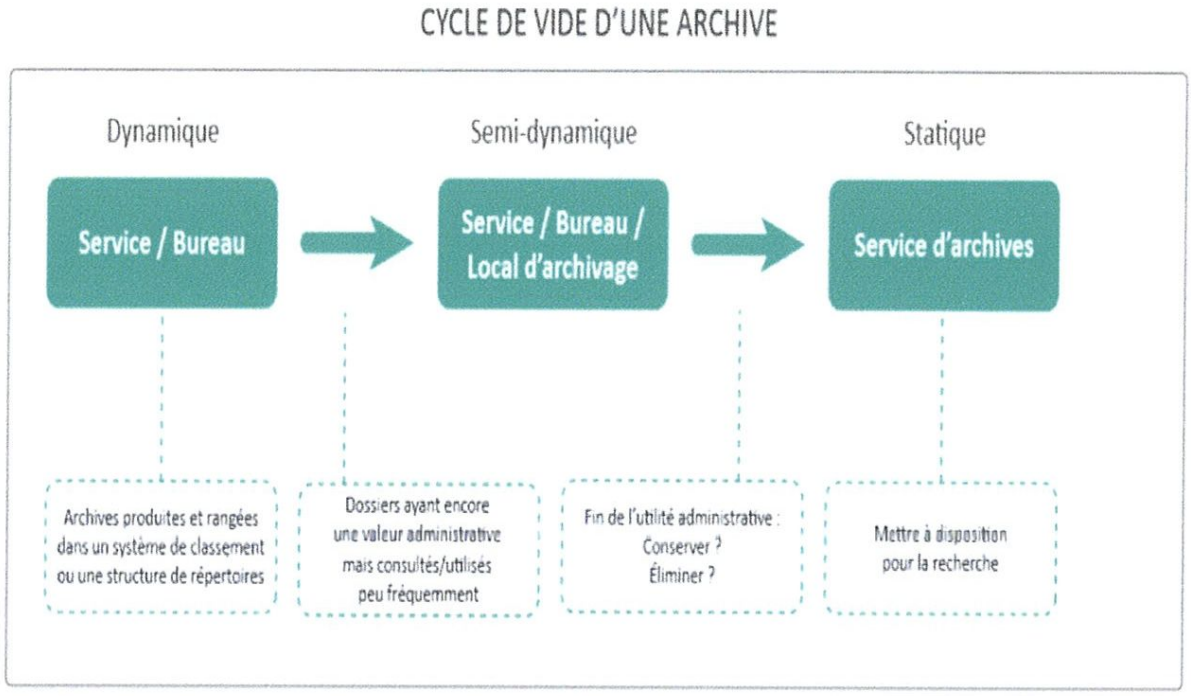


Figure I.1: Le cycle de vie d'un archive [SOY 11].

a) Les documents dynamiques et semi-dynamiques :

Ces documents sont ceux qui se trouvent dans leur phase d'utilité administrative. Ils sont classés au moment même de leur production en vue d'une utilisation immédiate ou à moyen terme; le classement est « actif », de nouveaux documents viennent sans cesse s'ajouter.

b) Les documents statiques :

Les documents ayant perdu leur utilité administrative sont définitivement classés. C'est à cette étape qu'ils seront triés sur base des recommandations de tri des Archives de l'État ; une partie d'entre eux pourra être éliminée ou transférée vers un centre ou un service d'archives

I.3. L'archivage électronique

I.3.1. Définition de l'archivage électronique

L'archivage électronique est l'ensemble des actions, outils et méthodes mis en œuvre pour réunir, identifier, sélectionner, classer et conserver des contenus électroniques, sur un support sécurisé, dans le but de les exploiter et de les rendre accessibles dans le temps, que ce soit à titre de preuve (en cas d'obligations légales notamment ou de litiges) ou à titre informatif. La conservation est l'ensemble des moyens mis en œuvre pour stocker, sécuriser, pérenniser, restituer, tracer, transférer voire détruire, les contenus électroniques archivés [WEB 3].

En comparaison avec l'archivage papier, l'archivage électronique gère les documents numériques, qui sont des « ensembles composés d'un contenu, d'une structure logique, d'attributs de présentation permettant leur représentation, dotés d'une signification intelligible par l'homme ou lisible par une machine ». Il peut être créé à l'état natif ou obtenu par un processus de transformation d'un document physique, par exemple par numérisation. Les documents bureautiques, les bases de données, les messages électroniques, les dossiers numérisés sont considérés comme des documents numériques.

Figure I.2: Les déférents formats de l'archivage électronique [SOY 11].

I.3.2. Utilité de l'archivage électronique

L'archivage électronique a pour objectif de faire en sorte que les documents électroniques remis à des Archives, qu'ils soient d'origine publique ou privée, restent durablement compréhensibles, mais aussi de garantir leur authenticité et de les rendre accessibles. Par durablement on entend une durée illimitée, au minimum plusieurs générations de matériel et de logiciels informatiques [WEB 4]. L'archivage électronique permet ainsi de :

- Faciliter l'accès à l'information,
- Répondre aux exigences légales de conservation et de communication,
- Relever le défi de l'obsolescence technologique récurrente.

I.3.3. Archivage électronique vs numérisation

I.3.3.1. Qu'est-ce que la numérisation ?

La numérisation de documents consiste à transformer un document physique en fichier électronique (***Figure I.3***).

Ainsi, l'objectif final de la numérisation est de faciliter le traitement du contenu : Reconnaissance ; Copie ; Stockage ; Indexation ; Recherche.

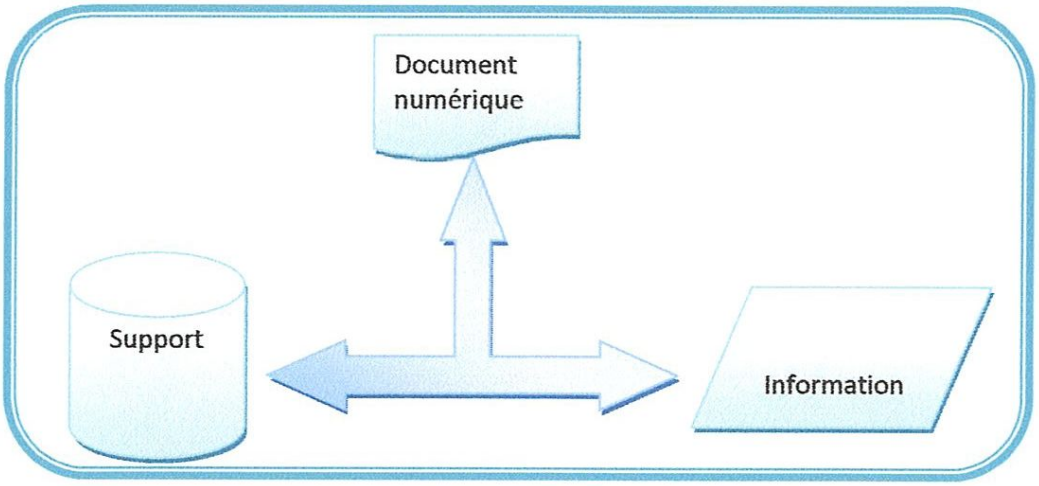


Figure I.3: Le document numérique dissocie le support et l'information contenue [WEB 5].

I.3.3.2. Différence entre l'archivage électronique et la numérisation

L'opération de numérisation consiste à reproduire le document original avec une qualité suffisante sous forme de document électronique pour une conservation et une communication à long terme. Le document original est conservé. Or l'archivage électronique n'est pas une sauvegarde électronique. La sauvegarde informatique est une copie de sécurité d'un ensemble d'informations électroniques dans le but de se prémunir contre les incidents, les pertes ou les vols... Sa durée de vie est limitée et son support est inexploitable en dehors de son environnement technique. En France, le terme « numérisation » est généralement conçu comme le procédé de conversion d'un document d'un support classique vers un support électronique. Tandis que ce terme exprime une autre chose que l'archivage électronique, les pays non-francophones n'ont pas de termes correspondants pour ces deux mots.

Contrairement aux pays africains, les pays européens connaissent déjà des définitions des documents électroniques, la numérisation, l'archivage électronique…etc. Qui dérivent souvent de la législation. Ces définitions sont plus ou moins semblables [ABB 09].

I.3.4. Archivage électronique vs la GED / GEIDE

L'archivage n'est pas, non plus, la Gestion Electronique d'Informations et de Documents pour l'Entreprise (GEIDE ou GED) car celle-ci autorise la modification des documents électroniques.

La GED : les systèmes de gestion électronique de documents « GED » sont largement utilisés pour la gestion et le contrôle des documents électroniques.

Contrairement à l'archivage électronique, permet la modification de documents et la coexistence de plusieurs versions d'un même document par leurs propriétaires. Elle peut également permettre la destruction des documents.

Le tableau suivant présente les différences existantes entre un GED et un Archivage électronique :

Différences GED / Archivage électronique	
Permet la modification des documents et la production de plusieurs versions	Empêche la modification des documents
Permet la destruction des documents par leurs auteurs	Peut comporter la gestion de délais de conservation
Empêche la destruction de documents en dehors d'un contrôle strict soumis à validation	Comprend obligatoirement un contrôle rigoureux des délais de conservation
Peut comprendre une structure organisée de stockage sous le contrôle des utilisateurs	Comprend obligatoirement une structure rigoureuse de classement pour la conservation et le stockage, gérée par l'administrateur du système (selon un plan de classement des activités)
Est a priori dédié à la gestion quotidienne des documents pour la conduite des affaires.	Peut faciliter les tâches quotidiennes mais est surtout destiné à la constitution d'un fonds sécurisé des documents probants de l'entreprise.

Tableau I.1: Les différences existantes entre un GED et un Archivage électronique [WEB 6].

I.3.5. Archivage électronique vs Archivage traditionnel

Le tableau suivant présente les différences existant entre l'archivage électronique et l'archivage traditionnel:

Critères	Archivage traditionnel	Archivage Électronique
Pérennité	Qualité des supports et conservation d'un exemplaire unique.	Écritures en multiples exemplaires, utilisation de formats informatiques non propriétaires, etc.
Intégrité	Méthodes de protection des objets (en limitant leurs sorties).	Catalogue des objets conservés, outils permettant de détecter toute modification des objets conservés.
Sécurité	Contrôle des accès, protection des locaux et de leur contenu (contre l'incendie, les dégâts des eaux, les nuisibles, etc.).	Contrôle des accès physiques, protection des locaux (contre l'incendie, les dégâts des eaux, etc.), gestion des droits d'accès informatiques, administration du système, réplications, sauvegardes des systèmes, etc.
Traçabilité	Journal des évènements.	Journal des évènements.
Authenticité	Signature et date.	Signature électronique, horodatage, calcul et gestion d'empreintes, etc.
Lisibilité / Intelligibilité	Implicite. Attention, certains documents peuvent s'estomper avec le temps (carbone, papiers chimiques, etc.).	Dispositifs matériels (lecteurs), formats de stockage, métadonnées spécifiques.
Disponibilité	Organisation des moyens et des ressources.	Organisation des ressources, plan de continuité, solutions de back-up, Plan de Reprise d'Activité.

Tableau I.2: Différences entre l'archivage électronique et l'archivage traditionnel [GRO 10]

I.4. Les systèmes d'archivage électronique

I.4.1. Définition

Un système d'archivage électronique « SAE » est un outil informatique permettant la conservation pérenne et sécurisée des documents électroniques. Une fois intégré dans un SAE, un document n'est plus modifiable et conserve donc sa valeur probante [SAE 14].

I.4.2. Fonctionnalités d'un SAE

Lorsque les archives sont trop nombreuses, les entreprises ou les collectivités sont amenées à investir dans un SAE. Le SAE donne en temps réel et avec précision :

- Le contenu des archives ;
- L'emplacement de chaque document ;
- Les dates de création, d'utilisation et d'archivage de chaque document ;
- Les mouvements de consultations des archives.

Il permet donc de :

- Rechercher un document d'archive ;
- Obtenir des statistiques ;
- Sécuriser le contenu des archives.

I.4.3. Objectifs d'un SAE

Les quatre objectifs d'un SAE sont les suivants (***Tableau I.3***):

Objectifs	Définitions	Comment
Pérennité	Assurer la lisibilité du document dans le temps.	- Vérification régulière de d'état et de l'exploitabilité des données. - Migration des données vers de nouveaux supports et formats. - Information sur le format du document pour en permettre la restitution indépendamment de son environnement d'origine (inclus dans les métadonnées). - Conservation des données dans un format standard pérenne (ex : PDF, XML).
Intégrité	Assurer l'authenticité du document c'est-à-dire la non modification possible du contenu et de la forme	- Vérification de l'identité de l'auteur par la signature électronique. - Horodatage (association d'une date et d'une heure au document électronique permettant de le sceller). - Vérification du non modification du document par la gestion des métadonnées. - Traçage de toute modification d'une donnée/document dans un journal des évènements.
Traçabilité	Déterminer l'historique du cycle de vie du document : description de toutes les opérations effectuées (consultations, migrations, etc.)	- Horodatage. - Enregistrement des transactions relatives aux documents dans un journal des évènements horodaté et signé pendant toute la durée de conservation
Sécurité	Assurer la conservation du document à long terme. Limiter la communication suivant la législation en vigueur.	- Définition des règles d'accès suivant la communicabilité (dans le temps) et les utilisateurs. - Duplication du document. - Conservation dans des coffres distants ou sur serveur sécurisé.

Tableau I.3: Les objectifs d'un SAE [SAE 14].

I.4.4. Avantages apportés par les SAE

En effet l'archivage, en conversant l'information à long et moyen terme, permet d'assurer sa continuité dans le temps ainsi que son exploitation ultérieure. La tendance actuelle vers la dématérialisation, poussent les entreprises à remplacer leurs supports traditionnels d'archivage (papier en autre ...) par des SAE (système d'archivage électronique). Les entreprises ont pour assurer leurs SAE. Soit en « interne » avec la création de leur propres services d'archivages, soit en « externe » en confiance leurs données à un tiers archiveur.

I.4.4.1. Les avantages d'un SAE dans l'entreprise

Dans l'entreprise certaines informations peuvent revêtir un caractère stratégique. Ainsi un SAE permet de mettre en place différents moyens de chiffrement ou signature électronique garantissant ainsi la confidentialité de l'information stockée.

La dématérialisation enduite par la mise en place d'un SAE permet un gain de place significative dans l'entreprise (et donc une réduction de charge). Ainsi même si l'évolutivité des supports d'archivages est à prévoir dans le coût de mise en place d'un SAE, ce dernier reste une solution plus attractive pour l'entreprise qu'un archivage classique.

Enfin un SAE permet de faciliter grandement la consultation des donnés archivées en optimisant le délai d'accès, la fréquence d'accès, le délai de restitution...etc. De plus l'insertion des documents dans un SAE est également facilité grâce notamment à l'intégration d'un plan de classement permettant de structurer automatiquement le contenu déposé sans l'intervention d'une personne physique.

I.5. Conclusion

L'archivage consiste à conserver, à moyen ou long terme des informations afin de pouvoir les exploiter ultérieurement. L'archivage permet ainsi d'assurer la fidélité et la durabilité de l'information conservée. En raison du développement considérable des données sous forme dématérialisée, les supports traditionnels d'archivage des documents (papier ou microforme) sont progressivement remplacés par des systèmes d'archivage électronique.

Donc, après cette étude théorique, on peut dire que l'archivage numérique est la meilleure solution pour une gestion bien organisée. Il reste à bien exploiter les documents numérisés en faisant un travail d'extraction ou de reconnaissance de la structure et du contenu de ces documents.

Chapitre II

Analyse de documents

II.1. Introduction

Il y a quelques années encore, le papier était le seul support d'enregistrement, de diffusion et de conservation de l'information. L'archivage était une étape finale dans le cycle de vie des documents.

Cependant, le développement des technologies de l'information et de la communication a totalement révolutionné les modes de production, de diffusion et de conservation de l'information numérique.

L'archivage électronique commence dès la création des documents, ce qui a donné naissance à un champ de recherche important dans le domaine du traitement d'images, à savoir l'analyse des documents.

L'analyse des documents ou la reconnaissance de la structure physique des documents est un processus complexe qui sert à extraire la structure physique afin de déterminer tous les régions homogènes dans le document (texte, graphique...) et sa structuration hiérarchique (paragraphe, mot, lettre..). L'analyse des documents reste un champ très vaste et très compliqué, dans le domaine du traitement d'images, et malgré les efforts des chercheurs, plusieurs difficultés sont encore ouvertes.

Ce chapitre est consacré à l'analyse ou la reconnaissance de la structure physique d'un document. Nous présentons tout d'abord la notion dc document et de structure. Ensuite, l'analyse des documents, ses objectifs, et les étapes qui le composent, ainsi que les différentes approches de reconnaissance de la structure physique, avant de conclure.

II.2. Notion de document

II.2.1. Définition

Plusieurs articles scientifiques abordent des problématiques liées aux documents, mais peu d'entre eux tentent de donner des définitions génériques pour ce terme. Nous présentons dans cette section quelques définitions génériques:

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

- Une définition plus générale donnée par Karim HADJAR [HAD 06] dans sa thèse de doctorat indique qu'un document peut avoir plusieurs types (textuel, sonores, vidéo, graphique...etc.), selon le support choisi. Pour lui, « Un document est le support physique pour conserver et transmettre de l'information ».
- Le document peut aussi être défini par tout ce que l'on réalise, utilise ou transfert lors d'un processus de communication écrit ou électronique [AZO 95].

II.2.2. Document électronique

Partant des définitions de Bachimont et de Hadjar présentées précédemment, un document est dit numérique (électronique) lorsque le support physique est numérique.

« Un document électronique est un document représenté sous la forme d'une structure de données ou d'une séquence d'octets stockée en mémoire ou sur un support informatique qui peut être transmis entre les ordinateurs » [HAD 06].

II.3. Document et structure

Décrire la structure d'un document consiste à identifier et décrire chacun des éléments textuels - ou non textuels - qui le constituent. Ceci dit, cette description peut prendre plusieurs formes. En effet, on distingue, en général, deux types de structure : la structure physique et la structure logique [SOU 02].

La structure physique définit la présentation du document sur le support (le papier) tandis que la structure logique définit une organisation hiérarchique de l'information contenue dans lui.

Plusieurs autres structures peuvent être définies afin de répondre à certains besoins de traitement et d'accès aux documents. Nous pouvons trouver [SOU 02]:

- **Structure conceptuelle :** constitue le contenu du document issu de sa conception.
- **Structure concrète :** regroupe les deux structures physique et logique.
- **Structure fonctionnelle :** représente le document décomposé en un ensemble de composants étiquetés selon leur fonction.

II.3.1. La structure physique d'un document

La structure physique est définie par l'organisation hiérarchique des blocs (qui servent à structurer l'aspect graphique d'un document) composant les pages d'un document (entête, bloc, contenu, colonne, ligne...) [AZO 95]. Elle peut également être définie par la typographie et l'organisation du document. La typographie définit le style des éléments graphiques

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

La Figure suivante (***Figure II.1***) présente un exemple d'une structure physique d'une page d'un article scientifique, exprimée sous forme d'arbre :

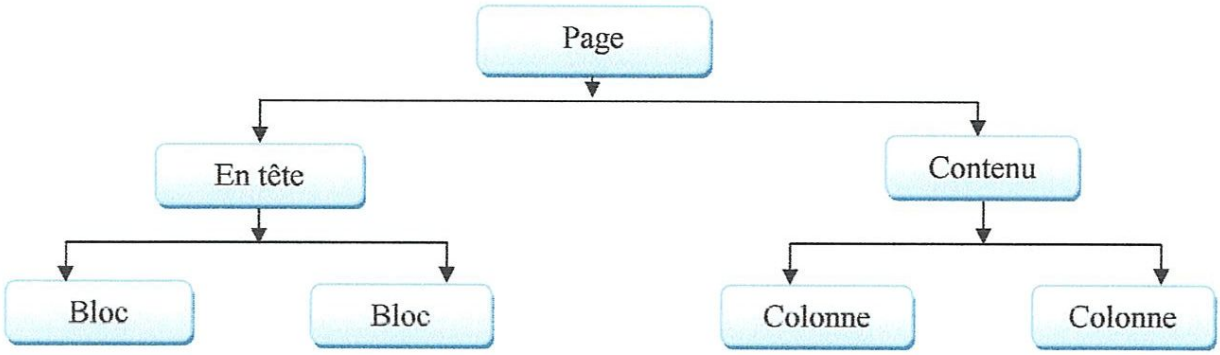


Figure II.1: Exemple de structure physique [HAD 06].

II.3.2. La Structure logique d'un document

La structure logique s'intéresse à décrire le rôle et la nature de chaque élément d'un document ainsi que l'ensemble des liens hiérarchiques et/ou logiques qui les lient les uns aux autres. La figure suivante (***Figure II.2***) illustre un exemple d'une structure logique correspondante à la structure physique de la figure précédente :

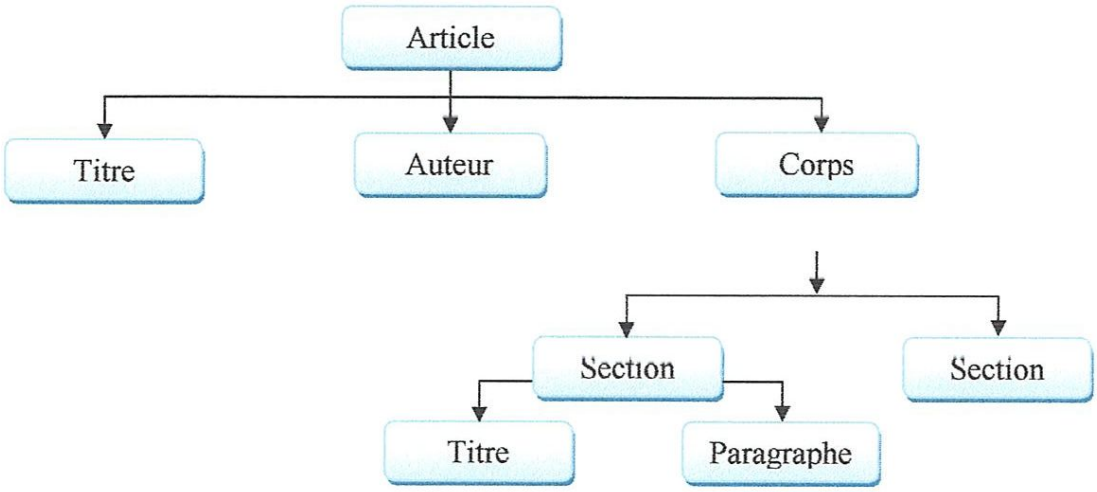


Figure II.2: Exemple de structure logique [HAD 06].

II.3.3. Production de document

Un document imprimé est le résultat d'un processus de production en plusieurs étapes. La première étape est la saisie du contenu, Si l'édition est structurée, elle aboutit à la forme logique du document qui contient des éléments textuels ou graphiques aux quels on a associé des étiquettes logiques comme « titre », « liste » ou « tableau ».

La deuxième étape est une transformation de la forme logique en forme physique appelée formatage. La forme physique ne contient plus d'étiquettes, le sens véhiculé par les étiquettes est traduit en attributs typographiques et dans la mise en page.

La restitution, l'étape suivante, transforme la forme physique du document en une image. En fin le processus peut se terminer par l'impression afin d'obtenir un document imprimé.

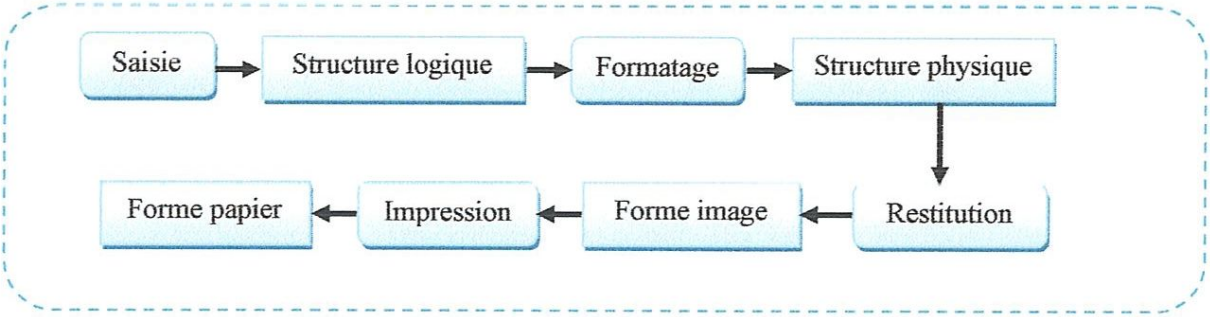


Figure II.3: Etapes de production et les différentes formes d'un document [HAD 06].

II.4. L'analyse de documents

L'analyse de documents ou l'analyse d'images de documents, ou encore la reconnaissance de la structure physique de documents est un thème de recherche du domaine du traitement d'images qui avait pour objectif principal de convertir des images de documents en vue de la modification, l'archivage, la recherche, la réutilisation et la transmission de l'information que ces images contiennent [TRU 05].

II.4.1. Définition

Nagy [NAG 00] propose la définition suivante de l'analyse de documents: « L'analyse d'images de documents : est une théorie et une pratique de reconstruction de la structure symbolique des images numériques directement produites par l'ordinateur ou simplement numérisées à partir du papier ».

Une autre définition plus précise est celle décrite dans [KET 10]. Ainsi [KET 10] définit l'analyse de documents comme « l'analyse de la mise en page pour trouver la structure

physique. Il s'agit de segmenter l'image de document en composantes homogènes et de classifier chaque zone en texte, image, graphique, etc. ».

II.4.2. Objectif de l'analyse de documents

L'analyse d'images de documents a pour but de découper l'image du document en régions ou blocs homogènes (texte, graphique, photographies...) puis, en lignes de texte et en caractères à l'intérieur de ceux-ci. Ces éléments constituent la structure physique du document.

II.4.3. Analyse de documents vs reconnaissance de documents

La reconnaissance de documents est le processus inverse de la production de la forme papier, elle essaie de remonter à la forme logique. Le document papier est saisi à l'aide d'un scanner de manière à obtenir une image sous la forme électronique, c'est à dire une matrice de pixels avec des méta-informations renseignant sur l'interprétation des pixels (couleur, résolution).

Cependant, la reconnaissance de documents est un processus qui inclut plusieurs étapes de traitement y compris la reconnaissance de la structure physique ou l'analyse de documents.

La figure suivante (***Figure II.4***) illustre la position de l'analyse de documents dans le processus de reconnaissance de documents :

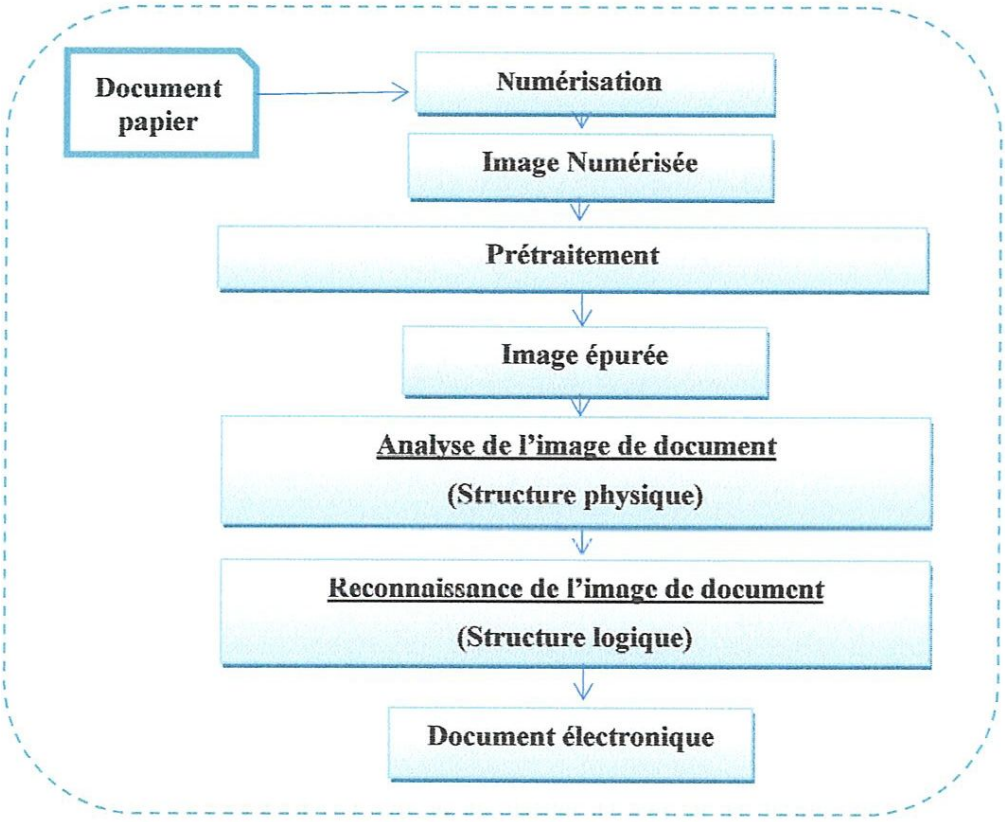


Figure II.4: Processus de reconnaissance de documents [KET 10].

II.5. Etapes d'analyse de documents

L'analyse de document joue un rôle très important dans le processus de reconnaissance et de compréhension de documents. Cependant, la reconnaissance sert à construire une version électronique exploitable à partir d'un document papier. Pour atteindre cet objectif, plusieurs étapes sont nécessaires.

La figure ci-dessous (***Figure II.5***) illustre de manière plus détaillée ces étapes :

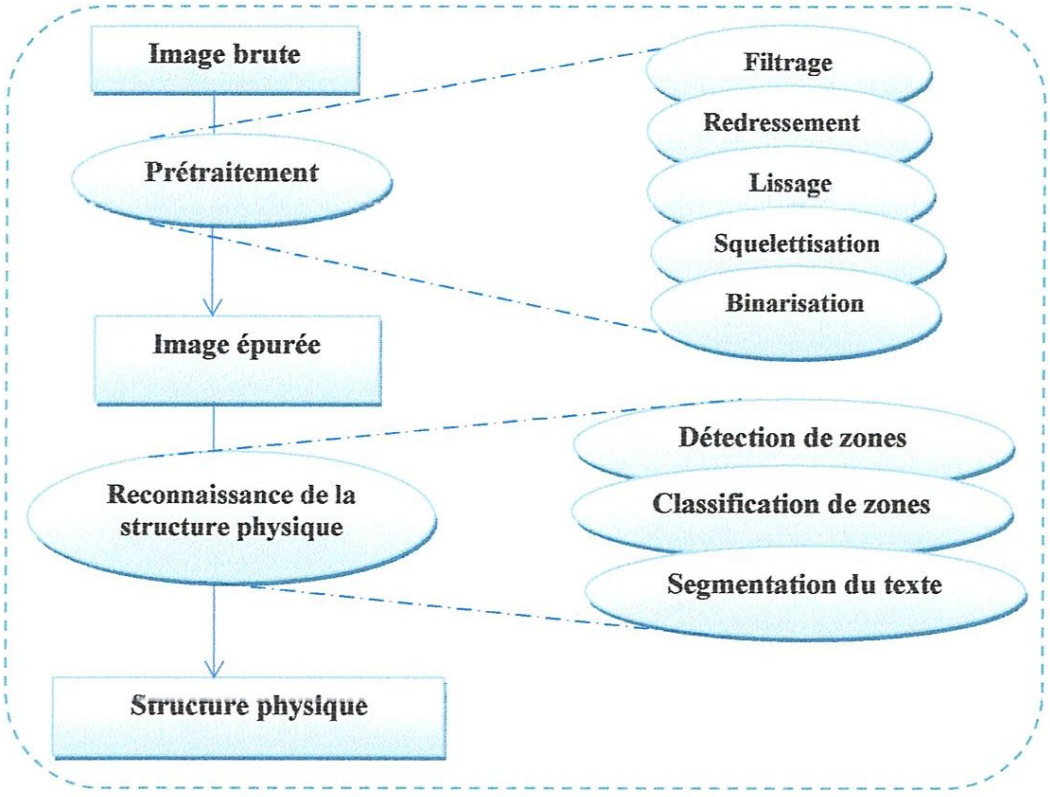


Figure II.5: Etapes de l'analyse de documents [ROB 01].

II.5.1. Prétraitements

L'image électronique obtenue par scannage est une image partiellement bruitée et biaisée appelée image brute.

Le bruit peut provenir de distorsions ou de poussières accumulées à divers endroits de l'appareil et le biais est du à une mauvaise position du document papier.

II.5.1.1. Définition

Le prétraitement consiste en une série d'opérations dont le but est la correction des imperfections et la préparation aux traitements futurs, on applique successivement des opérations de filtrage, de redressement, de lissage, de squelettisation ou de binarisation.

L'objectif du prétraitement est de, se rapprocher le plus possible de l'image idéale pour faciliter les traitements futurs.

II.5.1.2. Binarisation

Elle permet de passer d'une image en couleurs ou en niveaux de gris à une image binaire composée uniquement de 2 valeurs 0 et 1, plus simple à traiter. En général, on utilise un seuil de binarisation approprié qui traduit la limite des contrastes fort et faible dans l'image. Mais pour des images peu contrastées ou à contraste variable, il est difficile de fixer ce seuil à une valeur précise.

Différentes méthodes de binarisation ont été proposées dans la littérature telle que la méthode de Fisher (1958), d'Otsu (1979), de Kapur (1985), etc. Ces méthodes sont rapides et simples mais elles sont applicables sur des images non bruitées et possédant un fond uniforme.

Dans la méthode d'Otsu, considérée comme l'une des méthodes de binarisation les plus connues, un simple seuillage global suffit après l'analyse de l'histogramme des niveaux de gris, lorsque les documents papier sont de bonne qualité et de fond blanc. Mais si le fond est texturé ou le document est dégradé (pliures, taches ...), une analyse plus fine est nécessaire.

Ainsi, on peut trouver dans Trier et al [TRI 95], une bonne synthèse des méthodes de binarisation, proposant des seuils adaptatifs. Parmi les méthodes adaptatives les plus connues, nous citons : la méthode de Niblack, la méthode de Sauvola, la méthode de Bernsen, la méthode de Wolf, etc. Toutes ces méthodes procèdent à la binarisation d'image en calculant un seuil local pour chaque pixel de l'image en se basant sur des informations globales (sur l'image entière) et locales (sur le voisinage du pixel). La différence entre ces méthodes réside dans la façon de calculer les seuils locaux et dans les informations utilisées pour le calcul. Cependant, le défi reste total pour les fonds texturés où il est difficile de trouver une modalité claire dans la distribution.

En fait, il existe d'autres méthodes de binarisation basées sur le regroupement. Dans ces méthodes, les pixels de l'image sont attribués à l'une des deux classes Objet ou Fond en fonction d'un critère d'homogénéité. Cependant plusieurs techniques ont été employées pour atteindre cette classification : l'algorithme Kmeans (nuées dynamiques), le Réseau de neurones artificiel, la machine à support vectoriel (SVM), etc.

II.5.1.3. Redressement

Le redressement est une opération fréquente en analyse de documents, souvent due à un mauvais positionnement du document sur le scanner, conduisant à une inclinaison de l'image. Plusieurs méthodes de redressement ont été proposées ses deux dernières décennies, et la plupart des systèmes de reconnaissance optique de caractères (OCR) s'en trouvent pourvus actuellement tellement ces opérations sont courantes.

Les meilleurs algorithmes proposés sont sans doute ceux qui sont moins affectés par la présence de graphiques, de zones noires dans le texte présentant des inclinaisons différentes, ou de zones d'ombre près des marges provenant d'un phénomène de bombage du à la saisie de livres ou de magazines.

Une méthode, proposée par Agajan et Kailath [AGH 94], basée sur une analogie entre les lignes de texte et les ondes d'antenne radar. La distance entre une ligne de référence de chaque pixel du fond est convertie en une phase d'une onde sinus complexe. L'algorithme de détection détermine la cohérence spatiale entre les différentes contributions à partir de lignes différentes. Bien que similaire à l'idée de la transformée de Hough, cette méthode semble être, au dire de ses auteurs, plus efficace. Plus récemment, Chauduri & Pal ont proposé une méthode pour des documents indiens multi-scripts (Devanagari et Bangla).

II.5.2. Reconnaissance de la structure physique

La reconnaissance de la structure physique (ou forme physique) consiste d'une part en la détection et la classification des différentes zones de l'image en texte, graphique, table, formule, dessin ou photo et d'autre part en la découpe du texte en colonnes, paragraphes, lignes, mots et signes . A chaque objet de la structure physique est associé un ensemble d'attributs qui décrit l'apparence de l'objet (taille, fonte ou position)

La reconnaissance de la structure physique se fait en deux étapes principales ; la détection (la segmentation) et la classification (l'étiquetage). La détection comprend la segmentation des différentes zones de l'image afin de déterminer toutes les régions homogènes de l'image. La classification consiste à étiqueter les différentes zones extraites de la segmentation du document et de les regrouper ensemble. La classification peut être effectuée à plusieurs niveaux de granularité : blocs, colonnes, lignes ou mots, etc.

Généralement, la reconnaissance de la structure physique commence par la séparation des blocs de textes de tous les autres objets graphiques non textuels (illustrations, dessins, figures,

graphiques, ornements, cadres…). Ensuite, les zones de textes sont segmentées en colonnes, paragraphes, lignes, mots, caractères et ponctuations et le sens de lecture est retrouvé dans chaque paragraphe [EMP 03]. Pour les documents à structure complexe notamment les pages de journaux et de magazines, la reconnaissance de la structure physique, comprend à côté de la segmentation en régions (mot, ligne, colonne, etc.) et le regroupement de ces régions pour former des blocs, la détection des filets.

II.5.3. Approches de reconnaissance de la structure physique

Plusieurs classes de méthodes sont proposées en fonction de la nature de l'image (binaire, ou multi-niveaux de gris ou couleur), de la séparabilité de ses régions et de la régularité de sa structuration. On distingue : les méthodes descendantes et ascendantes (***Figure II.6***). Une troisième approche hybride peut être ajoutée.

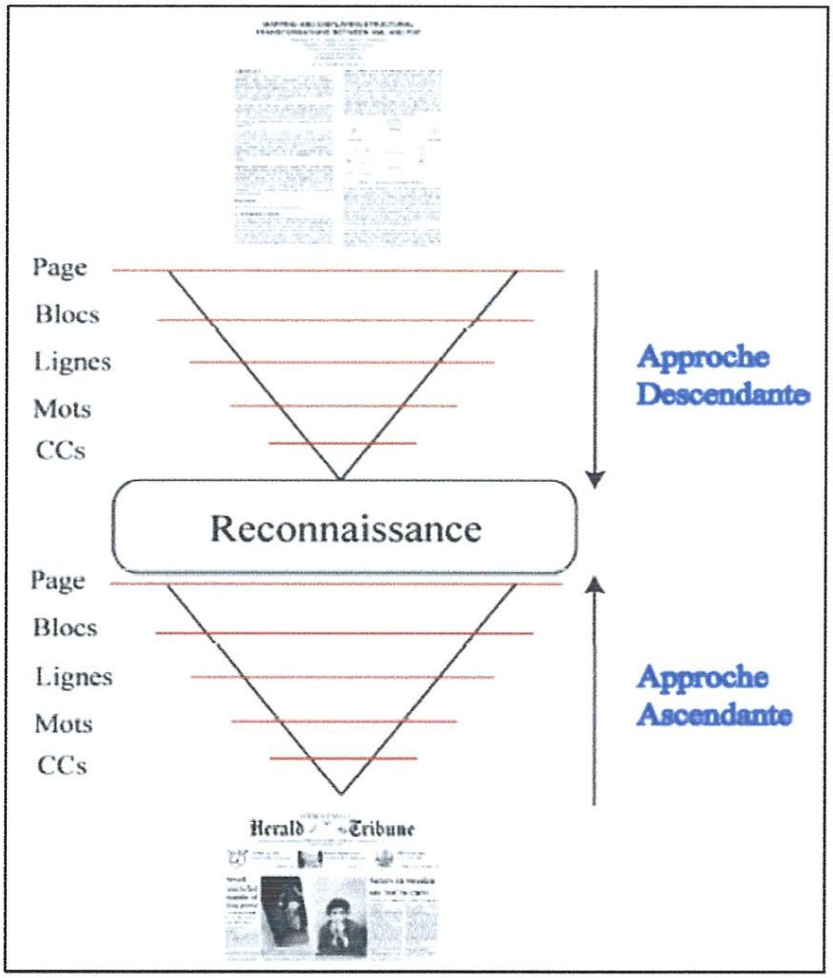


Figure II.6: Approche descendante et ascendante de reconnaissance de la structure physique [HAD 06].

II.5.3.1. Méthodes descendantes (top down)

L'approche descendante est souvent utilisée pour des documents à structure bien définie. Ces méthodes commencent par le niveau le plus élevé à savoir la page et descendent d'un niveau à un autre jusqu'à arriver au niveau des composantes connexes ou au niveau pixel (niveau le plus bas). Plusieurs méthodes descendantes ont été proposées dans la littérature : le célèbre algorithme de découpage (découpe en arbre X-Y), des méthodes utilisant l'algorithme de lissage RLSA, des méthodes basées sur l'analyse du fond blanc de l'image.

a) Méthodes utilisant l'algorithme de découpage X-Y :

L'algorithme de découpage en X-Ya été introduit par Nagy et al dans [NLA 88]. Le principe consiste à découper un document binaire horizontalement et verticalement en plusieurs rectangles. Le découpage continue dans chaque rectangle d'une manière récursive jusqu'à une condition soit satisfait. La condition d'arrêt est définie selon l'application souhaitée. Dans ce travail, les profils de projection horizontaux et verticaux sont étudiés pour définir les conditions d'arrêt afin d'extraire les lignes. La méthode a été appliquée sur des articles latins imprimés. Les auteurs donnent quelques résultats d'extraction de lignes sans toutefois mentionner le taux d'extraction. L'algorithme de découpage en XY a été utilisé pour des documents qui ne contiennent pas beaucoup des variations.

Les principales limites de cette classe de méthodes sont :

- La sensibilité aux inclinaisons ;

- L'inadaptation à segmenter des blocs mosaïques.

b) Méthodes utilisant l'algorithme de lissage RLSA :

Le RLSA (Run Length Smoothing Algorithm) [WON 82] consiste à relier les pixels noirs d'un document entre eux si leurs distances est inférieure à un certain seuil. Suivant le seuil choisi, cela permet de segmenter une lettre, un mot, une ligne ou un paragraphe. La figure suivante illustre ce principe.

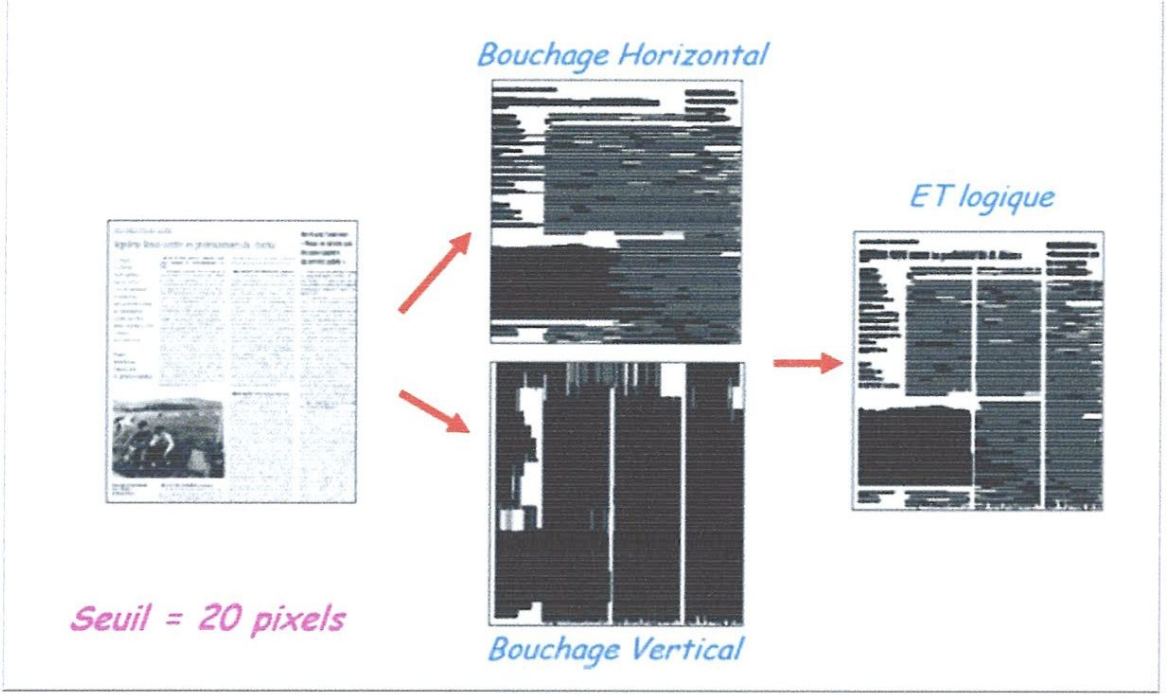


Figure II.7: Exemple d'application de RLSA [WEB 7].

Wang [WAN 89] propose une méthode basée sur la combinaison entre l'algorithme de lissage RLSA et l'algorithme de découpage X-Y récursif pour accomplir la segmentation d'image en blocs. Cette segmentation est suivie par une étape d'analyse de la texture pour classer les blocs résultants.

Les principales limites des méthodes basées sur l'algorithme RLSA sont [AZO 95] :

- Le choix arbitraire des seuils de lissage ;
- La sensibilité aux inclinaisons ;
- L'inadaptation à segmenter des blocs graphiques, formules et tableaux.

c) Méthodes utilisant l'analyse du fond blanc de l'image :

La segmentation par analyse du fond de l'image se base sur la détection et l'analyse de grandes zones d'espaces blancs. Ces algorithmes reposent sur l'hypothese que les entites sont délimitées par des flux de zones blanches verticales et horizontales plus grandes que les zones blanches comprises dans une entité. Ces approches cherchent à déterminer la structure de l'arrière plan correspondant aux zones blanches et non la structure du premier plan. Le principe général de ce type de méthodes est de rechercher un ensemble de rectangles maximaux qui ne contiennent pas de pixels du premier plan (pixels noirs). Ces rectangles une fois fusionnés permettent de construire les régions qui délimitent les blocs. Nous pouvons citer les travaux de [PAV 91] qui cherchent les colonnes blanches les plus larges possible

localement. Ces colonnes blanches sont ensuite fusionnées selon deux critères : si ces colonnes ont des tailles similaires et si elles sont suffisamment proches alors elles sont fusionnées. L'approche proposée dans [BAI 90] est basée sur la recherche des rectangles blancs maximaux. L'algorithme fusionne des rectangles blancs non maximaux fournis en entrée tant que le critère d'arrêt basé sur des heuristiques n'est pas atteint. Dans [ANT 98], le principe est le même que dans [PAV 92], mais il améliore le pavage des zones blanches. En effet, l'algorithme permet de traiter des zones ayant subi des rotations et pouvant être polygonales.

II.5.3.2. Méthodes ascendantes (bottom-up)

Les méthodes ascendantes sont basées sur l'analyse d'une image à partir des pixels.

Le principe des méthodes ascendantes est le suivant : elles commencent au niveau des composantes connexes, les fusionner en formant les mots, puis en fusionnant ces mots en lignes, les lignes en blocs, jusqu'à ce que la page soit complètement reconstituée.

Plusieurs méthodes ascendantes existent à savoir les méthodes utilisant les composantes connexes, les méthodes utilisant le filtrage à base de fenêtres, les méthodes utilisant la technique docstrum, etc.

a) Méthodes utilisant les composantes connexes :

Ces méthodes fusionnent les composantes connexes jusqu'à l'assemblage complet de la page du document. Fisher [FIS 90] a proposé une combinaison de l'algorithme de lissage avec l'extraction des composantes connexes. Cette approche permet d'identifier les zones textes et non textes mais reste cependant sensible à la rotation de l'image du document.

Drivas [DRI 95] a proposé une méthode de segmentation de pages d'images de documents composée d'un ensemble d'algorithmes. Le premier algorithme permet de déterminer l'angle de rotation, le deuxième permet la segmentation, et le troisième permet l'étiquetage des blocs obtenus en texte et en image. L'algorithme de segmentation consiste à extraire les composantes connexes et ensuite à appliquer la fusion de ces derniers. Cette fusion repose sur la recherche des plus proches composantes connexes et le regroupement des composantes connexes ayant une même dimension

La ***figure II.8*** présente un exemple de l'étiquetage des composantes connexes:

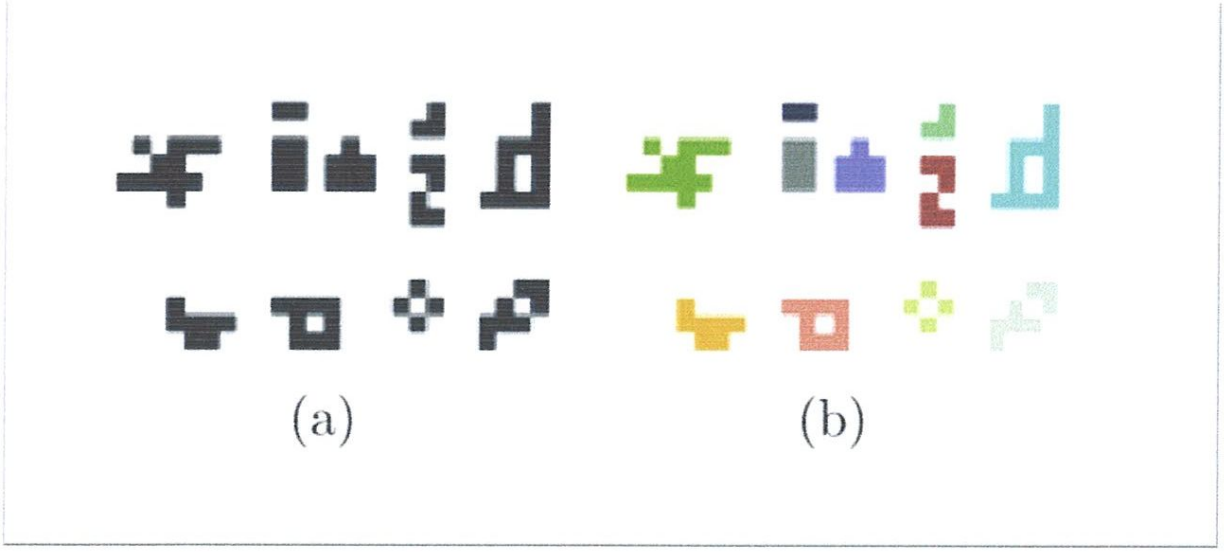


Figure II.8: Exemple de l'étiquetage des composantes connexes : (a) Image d'origine; (b) Composantes connexes.

b) Méthodes utilisant le filtrage à base de fenêtres :

Les méthodes ascendantes, utilisant le filtrage à base de fenêtres, reposent sur un balayage d'une fenêtre de taille quelconque sur l'image entière de document.

Le bourgeois [LEB 92] utilise un filtre de 8 × 3 pixels. L'image échantillonnée est dilatée par un élément de structure horizontale pour rassembler les caractères adjacents l'un vers l'autre. Notons que chaque composante connexe de l'image est définie par un rectangle englobant et par la moyenne des longueurs de plages de valeurs de pixels noirs. Si la composant connexe est à l'intérieur de l'intervalle, celle-ci sera classée en une zone de texte qui sont après fusionnées verticalement en blocs, sinon elle sera classée en zone non texte [HAD 06].

c) Méthodes utilisant les diagrammes de Voronoi :

L'utilisation du diagramme de Voronoï [CHA 91] permet un partitionnement de l'image en polygones. A l'initialisation, le diagramme est construit sur un ensemble de germes sélectionnés de façon aléatoire sur l'image via un processus de Poisson. L'application des phases de fusion et de décomposition permet la suppression de régions (et de germes) superflus ainsi qu'une décomposition des régions hétérogènes au sens du critère adopté initialement.

Il est à noter que cette méthode est efficace pour l'extraction des zones de texte et possède un taux de reconnaissance comparable à celui obtenu par les méthodes basées sur l'analyse des composantes connexes.

II.5.3.3. Méthodes mixtes

Les méthodes mixtes proposent d'utiliser les avantages des méthodes ascendantes et des méthodes descendantes par la mise en place d'architectures d'analyse mixant ces deux types d'approches. Nagy et al. [NLA 88] ont par exemple, développé une architecture logicielle utilisant une grammaire pour spécifier des règles explicitant comment regrouper les pixels d'une image pour construire des objets de niveaux sémantiques de plus en plus élevés. La construction et l'étiquetage des blocs sont alors réalisés simultanément en utilisant une approche mixte. Il devient possible de produire des grammaires adaptées aux différentes classes de documents mais la mise en place de ces critères reste très complexe et rend ce système peu générique.

L'analyse des approches ascendantes, descendantes et mixte montre que ces approches manque de généricité. De plus, ces méthodes utilisent des paramètres libres qui nécessitent une bonne connaissance de la structure du document.

II.6. Conclusion

Dans la majorité des cas, la structure physique est le point de départ de la reconnaissance d'un document. Sa reconnaissance à généralement deux buts ; la première est la segmentation du document en zones homogènes, et la deuxième l'étiquetage de ces zones. Parmi les méthodes de segmentation, on trouve des méthodes ascendantes, descendantes et mixtes.

Dans ce chapitre, Nous avons essayé de présenter une vue globale sur le domaine de l'analyse de documents. Nous avons défini dans ce chapitre les notions de document et structure, et cité les différentes structures d'un document tout en focalisant sur la structure physique. La majeure partie du chapitre a été réservé à la présentation de tous les points relatifs aux différentes approches et méthodes existantes dans la littérature pour la reconnaissance de la structure physique. Cette dernière est considérée comme l'entrée de la prochaine étape de reconnaissance de la structure logique.

Chapitre III

Conception du système

III.1. Introduction

Le présent chapitre est consacré à la présentation de la conception ou bien la structure générale et la modélisation de notre application de **localisation du matricule dans les relevés de notes du bac**.

Au départ, nous présentons dans ce chapitre l'étude physique des relevés de notes du baccalauréat ; Comme plusieurs formats existent, nous commençons par la présentation (l'extraction) des caractéristiques de chaque format de relevés. Après, nous décrivons l'approche proposée et nous détaillons les différentes étapes impliquées, et les méthodes utilisées, avant de conclure.

Nous essayons au cours de ce chapitre de démontrer comment nous avons pu faire passer d'une image brute de document à un ensemble d'informations structurées exploitables sur le document.

III.2. Objectif du projet

L'objectif de ce projet de fin d'étude est donc de développer un système d'analyse des relevés de notes du bac numérisés pour en extraire et localiser une information très importante qui est **le matricule de l'étudiant**. Une fois le matricule est localisé sur l'image du relevé de notes, nous le pouvons passer à un autre système de reconnaissance de chiffres pour le reconnaitre.

Ce matricule (après sa reconnaissance) nous serve par la suite comme clé de recherche dans la base de données des étudiants, s'il est reconnu pour ramener le reste des informations de l'étudiant (Nom, prénom, date de naissance, ...etc.).

III.3. Analyse physique des relevés de notes du baccalauréat

III.3.1. Caractéristiques des relevés de notes

Après l'étude visuelle du relevés de notes de baccalauréat algériens depuis l'année « 1997 » jusqu'à « 2016 », nous avons remarqué qu'ils varient d'une année à l'autre et ne prennent pas une forme standard (La ***figure III.1***), et qu'ils possèdent une structure complexe (cadre, entête, tableau, texte arabe, texte latin, graphique ... etc.)

Les variations se situent dans :

- **La qualité du papier :** y a des années où la qualité de papier n'est pas la même par exemple dans le cas de papier standard il se plie ou se déchiré ce qui fait perdre les informations.
- **Polices de caractères :** l'utilisation de plusieurs types.
- **Taille et forme de cadre :** qui se diffèrent d'une année à une autre.
- **Arrière plan flagrant :** qui perturbe la vision des informations utiles.

Cette variabilité exige une étude approfondie avant le passage à l'analyse physique des images.

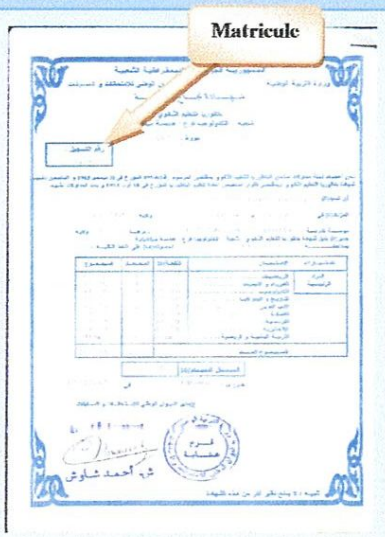
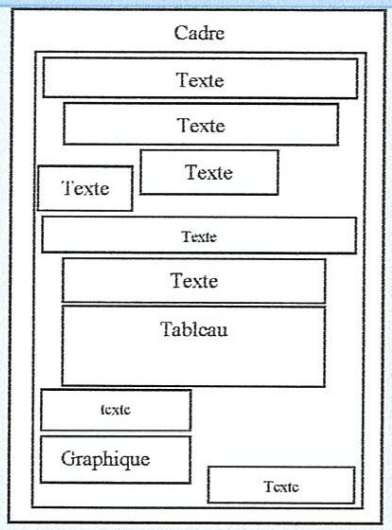
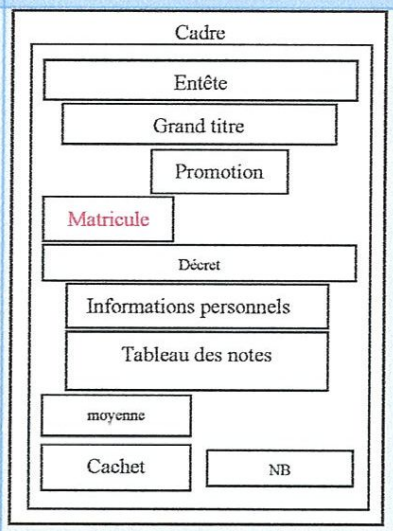
Bac « 1998 » Bac « 2002 » Bac « 2013 » Bac « 2015 »

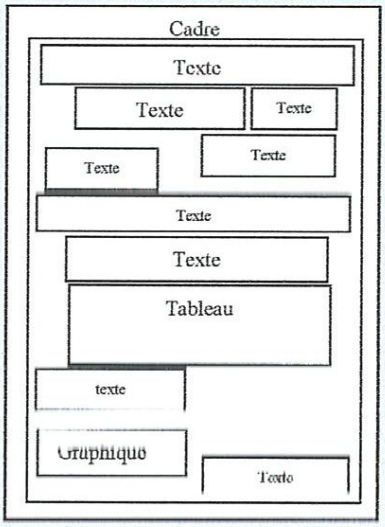
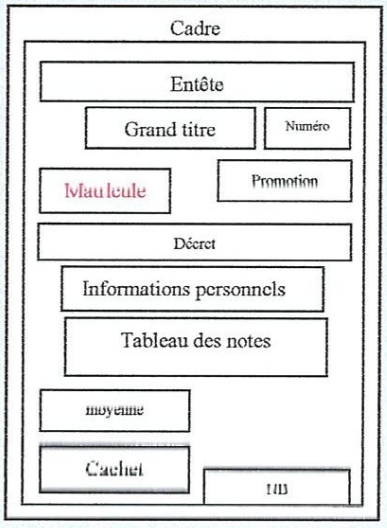
Figure III.1 : Différents types de document.

III.3.2. Structure des relevés de notes

Selon les images de la figure **III.1**, on peut remarquer que généralement le matricule se situe dans la partie supérieure (premier tiers de document) et dans le coin inférieur gauche dans cette partie. Il est à noter aussi que le matricule est parfois encadré seul ou avec un texte arabe ("رقم التسجيل") au-dessus, ou bien il est délimité par une ligne droite au dessus.

Dans la suite, nous allons présenter les différents niveaux de structures des relevés de notes :

Document original	Structure physique	Structure logique
Bac 99 et antérieur, 1cr cas :		
		

Document original	Structure physique	Structure logique
Bac 2000 et suivant, 2ème cas :		
		

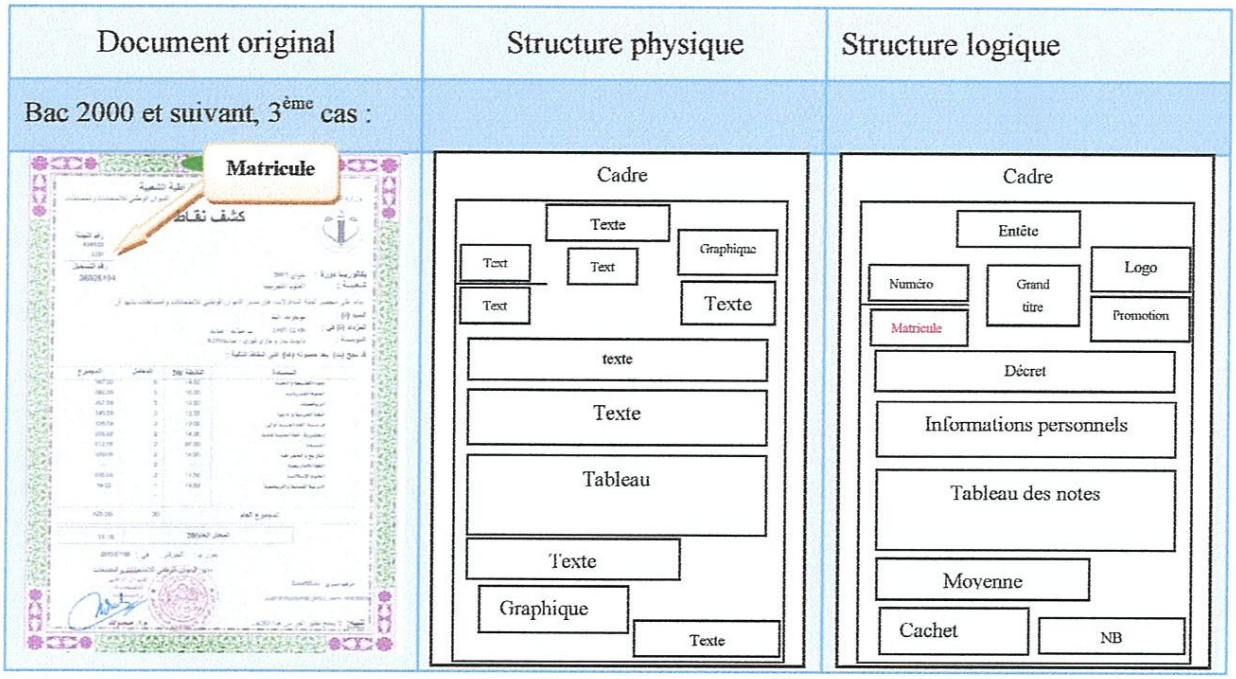


Figure III.2 : Différentes niveaux de structures d'un document.

III.4. Description de l'approche proposée

La réalisation de notre système se fait par plusieurs étapes de traitement que nous pouvons les regrouper en trois parties :

1. Une première partie d'extraction de la structure physique.
2. Une deuxième partie d'étiquetage logique (extraction du « Matricule »).
3. Une troisième partie de reconnaissance du matricule.

La partie d'extraction de la structure physique inclue deux phases, regroupant chacune un ensemble de traitements :

1. La phase de prétraitement : afin d'améliorer la qualité de l'image d'entrée.
2. La phase de segmentation : permettant de séparer les entités physiques du document.

La deuxième partie est basée sur l'étiquetage logique de certaines entités physiques extraites précédemment : celles composants le matricule du relevé de notes.

La troisième partie du système reçoit en entrée une sous-image incluant le matricule de l'étudiant extrait lors de la phase précédente, cette image enregistrée avec le relevé originale dans une base de données, par la suite, fait intervenir un algorithme de reconnaissance de chiffres pour fournir en résultat le matricule sous forme d'une chaine de caractères.
Le schéma présenté dans la figure ci-dessous, illustre les différentes étapes de traitement impliquées dans notre système :

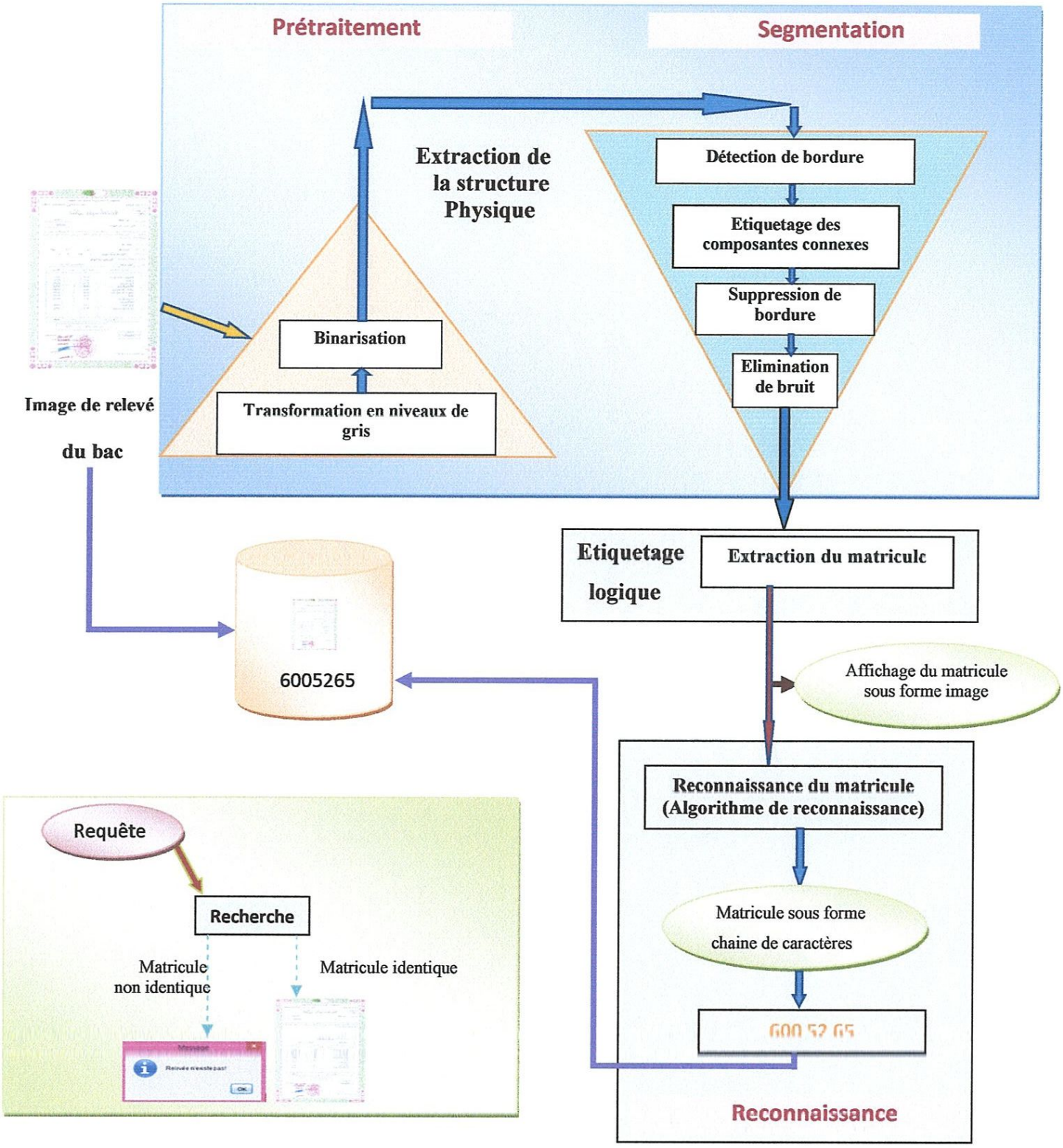


Figure III.3 : Schéma du processus général du système proposé.

III.4.1. Première partie : extraction de la structure physique

Cette partie a pour but d'extraire la structure physique du document sans se préoccuper de son format (défini lors de l'analyse physique). Ainsi, la structure physique se diffère selon le type

du document considéré dans l'étude. Pour notre cas, on s'arête au niveau des composantes connexes.

Cette partie regroupe deux phases : prétraitement, et segmentation. Dans cette section nous allons détailler chacune de ces deux phases.

III.4.1.1. Prétraitement de l'image

Dans tout système d'analyse et de reconnaissance de documents le prétraitement joue un rôle primordial et considéré comme une phase fondamentale. Le prétraitement regroupe un ensemble d'opérations tel que: « l'élimination de bruit », « réduction des dégradations »... etc. Cette phase consiste à préservé que les informations utiles de l'image, de manière à préparer le terrain aux prochains traitements dans le processus d'analyse et de reconnaissance.

Dans notre système, le prétraitement regroupe les étapes suivantes: la transformation en niveaux de gris, la binarisation, et l'élimination de bruit par lissage.

III.4.1.1.1. Transformation en niveaux de gris

Le premier traitement à faire est de transformer l'image en niveaux de gris. Cette transformation est nécessaire car la méthode de binarisation que nous utiliserons dans l'étape suivante n'est applicable que sur des images en niveaux de gris. Ce passage peut être effectué en utilisant le pseudo code suivant :

Algorithme de transformation en niveaux de gris

Entrée : image I en couleurs
Sortie : image I' en niveaux de gris
Début
  Pour chaque pixel p de l'image faire :
    $r \leftarrow$ la quantité de la couleur rouge du pixel p ;
    $v \leftarrow$ la quantité de la couleur vert du pixel p ;
    $b \leftarrow$ la quantité de la couleur bleu du pixel p ;
    $g \leftarrow (r + v + b) / 3$,
  Fin Pour
Fin.

III.4.1.1.2. Binarisation

La binarisation appelée aussi seuillage, est la technique de classification la plus simple. Elle permet de classifier les pixels de l'image à l'aide d'un seuil (dans le cas de seuillage global)

en deux classes. En général, ils sont représentés par une classe de pixels noirs et une autre classe de pixels blancs. L'image est alors séparée en deux classes, une classe représentant le fond de l'image et une autre classe représentant la scène de l'image (L'objet). La binarisation permet alors de conserver les pixels ayant un niveau de gris compris entre 0 et **T** ou entre **T**+1 et 255. Le reste est par conséquent ignoré.

Un très grand nombre de techniques de binarisation ont été proposées et chacune d'entre elles a des caractéristiques différentes. Dans notre système nous avons proposé d'intégrer une méthode de seuillage global, et nous avons opté pour la méthode d'Otsu [OTS 79]. C'est une méthode très populaire et elle a montré des bonnes performances dans plusieurs études comparatives antérieures, surtout pour les images de documents qui sont de qualité suffisante comme les images de relevés de notes.

- **La méthode d'OTSU**

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

Algorithme de la méthode d'Otsu

Entrée : Image I en niveaux de gris
Sortie: Image I' binaire avec **0 = Noir** et **1 = Blanc** ;
Début
Calculer h l'histogramme de niveaux de gris de I ;
Calculer $h2$ l'histogramme normalisé ;
Pour chaque niveau de gris S Faire

$$q_1(S) = \sum_{i=0}^{S-1} h2(i) \; ; \; q_2(S) - \sum_{i=S}^{255} h2(i) \; ;$$

$$\mu_1(S) = \frac{1}{q_1(S)} \sum_{i-0}^{S-1} h2(i) \times i \; ; \; \mu_2(S) = \frac{1}{q_2(S)} \sum_{i-S}^{255} h2(i) \times i \; ;$$

$$\sigma_{inter}{}^2 = q_1(s) \times q_2(s) \times [\mu_1(s) - \mu_2(s)]^2 ;$$

Fin pour
$T \leftarrow$ le niveau de gris dont la variance est minimale ;
Pour chaque pixel (x, y) de l'image Faire
Si $I(x, y) < T$ alors $I'(x, y) \leftarrow 0$;
Sinon $I'(x, y) \leftarrow 1$;
Fin Pour
Fin.

La figure (***Figure III.4***) illustre le résultat de binarisation d'une image du relevé du Bac en niveaux de gris en utilisant la méthode d'Otsu.

(*a*) Image en niveau de gris (*b*) Résultat de la méthode d'OTSU

Figure III.4 : Binarisation du relevé du bac.

III.4.1.2. Segmentation

L'objectif principal de l'extraction de la structure physique est de déterminer les frontières des déférentes régions de l'image du document. Cette extraction a pour but de décomposer l'image en une hiérarchie de régions homogènes.

Généralement, comme nous avons vu dans le deuxième chapitre les trois approches de segmentation de documents : ascendantes, descendantes, et mixtes.

Dans notre système, nous appliquons une segmentation mixte. Tout d'abord, nous commençons par une segmentation ascendante afin de regrouper les pixels ayant les mêmes propriétés en composants connexe. Ensuite, nous utilisons les caractéristiques de ces composantes connexes pour séparer les composantes textuelles des autres composantes.

III.4.1.2.1. Application de l'algorithme RLSA pour la détection de la bordure

La bordure de relevé des notes qui entoure les informations du document est sans aucune importance ; Elle porte une forme non régulière ce qui perturbe l'analyse de document (trop d'objets graphiques à étiqueter, et non unifiée pour tous les relevés).

Notre travail consiste à localiser le matricule qui se trouve généralement dans un rectangle, et ce rectangle devient la plus grande composante connexe après la bordure dans la partie

supérieure (premier tiers du document). Donc c'est pour ces raisons qu'il faut éliminer la bordure définitivement.

Selon l'étude physique que nous avons effectué plusieurs types de bordures existent : bordure sous forme de cadre et donc il est formé d'une seule composante, bordure sous forme d'une série d'étoiles ou d'autres formes géométriques, etc. La diversité de types de bordures complique l'opération de détection et la méthode employée qui doit donc être générique et applicable pour tous les types de bordure.

La méthode que nous avons appliquée pour la détection de bordure est basée sur l'algorithme RLSA (Run Length Smoothing Algorithm).

L'idée de cet algorithme consiste à relier les pixels noirs successifs séparés par moins de *n* pixels blanc dans les deux directions horizontale et verticale. Cependant, comme notre objectif dans cette étape est de relier uniquement les pixels composant la bordure de l'image, nous avons proposé de ne pas appliquer l'algorithme RLSA sur l'image entière mais uniquement sur la partie de l'image contenant le cadre.

Le seuil *n* est fixé par expérimentations égale à 10% pixels de largeur de l'image de relevé. De plus, d'après l'étude physique des relevés de bac, on a trouvé que plusieurs types de cadres existent, et la position du cadre dans le document peut se diffère légèrement d'un document à un autre, mais l'épaisseur du cadre ne dépasse jamais la valeur (*largeur du document* / 7). Cependant, le cadre prend la forme d'un rectangle formé de quatre cotés : deux cotés horizontaux (un en haut de l'image et l'autre en bas) et deux cotés verticaux (à droite et à gauche de l'image respectivement). En divisant l'image horizontalement en 7 sous-images, les deux cotés horizontaux du cadre se trouvent respectivement dans la première et la dernière sous-image. De même, en divisant l'image verticalement en 7 sous-images, les deux cotés verticaux du cadre se trouvent dans la première et la dernière sous-image respectivement.

La détection de la bordure peut être résumée par le pseudo-code suivant :

Algorithme de la détection de la bordure

Entrée : Image Binarisée (*I*)

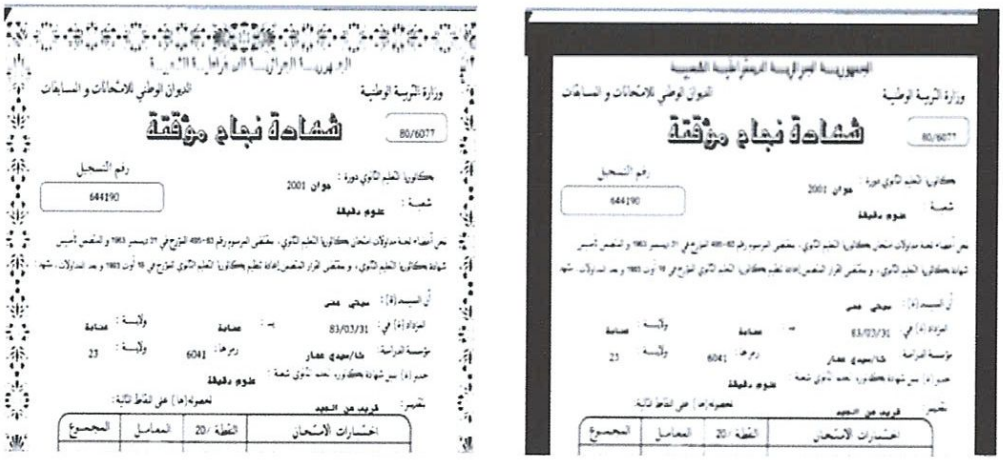
Sortie : Image sans bordure (*I`*)

Début

- *Epaisseur* ← largeur de l'image / 7 ;
- Diviser l'image horizontalement en plusieurs sous-images chacune d'hauteur = *Epaisseur* ;
- Appliquer l'algorithme RLSA horizontalement sur la première et la dernière sous-image séparément en prenant n = 10% de largeur de l'image. Ce lissage permet de relier les pixels noirs du cadre proches (dont la distance inférieure à n) selon l'axe horizontal.
- Diviser l'image verticalement en plusieurs sous-images chacune de largeur = *Epaisseur* ;
- Appliquer l'algorithme RLSA verticalement sur la première et la dernière sous-image séparément en prenant n = 10% de largeur de l'image. Ce lissage permet de relier les pixels noirs du cadre proches (dont la distance inférieure à n) selon l'axe vertical.
- A la fin des deux passes de l'algorithme RLSA (horizontal et vertical) le cadre devient composé d'une seule unité ou composante.

Fin.

La figure suivante présente le résultat de cette étape :



(a) Image binaire (b) Détection du cadre

Figure III.5 : Détection de la bordure.

III.4.1.2.2. Etiquetage des composantes connexes

L'étiquetage des composantes connexes est une opération qui permet de passer du niveau d'analyse lié à l'échelle du pixel à un niveau d'analyse lié à l'information des différentes régions de l'image. Elle consiste à regrouper tous les pixels noirs voisins dans une unité distincte. Et pour cela nous utilisons la méthode d'agrégation des pixels.

Nous avons conçu dans notre système deux versions de telle méthode: une version récursive et une autre itérative. Pour la version récursive, elle est plus rapide mais devient inefficace dans le cas des grandes composantes tel que les composantes graphiques. Pour cette raison, on à utilisé une deuxième version « itérative » malgré sa lenteur d'exécution.

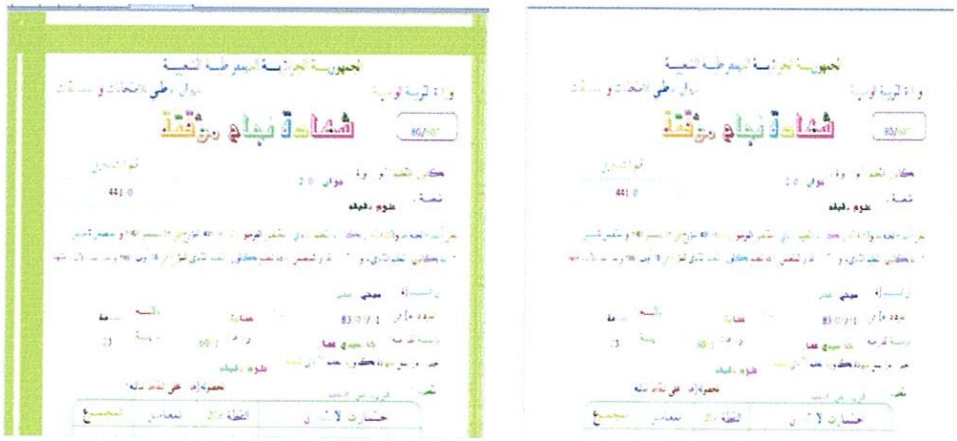
Le résultat de l'étiquetage des composantes connexes est une image en couleurs dont chaque composante connexe est affichée par une couleur différente. La figure suivante (***Figure III.6***) illustre le résultat de l'étiquetage des composantes connexes.

Figure III.6 : Etiquetage des composantes connexes.

III.4.1.2.3. Suppression de la bordure

Après l'application de l'algorithme RLSA, tous les pixels de la bordure sont devenus interconnectés et forment une seule composante connexe lors de l'étiquetage.

La bordure est simplement la composante connexe plus grande. Pour supprimer la bordure il suffit donc de chercher la composante connexe dont la taille est maximale et de la supprimer par la suite. La figure suivante présente le résultat de cette étape :



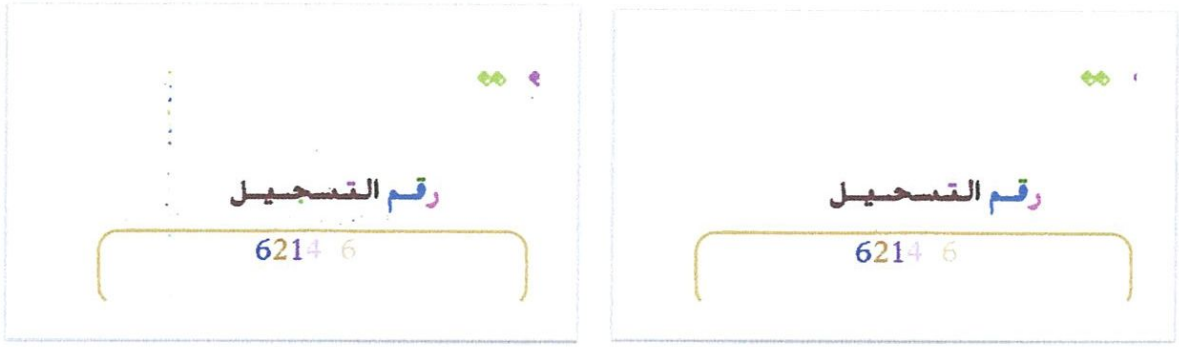
(a) Image des composantes étiquetées (b) Suppression du cadre

Figure III.7: Elimination de la bordure.

III.4.1.2.4. Elimination de bruit (Lissage)

Le lissage est une opération important dans cette phase, parce que la binarisation peut introduire du bruit dans l'image, ce qui pose des problèmes sur les autres traitements suivants et diminue les performances de notre système. Pour remédier ce problème, nous avons proposé d'appliquer un lissage à l'image par nettoyage, qui sert à éliminer tous les pixels isolés. Le principe est de recherchez la taille de la plus petite composante connexe (notons *tailleMin* cette taille), puis supprimer toutes les composantes connexes ayant une taille inférieure ou égale à *tailleMin* × 2% de largeur de l'image.

La figure suivante montre le résultat d'application de cette méthode :



(a) **(b)**

Figure III.8: Résultat de lissage, (a) image étiquetée, (b) image lissée.

III.4.2. Deuxième partie : Etiquetage logique du matricule

Le matricule est composé d'un ensemble de composantes connexes possédant certaines caractéristiques. Ces caractéristiques sont déduites de l'étude physique des documents.

Dans les deux cas premiers présentés dans la figure III.2

Dans les deux premiers cas, le matricule est entouré par un rectangle dans la partie haute de l'image. Cependant, après l'élimination de la bordure, la partie haute de l'image ne contient qu'une seule rectangle : celle qui englobe la matricule. Cette caractéristique peut donc être exploitée pour localiser le matricule dans le cas des relevés comme ceux des deux cas premiers présentés dans la ***figure III.2***.

Au départ, on cherche dans la partie supérieure et plus précisément dans le premier tiers de l'image un rectangle, qui est la composante connexe la plus large.

Une fois le rectangle est trouvé, on cherche à l'intérieur de ce rectangle pour identifier les composantes du matricule. Deux cas de figures existent :

- Le rectangle contient uniquement les composantes correspondantes aux chiffres du matricule sur une seule ligne de texte (2ème cas de la figure III.2).
- Le rectangle englobe deux lignes de texte. La première contient deux mots arabes (رقم التسجيل) et la deuxième contient les chiffres du matricule comme dans le 1er cas de la figure III.2.

En effet, pour ces deux cas, les chiffres du matricule ont généralement la même hauteur et ils sont bien alignés horizontalement.

Cependant, pour identifier les composantes du matricule dans ces deux cas premiers, et en tenant compte de la remarque précédente, on procède comme suit : on commence par parcourir toutes les composantes connexes qui se trouvent à l'intérieur du rectangle. Parmi ces composantes on élimine celles qui correspondent à des points parasites (bruit) éventuels. Ainsi, les composantes connexes dont la largeur est inférieure à 5% de la largeur de l'image de relevé sont éliminées (ce pourcentage est choisie par experience).

Ensuite, on cherche la composante connexe ayant la plus grande coordonnée verticale (*Xmax*). Puis, on parcoure la partie de l'image du relevé contenant la rectangle du matricule à partir de cette coordonnée (*Xmax*) et montons jusqu'à trouver un vide (ligne blanc) ; La figure suivante explique la position du *(Xmax)*:

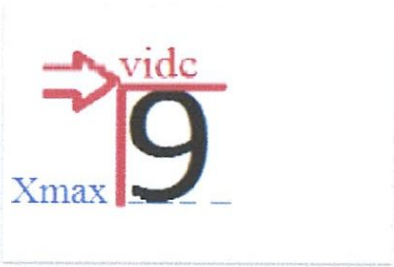


Figure III.9 : Exemple expliqué la position du (Xmax).

Ensuite, il suffit de dessiner une ligne séparatrice le long du vide trouvé dans le rectangle.

Le matricule est donc composé des composantes connexes inclues dans la rectangle et se trouvant au dessous de cette ligne séparatrice (***Figures III.10*** et ***III.11***).

Figure III.10 : Exemple de détection du matricule dans le premier cas.

Figure III.11 : Exemple de détection du matricule dans le deuxième cas.

Dans le troisième cas présenté dans la figure III.2

Dans ce cas, le matricule n'est pas entouré par un rectangle mais il se trouve au dessous d'une ligne droite de largeur moyenne. Il sera facile donc de localiser le matricule si on arrive à détecter cette ligne droite.

Pour détecter cette ligne, on recherche toujours dans le premier tiers de l'image la composante connexe la plus large. Après on dessine un rectangle virtuel à partir de cette ligne afin de se transformer au deuxième cas présenté précédent.

Ainsi, la ligne détectée constitue le côté longue haut du rectangle. La largeur du rectangle est choisi par expérimentations égale à (longueur /2).

A la fin de cette étape nous tombons directement dans le 1er cas, et on répète les mêmes traitements présentés dans la section précédente afin de séparer la ligne de l'écriture arabe et celle des chiffres du matricule (Figure suivante), et localiser le matricule par la suite.

Figure III.12: Exemple de détection du matricule dans le troisième cas.

III.4.3. Troisième partie : Reconnaissance du matricule

Comme nous avons signalé dés le départ, l'objectif de notre travail est la localisation du matricule dans les relevés du bac. Cette localisation peut être exploitée ultérieurement pour divers traitements. Afin de mieux éclaircir l'utilité et l'importance de notre travail, nous avons proposé d'ajouter un module de reconnaissance dans notre application qui exploite le résultat de détection (qui est une sous-image contenant uniquement le matricule) et produit le matricule, après sa reconnaissance, sous forme de chaine de caractères.

Le module de reconnaissance dans cette partie utilise la bibliothèque java «Tess4J[2]», qui inclut une technique de reconnaissance de caractères (OCR). Cette technique, permet

[2] http://tess4j.sourceforge.net.

comme son nom l'indique, de reconnaître les caractères d'un document imprimé ou manuscrit. L'intégration de « Tess4J » dans le projet et son utilisation ont été plutôt simple.

Pour pouvoir utiliser cette bibliothèque, il faut l'importer dans notre projet. Nous avons besoin d'ajouter les deux lignes suivant dans la classe principale :

import net.sourceforge.tess4j.Tesseract;

import net.sourceforge.tess4j.TesseractException;

Il faut également ajouter les instructions nécessaires pour la reconnaissance des chiffres du matricule :

Tesseract instance = Tesseract.getInstance();

String result = instance.doOCR(bi); // *bi* est une BufferedImage contenant l'image du matricule

La figure ci-dessous représente une partie du code montrant l'utilisation de la bibliothèque «Tess4J» :

```
private void jButton7ActionPerformed(java.awt.event.ActionEvent evt) {
    try {
        // TODO add your handling code here:
        ITesseract instance = new Tesseract();  // JNA Interface Mapping
        instance.setDatapath("C:\\Users\\LthMsd\\Documents\\NetBeansProjects\\api\\Tess4J\\tessdata");
        String result = instance.doOCR(newImage1);
        jTextField1.setText(result);
    } catch (TesseractException ex) {
        Logger.getLogger(fenetre.class.getName()).log(Level.SEVERE, null, ex);
    }
}
```

Figure III.13 : code montrant l'utilisation de la bibliothèque «Tess4J».

III.4.4. Stockage des données et module de recherche

Une fois la reconnaissance est terminée, l'image originale du relevé de notes, et le matricule sous forme de chaine de caractères sont stockés sur disque. Ces données prouvent être exploitées par la suite par un module de recherche où le matricule (sous forme de chaine de caractères) constitue la clé de recherche.

Le système de stockage conçu est composé d'un dossier nommé « relevé » pour le stockage des relevés originaux, et un fichier texte nommé « matricule.txt ». Ce dernier stocke les matricules après leur reconnaissance (sous forme de chaine de caractères) ainsi que le chemin du relevé de note correspondant à chaque matricule. La figure suivante présente un exemple de tel fichier :

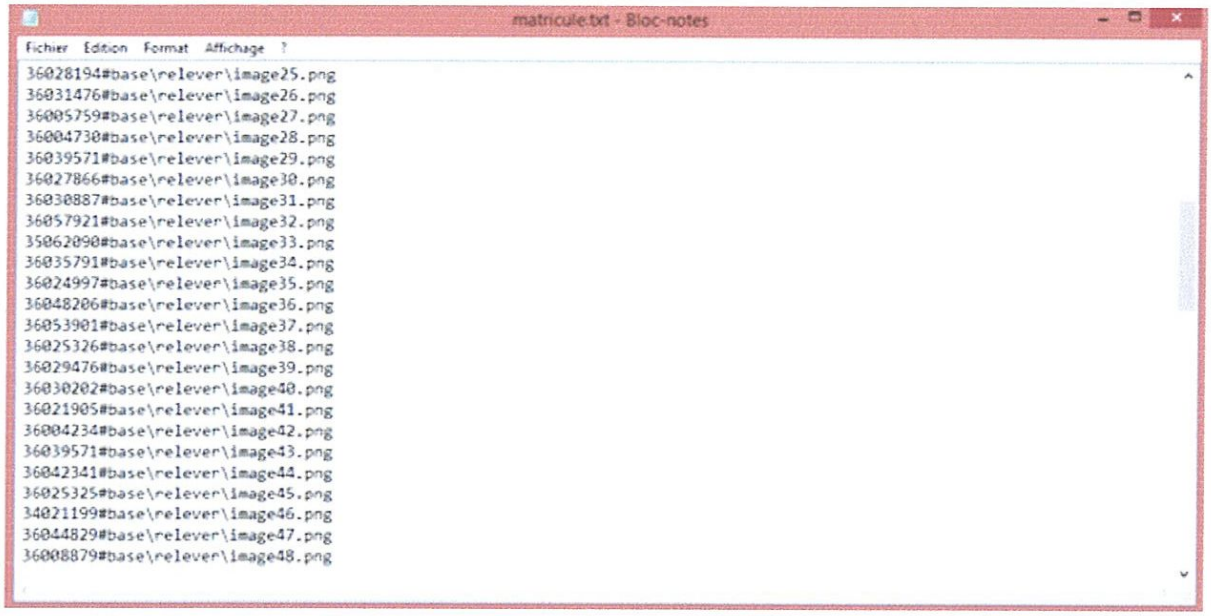


Figure III.14 : Fichier stockant les matricules reconnus sous forme de chaines de caractères.

Pour tester ce système de reconnaissance nous avons proposé de faire un simple module de recherche qu'il reçoit comme entrée une requête textuelle (chaine de caractères) correspondante au matricule recherché. Ce module cherche ensuite dans la base de données si le matricule recherché existe ou pas. S'il est trouvé, on affiche le relevé de note correspondant (relevé originale); sinon, on retourne dans une fenêtre « le matricule n'existe pas ».

III.5. Conclusion

Dans ce chapitre, nous avons détaillé l'approche proposée pour la réalisation du notre système, et bien présenté les différentes étapes de modélisation.

L'approche proposée regroupe plusieurs parties ; chaque partie inclut des différentes phases. Nous avons également détaillé chacune de ces parties ainsi que, ces différentes phases qui sont d'une part l'amélioration de la qualité de l'image de document et d'autre part la segmentation de cette dernière.

L'implémentation de la méthode proposée, les tests effectués, et les résultats obtenus seront l'objet du prochain chapitre.

Chapitre IV

Implémentation et résultats

IV.1. Introduction

L'implémentation est la phase la plus importante après celle de la conception. Ce chapitre prend en charge la plateforme matérielle et logicielle, ainsi que la présentation de l'application développée et les résultats obtenus à travers des captures-écrans.

IV.2. Présentation des outils de développement

IV.2.1. Plateforme matérielle

L'implémentation de l'application est réalisée sur un ordinateur portable ayant les caractéristiques suivantes :

- Machine : Toshiba
- Processeur : Intel(R) Core™ i3-3120(R) CPU P6200
- Fréquence : 2.50 GHz
- RAM : 4.00 Go
- Carte graphique : Intel(R)
- Système d'exploitation : Microsoft Windows 8 « 64-Bits »

IV.2.2. Plateforme logicielle

Vu le nombre de programmes, le choix d'un langage de programmation et d'un environnement de développement est donc une décision importante. Dans ce qui suit, nous présentons brièvement les environnements que nous avons choisis pour l'implémentation :

- Le langage de programmation « Java ».
- L'environnement de développement« NetBeans ».

IV.2.2.1. Langage de programmation

Java est un langage de programmation à usage général, évolué et orienté objet dont la syntaxe est proche du « C ». Ses caractéristiques ainsi que la richesse de son écosystème et de sa communauté lui ont permis d'être très largement utilisé pour le développement d'applications de types très disparates. Elle possède un certain nombre de caractéristiques qui ont largement contribué à son énorme succès : elle est simple et portable, indépendant de toute plate-forme, assure la gestion de la mémoire, multitâche, économe...etc. Java est notamment largement utilisée pour le développement d'applications d'entreprises et mobiles [WEB 8].

IV.2.2.2. Environnement de développement

Nous avons développé notre système sous l'environnement NetBeans version IDE 8.1, qu'il est un environnement de développement intégré(EDI), placé en open source, sous licence CDDL (Common Development and Distribution License) et GPLv2. NetBeans IDE est spécialement conçu pour les programmeurs pour les aider à développer facilement des applications Web, mobiles et bureautiques. Il prend en charge les dernières technologies Java et les améliorations de spécifications en plus d'autres technologies à la fine pointe de la technologie. L'IDE est livré avec une grande variété d'outils, de fonctionnalités, d'échantillons et de modèles. Elle possède nombreux caractéristiques y compris [WEB 9]:

- Édition rapide et intelligente de code
- Gestion de projet facile et efficace
- Développement rapide de l'interface utilisateur
- Ecriture de code sans erreur
- Support multilingue
- Support de plate-forme croisée
- Compatible avec de nombreux plugins fournis par la communauté

Voici son interface :

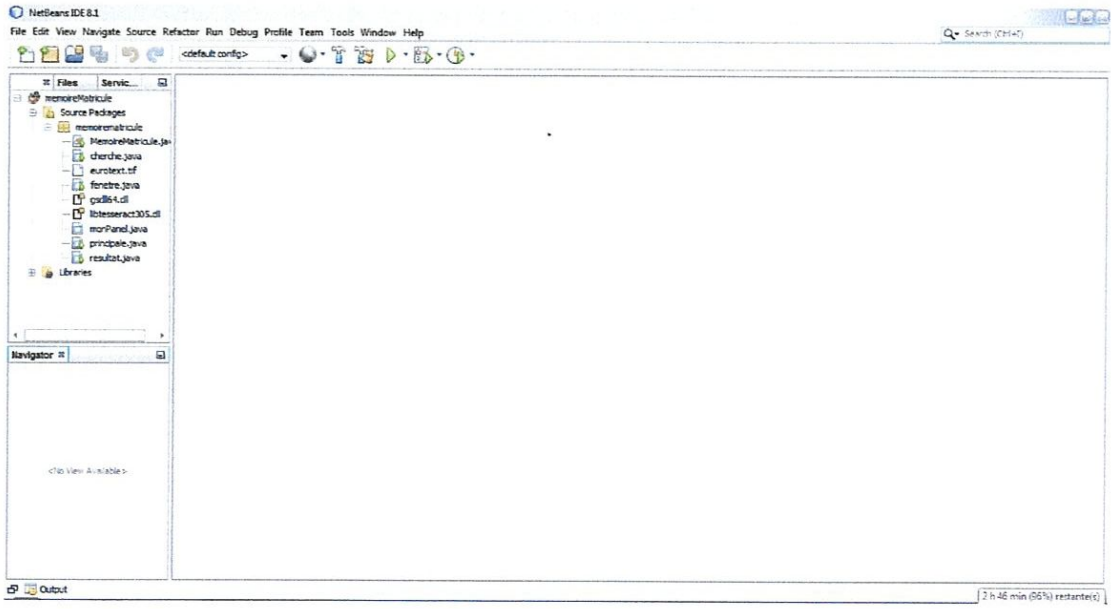


Figure IV.1: Interface de l'environnement de développement NetBeans version EDI.8.1

IV.3. Fonctionnement du système

Au démarrage le système reçoit comme donnée d'entrée une image d'un relevé du bac. Cette image est passée par la suite à une certaines étapes de « prétraitement » (transformation en

niveau de gris, binarisation), afin d'améliorer sa qualité. Après cela, l'image résultante est transférée au prochain module de « segmentation », qui inclut deux phases (l'élimination de bordure, et l'étiquetage des composantes connexes). A partir des composantes connexes étiquetées, le système sélectionne l'ensemble de composantes qui constituent la matricule. Le résultat obtenu est une sous-image incluant l'information importante qu'il est « le matricule ».

Puis, la sous-image résultante est transférée au module de « reconnaissance », afin de récupérer le matricule sous forme d'une chaine de caractère .Cette dernière est sauvegardée dans une base de données.

Au niveau du module de recherche, le système reçoit des requêtes textuelles correspondantes à des matricules. Dès qu'une requête est reçue, le système cherche dans la base de données le matricule identique à celle de la requête, puis retourne le relevé correspond.

IV.4. Description du corpus des documents utilisé

Pour l'expérimentation de notre application, nous avons utilisé un corpus composé d'images de relevés du bac numérisés au format « JPG ». Les documents de notre corpus de test sont de différentes structures et formats rendant leur traitement et analyse difficiles.

Cependant nous avons considéré uniquement des relevés de notes à partir de 1997. Les anciens relevés où le matricule est écrit en manuscrit ne sont pas pris en compte dans notre application. Les caractéristiques des pages de relevés de notes utilisées dans le test sont les suivantes : Chaque page peut contenir ;

- Un cadre
- Un entête
- Un titre
- Le numéro d'inscription « **Matricule** »
- Une partie du texte qui contient les informations personnelles de l'étudiant (nom, prénom, date et lieu de naissance, nom du lycée, filière du bac, mention) ;
- Un tableau qui contient les notes de l'étudiant et sa moyenne ;
- Partie graphique (le cachet + la signature).

Notre corpus est composé de 48 pages de relevé depuis bac « 1997 » jusqu'au bac « 2016 », le papier de ces documents est généralement de type épais ou standard et de format A4.

Voici quelques exemples d'images de notre corpus de test (***Figure IV.2***).

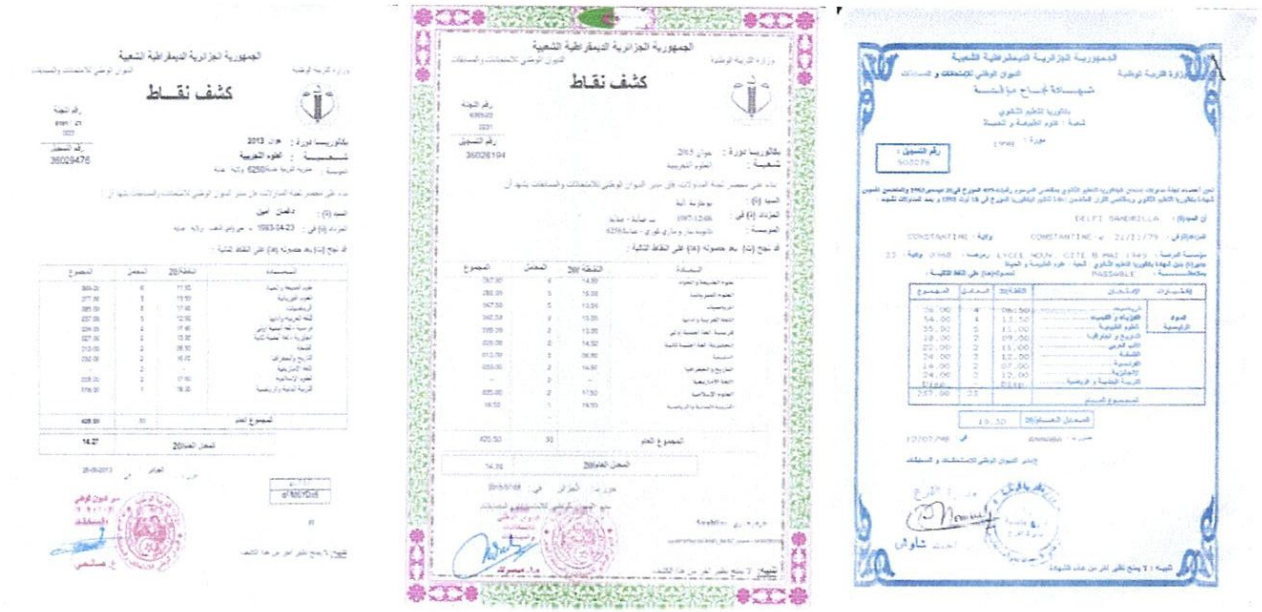


Figure IV.2 : Exemples d'images de notre corpus de test.

IV.5. Présentation de l'application

Dans cette partie on s'intéresse à la présentation de notre application, en se basant sur quelques exemples d'aperçus d'écran des différentes interfaces. Au lancement de l'application, l'interface suivante principale s'affiche (Voir la ***figure IV.3***):

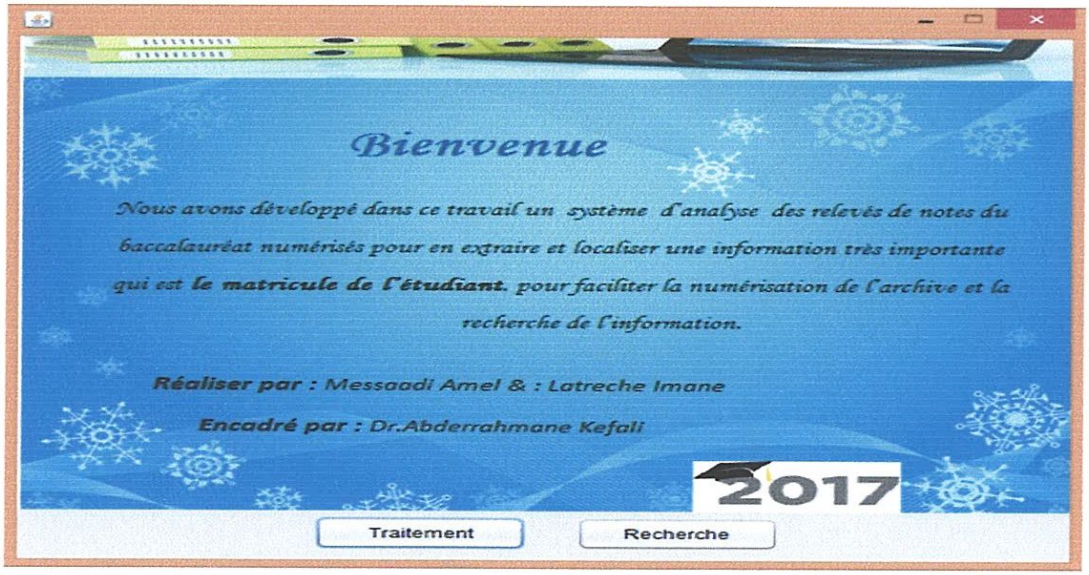


Figure IV.3 : Interface principale de notre application.

Cette interface nous permet d'accéder aux différentes fonctionnalités de l'application en cliquant sur les deux boutons :

1- **Bouton Traitement :** on l'utilise lorsqu'on veut procéder au traitement d'un nouveau relevé de bac. Il permet d'afficher la fenêtre qui regroupe toutes les traitements sur l'image de relevé de notes.

2- **Bouton Recherche :** on l'utilise lorsqu'on veut rechercher si un relevé de notes existe dans la base. La fenêtre correspondante contient un simple module de recherche.

IV.5.1. Traitement du relevé de bac

Lorsqu'on appuie sur le bouton « **Traitement** », la fenêtre illustrée par la figure suivante s'affiche (***Figure IV.4***):

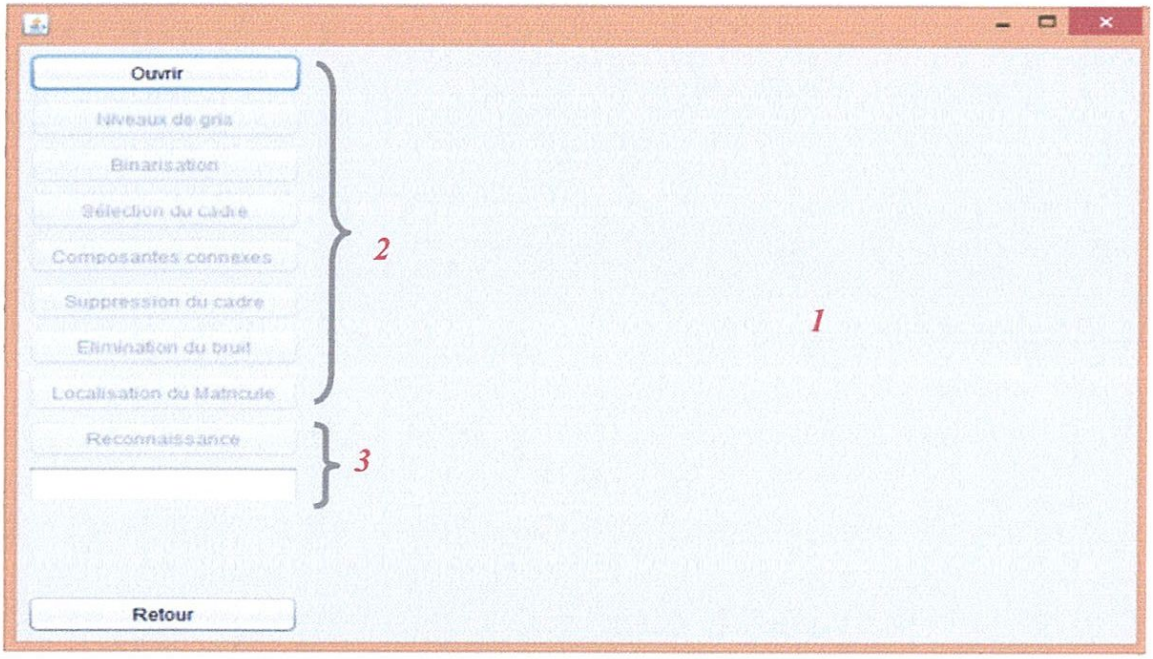


Figure IV.4 : Interface du traitement de relevés de bac.

La fenêtre contient les parties suivantes :

1- Zone d'affichage de l'image traitée.
2- Les bottons de traitement.
3- Bouton de **Reconnaissance** et zone d'affichage du résultat de reconnaissance.

IV.5.1.1. Chargement d'une image de relevé de notes

Nous commençons par le chargement de l'image à traiter à partir du bouton « Ouvrir ». Une boite de dialogue s'affiche nous permettant de sélectionner l'image de relevé à traiter. Cette dernière sera affichée dans la zone d'affichage (Zone 1), comme le montre la figure suivante (***Figure IV.5***) :

Figure IV.5 : Chargement d'une image.

IV.5.1.2. Transformation de l'image en niveaux de gris

Après le chargement de l'image, le bouton suivant « Niveaux de gris » devient activé. Ce bouton permet la transformation de l'image originale en niveaux de gris, et l'image résultante sera affichée dans un nouvel onglet (Voir la ***figure IV.6***).

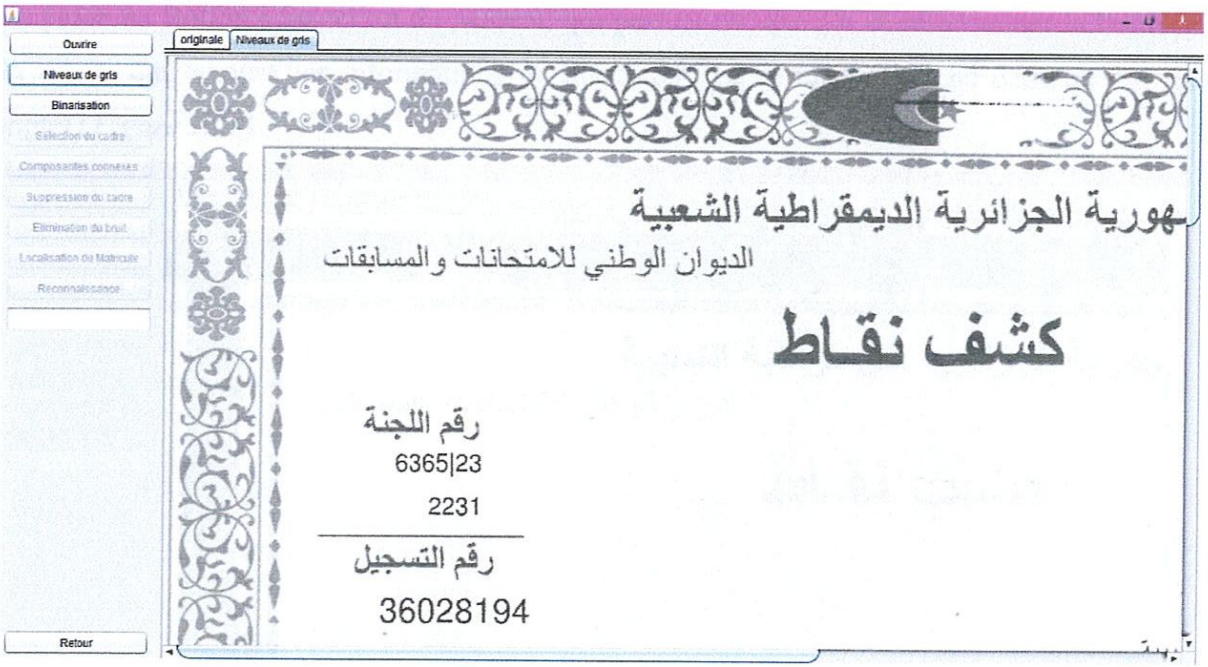


Figure IV.6 : Transformation de l'image en niveaux de gris.

IV.5.1.5. Etiquetage des composants connexes

En appuyant sur le bouton « Composantes connexes », les composantes connexes seront étiquetées chacune par une couleur distincte et l'image résultante est affichée dans un nouvel onglet (***Figure IV.9***).

Figure IV.9 : Etiquetage des composantes connexes.

IV.5.1.6. Suppression de la bordure

Lorsqu'on clique sur le bouton « Suppression du cadre », la bordure sera éliminée. Le résultat est affiché dans un nouvel onglet comme dans la figure suivante (***Figure IV.10***) :

Figure IV.10 : Suppression du cadre de l'image.

IV.5.1.7. Elimination du bruit

Le lissage de l'image est accessible depuis le bouton « Elimination du bruit ». Le résultat (dans un nouvel onglet) est la suivante (***Figure IV.11***):

Figure IV.11 : Elimination du bruit.

IV.5.1.8. Extraction du matricule

En cliquant sur le bouton « Localisation du matricule », le matricule sera sélectionné. Voici le résultat (***Figure IV.12)*** :

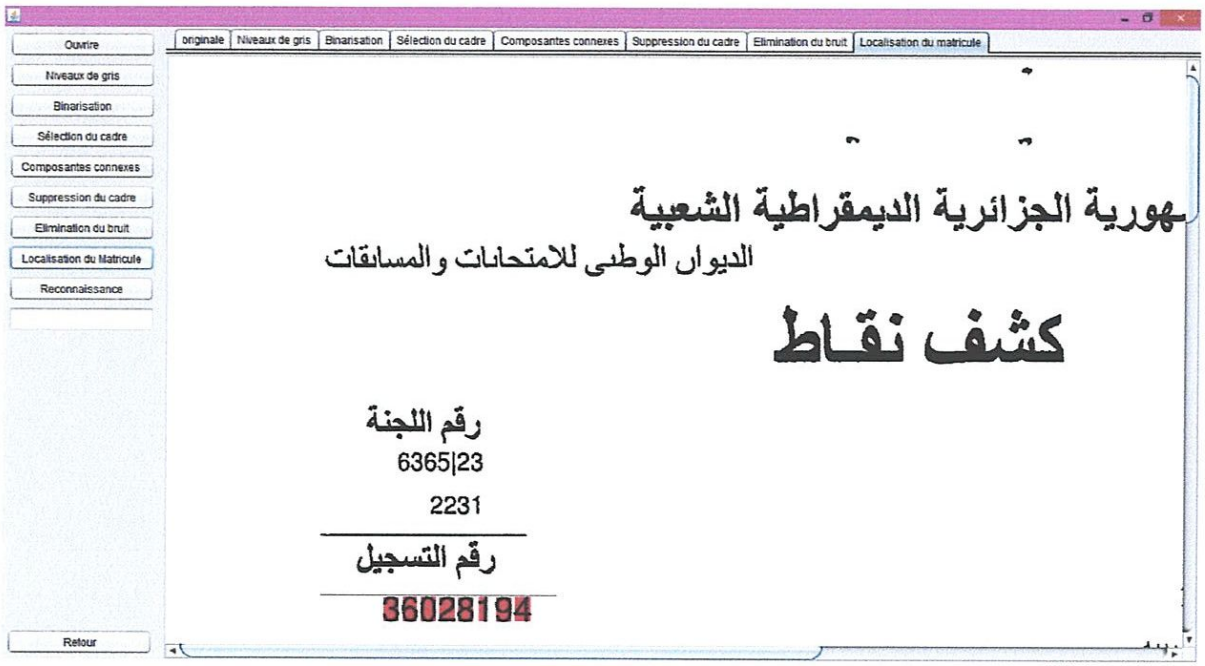


Figure IV.12 : Extraction du matricule.

IV.5.1.9. Reconnaissance du matricule

Le dernier bouton « Reconnaissance » fait appel à la bibliothèque de reconnaissance optique de caractères « Tess4J » pour procéder à la reconnaissance du matricule détecté précédemment. Le résultat de reconnaissance qui est le matricule sous forme de chaine de caractères est affiché dans la zone de texte correspondante comme dans la figure suivante (***Figure IV.13)***.

Figure IV.13 : Reconnaissance du matricule et affichage du matricule reconnu.

IV.5.2. Recherche du matricule

Lorsqu'on appuie sur le bouton « **Recherche»** dans l'interface principale**,** la fenêtre illustrée par la figure suivante s'affiche (***Figure IV.14***):

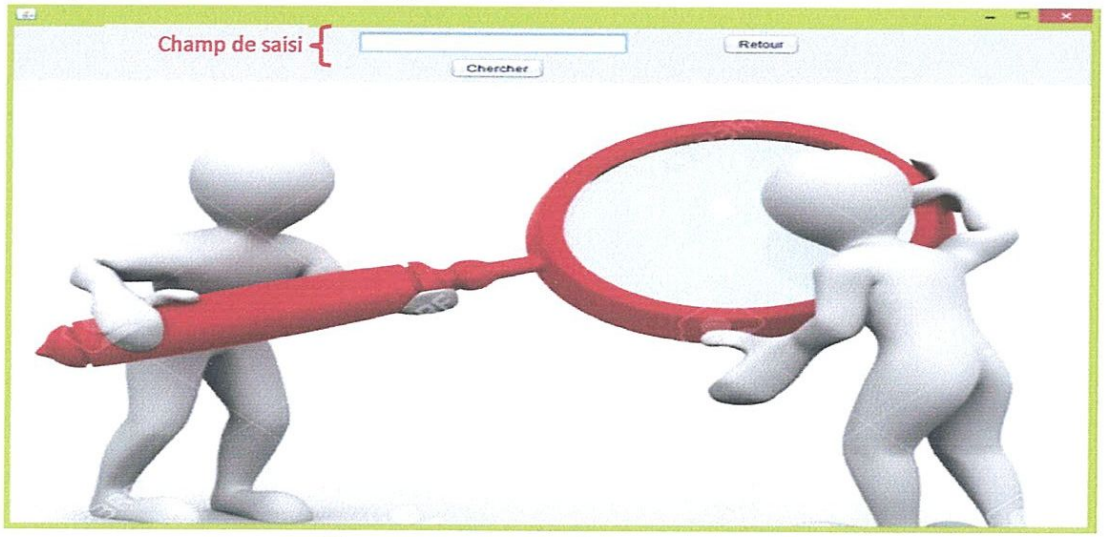


Figure IV.14 : Interface de recherche.

L'interface de recherche permet de retourner le relevé correspondant au matricule donnée en entrée par l'utilisateur dans le champ de saisi. La fenêtre suivante présente la requête de l'utilisateur (***Figure IV.15) :***

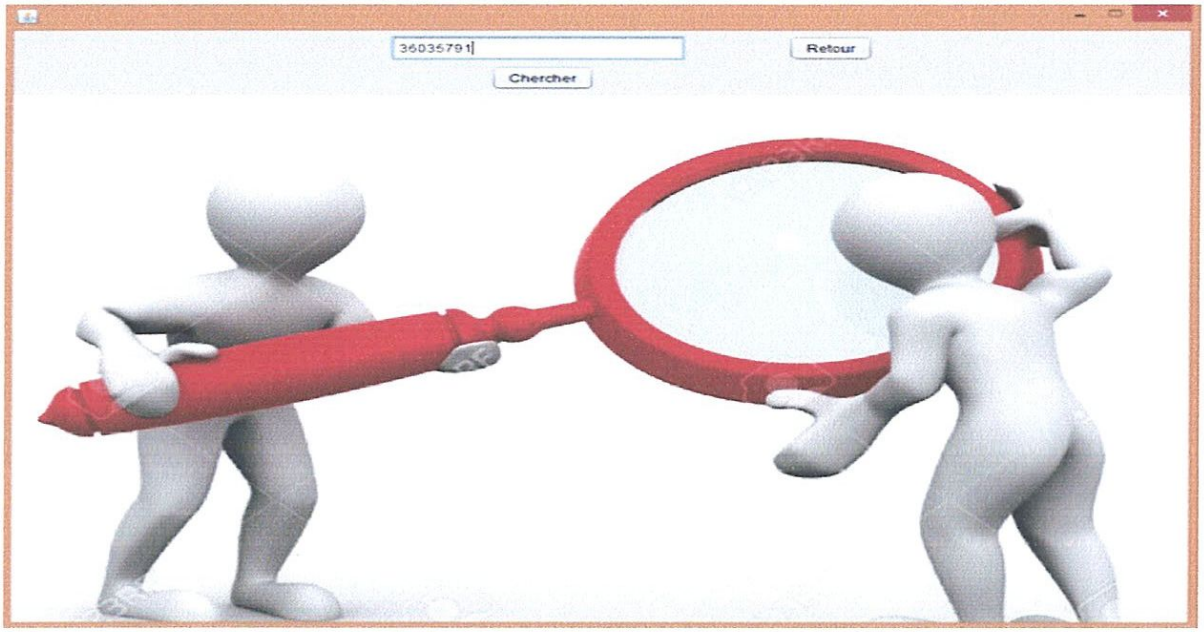


Figure IV.15 : Requête de l'utilisateur.

Après avoir appuyé sur le bouton « **chercher** », le système cherche dans la base si un relevé de notes correspond au matricule recherché ou pas.

Si le matricule cherché est trouvé dans la base, le système retourne le relevé correspond au matricule. Voici le résultat (***Figure IV.16)***

Figure IV.16 : Le relevé cherché.

Si le matricule n'est pas reconnu, un message sera affiché indiquant que le relevé n'existe pas. Voici le résultat (***Figure IV.17)*** :

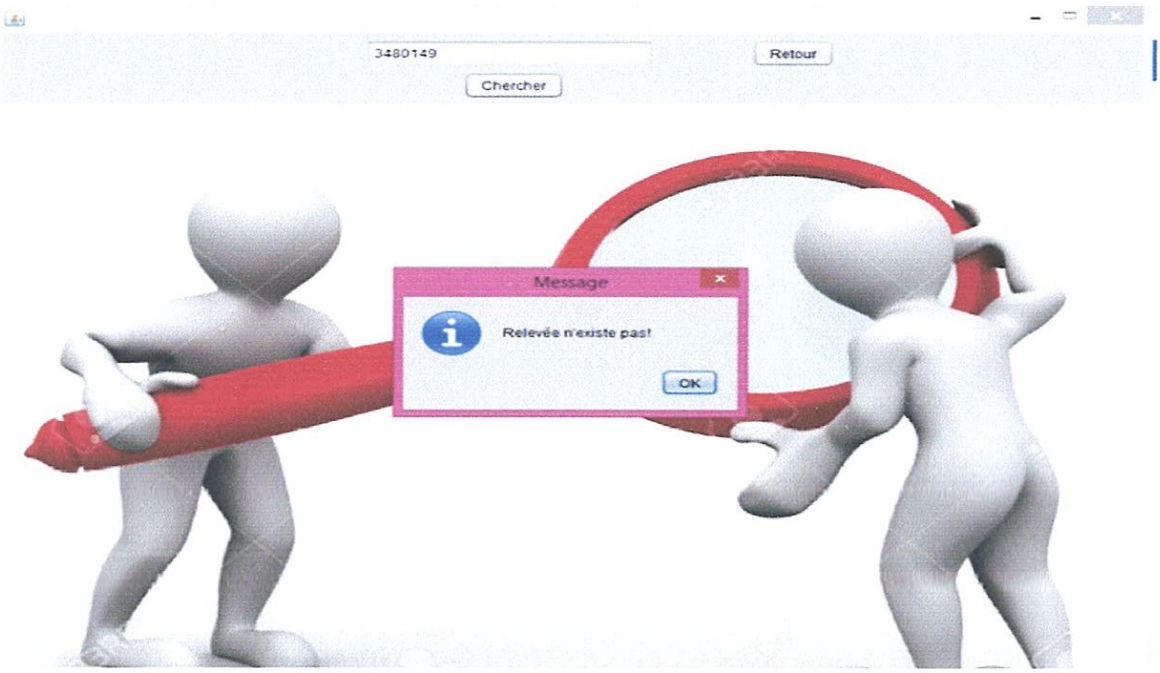


Figure IV.17 : Message d'erreur lorsque le matricule cherché n'existe pas.

IV.6. Expérimentations et résultats

Le système proposé a été appliqué sur les 48 images de relevés des notes du corpus de test en vue d'évaluer ses performances. L'évaluation a été effectuée en termes de : temps de réponse moyen, taux de détection des matricules, et taux de reconnaissance des matricules détectés.

Les tests ont montré que le temps de réponse varie entre 42 et 60 secondes, et dépend de la structure du relevé.

Le ***tableau IV.1*** (page suivante) récapitule les résultats de détection et de reconnaissance pour toutes images du corpus de test.

Relevé	Matricule bien détecté	Matricule réel	Matricule reconnu	Nb de chiffres reconnus
1-1997	×	511246	511246	6/6
2-1998	×	503276	503276	6/6
3-2000	×	654269	654269	6/6
4-2000	×	621406	621406	6/6
5-2001	×	668352	668352	6/6
6-2001	×	647516	647516	6/6
7-2001	×	644190	644190	6/6
8-2002	×	646867	646867	6/6
9-2007	×	6020266	6020266	7/7
10-2008	×	8047283	8047283	7/7
11-2009	×	6018993	6018993	7/7
12-2010	×	6005265	6005265	7/7
13-2011	×	6008557	6008557	7/7
14-2012	×	9051810	9051810	7/7
15-2013	×	36029176	36029176	8/8
16-2013	×	36030202	36030202	8/8
17-2013	×	36021905	36021905	8/8
18-2013	×	36021905	36021905	8/8
19-2013	×	36004234	36004234	8/8
20-2013	×	36039571	36039571	8/8
21-2013	×	36042341	36042341	8/8
22-2013	×	36025325	36025325	8/8
23-2013	×	34021199	34021199	8/8
24-2013	×	36044829	36044829	8/8
25-2013	×	36008879	36008879	8/8
26-2013	×	36051209	36051209	8/8
27-2013	×	39049815	39049815	8/8
28-2013	×	35062090	35062090	8/8
29-2013	×	37000293	37000293	8/8
30-2013	×	34014990	34014990	8/8
31-2013	×	38041454	38041454	8/8
32-2014	×	36035791	36035791	8/8
33-2014	×	36024997	36024997	8/8
34-2014	×	36048206	36048206	8/8
35-2014	×	36053901	36053901	8/8
36-2014	×	36025326	36025326	8/8
37-2014	×	36030324	36030324	8/8
38-2015	×	36028194	36028194	8/8
39-2015	×	36028194	36028194	8/8
40-2015	×	36031476	36031476	8/8
41-2015	×	36005759	36005759	8/8
42-2015	×	36004730	36004730	8/8
43-2015	×	36027866	36027866	8/8
44-2015	×	36030687	36030687	8/8
45-2015	--	36036200	36036200	0/8
46-2015	×	36057921	36057921	8/8
47-2015	×	34027389	34027389	8/8
48-2016	×	35052176	36052170	6/8
Nb de matricules bien détectés	47/48	**Nb de matricules bien reconnus**	46/48	352/362

Tableau IV.1 : Résultats de détection et de reconnaissance obtenus pour toutes les images.

A partir du tableau ci-dessus, nous pouvons tirer le taux final de détection et de reconnaissance :

- Le taux de réussite d'extraction des matricules est de : (47/48) × 100 = **97.91%**.
- Le taux de reconnaissance des matricules est de : (46/47) × 100 = **97.87%**.

En effet, le problème a été rencontré avec le relevé numéro « 45-2015 », car le matricule dans ce relevé et l'écriture arabe sont collés (***Figure IV.18***).

Figure IV.18 : Exemple d'un matricule non détecté (relevé 45-2015).

Aussi, l'autre problème a été rencontré avec le relevé numéro « 48-2016 ». Avec ce relevé, la détection a été effectuée correctement mais la reconnaissance est échouée ; c'est parce que la bibliothèque utilisée n'a pas pu reconnaître tous les chiffres du matricule : 35052176 ≠ 36052170.

IV.7. Conclusion

Au terme de ce chapitre, on a présenté une vue complète sur les étapes de réalisation de notre application à partir des différentes interfaces capturées. Tout au long du processus de développement. Nous avons découvert plusieurs outils de programmation tout en ayant l'occasion d'apprécier leurs intérêts au développeur.

Conclusion générale et Perspectives

Le travail adressé dans ce mémoire a pour but la localisation du matricule dans les relevées de notes du BAC. Cette tâche s'avère une nécessité pour offrir une manipulation plus aisée des données : archivage, indexation, recherche, etc. Comme nous avons dit précédemment, la compréhension d'un document nécessite la reconnaissance de sa structure en plus de son contenu textuel puis on peut localiser n'importe qu'elle information constituant ce document.

Dans ce mémoire, nous avons réalisé un système capable d'extraire le matricule des bacheliers à partir des relevées de bac de différents styles et formats.

A travers les résultats obtenus on peut dire que notre système est fiable ce qui implique que l'approche utilisé pour l'analyse de documents est efficace.

Comme tout travail n'est pas complet plusieurs extensions sont envisageables :

- Localisation d'autres zones d'intérêt (nom, prénom, date de naissance, moyenne de bac, mention, etc.
- Intégration d'autres étapes de prétraitement notamment afin de tenir compte d'autres problèmes : inclinaison de documents, dégradation des documents, etc.
- Développement d'un module de reconnaissance de chiffres et l'intégrer dans notre système.
- Etendre l'application développée à la reconnaissance de la structure complète des relevées de notes.
- Conception d'un réel système d'indexation et recherche des informations sur les étudiants.

Bibliographie

[ABB 09] R. Abbassi, C.A. Akono, S.D. Bouzidi, L. Dinomais, S. Longuet, D. Sarenac, « L'archivage électronique », *Direction des archives de France*, 2009.

[AGH 94] H.K. Aghajan and T. Kailath, « SLIDE: Subspace-based line detection», *Pattern Analysis and Machine Intelligence*, vol. 16, No.11, pp. 1057-1073, 1994.

[ANT 98] A. Antonacopoulos, « Page Segmentation Using the Description of the Background», *Computer Vision and Image Understanding*, vol. 70, No. 3, pp 350-369, 1998.

[AZO 95] A.S. Azokly, « Une approche uniforme pour la reconnaissance de la structure physique de documents composites fondée sur l'analyse des espaces», *Thèse de doctorat*, Université de Fribourg, 1995.

[BAC 98] B. Bachimont, « Bibliothèques numériques audiovisuelles: des enjeux scientifiques et techniques», *Document numérique*, pp. 2.3-4, 1998.

[BAI 90] H.S. Baird, S.E. Jones, S. Fortune, « Image Segmentation by Shape Directed Covers», *International Conference on Pattern Recognition*, pp. 820-825, 1990.

[CHA 91] J.M. Chassery, M. Melkemi, « Diagramme de Voronoï appliqué à la segmentation d'images et à la détection d'événements en imagerie multisources », *Traitement du signal*, pp 155–164, 1991.

[CHA 10] M.A. Chabin, « Nouveau glossaire de l'archivage [en ligne] », Archive 17, 2010.

[COL 04] A. Colin, S. Cacaly, Y.F. Le Coadic, P.D. Pomart, et al., « Dictionnaire de l'information », Paris : 2ème édition, pp 7, 2004.

[DRI 95] D. Drivas, A. Amin , «Page Segmentation and Classification Utilizing Bottom-Up Approach», *Proceedings of the 3rd International Conference on Document Analysis and Recognition*, pp. 610-614, Montreal, Canada, 1995.

[EMP 03] H. Emptoz, F. Lebourgeois, V. Eglin, Y. Leydier, « La reconnaissance dans les images numérisées: OCR et transcription, reconnaissance des structures fonctionnelles et des méta-données », *Journées d'étude Numérisation des textes et des images*, pp.105-129, 2003.

[FIS 90] J.L. Fisher, C.H.Stuart, P.D.Donald , «A rule-based system for document image segmentation », *Proceedings of the 10th International Conference on Pattern Recognition*, vol. 1, pp. 567-572, 1990.

[GRO 10] Groupe de travail «archivage électronique» de la Fédération Nationale des Tiers de Confiance, « Guide De L'archivage Electronique et du Coffre-Fort Electronique », *Collection Les Guides de la Confiance de la FNTC*, 2010.

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

[KET 10] D. Ketata, K. Maher, « Un survol sur l'analyse et la reconnaissance de documents: imprimé, ancien et manuscrit », *Colloque International Francophone sur l'Ecrit et le Document (CIFED)*, pp.12, 2010.

[LEB 92] F. Lebourgeois, Z. Bublinski, H. Emptoz, «A fast and efficient method for extracting text paragraphs and graphics from unconstrained documents», *Conference B: Pattern Recognition Methodology and Systems, Proceedings of the 11th IAPR International Conference on Pattern Recognition*, vol.2, pp. 272-276, 1992

[NAG 00] G. Nagy, «Twenty years of document image analysis in PAMI», *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, No.1, pp. 38-62, 2000.

[NLA 88] G. Nagy, J. Kanai, M. Krishnamoorthy, M. Thomas, M. Viswanathan, «Two complementary techniques for digitized document analysis», *DOCPROCS : Proceedings of the ACM conference on Document Processing Systems*, pp169–176, 1988.

[OTS 79] N. Otsu, « A threshold selection method from gray-level histograms », *IEEE Transactions on Systems, Man and Cybernetics*, vol. 9, No. 1, pp. 62–66, 1979.

[PAV 91] T. Pavlidis, J. Zhou, « Page Segmentation by White Streams», *International Conference on Document Analysis and Recognition*, pp 945-953, 1991.

[PAV 92] T. Pavlidis, J. Zhou, « Page Segmentation and Classification », *Graphical Models and Image Processing*, pp. 484-496, 1992.

[ROB 01] L. Robadey, « (CREM): Une méthode de reconnaissance structurelle de documents complexes basée sur des patterns bidimensionnels », *Thèse de doctorat*, Université de Fribourg, 2001.

[SAE 14] Service Archives, « Système d'archivage électronique (SAE) », *Service archives Fiche pratique*, vol. 3, pp. 825-837, Octobre 2014 [en ligne https://www.cigversailles.fr/download/file/a2111624-b811-41fb-9338-5abf35e95f31].

[SOU 02] S. Souafi-Bensafi, « Contribution à la reconnaissance des structures des documents écrits: approche probabiliste », Thèse de doctorat, Villeurbanne, INSA de Lyon, 2002.

[SOY 11] S. Soyez, M. Layeux, « Comment classer mes documents », *Brochures de recommandations et de conseil*, Version décembre 2011, [en ligne www.arch.be/docs/brochures/classer_documents.pdf]

[TRI 95] O.D. Trier, T.Tax, « Evaluation of binarization methods for document images », *Pattern Analysis and Machine Intelligence*, vol. 11, No. 12, pp. 312-314, December 1995.

[TRU 05] E. Trupin, « La reconnaissance d'images de documents: Un panorama », Traitement du signal, vol. 22, No.3, pp. 159-189, 2005.

[WAN 89] D. Wang, N.S .Sargur, N. Srihari, «Classification of newspaper image blocks using texture analysis», *Computer Vision, Graphics, and Image Processing*, vol. 47, No.3, pp. 327-352, 1989.

[WON 82] K.Y. Wong, R.G. Casey, F.M. Wahl, « Document analysis system », *IBM journal of research and development*, vol. 26, No. 6, pp. 647–656, 1982.

[WEB 1] http://dictionnaire.sensagent.leparisien.fr/Archivage%20%C3%A9lectronique/fr-fr/ Consulté le 02/02/2017.

[WEB 2] https://zagoarchivesblog.wordpress.com Consulté le 15/02/2017.

[WEB 3] http://dictionnaire.sensagent.leparisien.fr/Archivage%20%C3%A9lectronique/fr-fr/ Consulté le 15/02/2017.

[WEB 4] http://kost-ceco.ch/cms/index.php?minimal_specifications_fr Consulté le 10/02/2017.

[WEB 5] http://wiki.maarch.org/introduction_a_l'archivage_electronique Consulté le 15/02/2017.

[WEB 6] http://www.archivesfactory.com/blog/23-l-essentiel-sur-l-archivage-legal/389-quest-ce-que-larchivage-electronique Consulté le 16/02/2017.

[WEB 7] https://liris.cnrs.fr/~veglin/Master3_2011Numérisation des documents patrimoniaux/

[WEB 8] https://www.jmdoudoux.fr/java/dej/chap-presentation.htm Consulté le 15/05/2017.

[WEB 9] http://www.freewarefiles.com/NetBeans-IDE_program_93006.html Consulté 15/05/ 2017.