الجمهورية الجزائرية الديمقراطية الشعبية

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 : Systèmes et technologie de l'information et de la communication

Thème :

Vers la construction d'une base d'images de documents arabes dégradés synthétiques

Encadré Par :

Dr. Abderrahmane Kefali

Présenté par :

Houssam Houmeur

Octobre 2020

Abstract

Researchers in the field of image and writing processing have long been interested in building databases, the objective of which is to provide large learning and testing collections that allow researchers to test and evaluate their methods on a standard set of images. These databases must contain, in addition to images, reference information (so-called *ground truth*) expressing the ideal results expected from certain processing steps, in order to be able to objectively evaluate the results obtained. Most of the time, this ground truth information is prepared manually.

For old Arab documents, which are the heart of our work, the problem still remains. Almost no database is currently available to the general public and little research has gone into building a database of old documents. In fact, several difficulties persist, among them, *(a)* a large part of the Arab and Islamic documents are scattered around the world (in specific families or in mosques or *Zawaia*) and are not kept in specialized institutions, *(b)* The difficulty of manually establishing ground truth information, because of the high costs, *(c)* the degradation characteristics and the structural complexity of old documents in addition to the absence of sufficient information on these images, further complicate the construction of a database of ancient Arabic documents.

In this project, we are interested in the construction of an images database of synthetic degraded Arabic documents but in the opposite direction, i.e. from ground truth texts and images to synthetic images of degraded documents. We thus offer a tool allowing the automatic creation of synthetic document images by combining text images with old background images, in addition to noisy images with different types of degradation related to old documents, by inspiring from certain works of noise and degradation modeling. The constructed database also contains other ground truth information (number of text lines, their position, number of words in each line, etc.) obtained after several stages of analysis and segmentation. Thus, the constructed database contains three levels of ground truth information: ground truth texts, binary ground truth images, and ground truth annotation files, which makes our base suitable for several applications of documents analysis and recognition.

Résumé

Les chercheurs dans le domaine du traitement d'images et de l'écriture ont été intéressés depuis longtemps à la construction des bases de données, dont l'objectif de fournir des grandes collections d'apprentissage et de test permettant aux chercheurs de tester et d'évaluer leurs méthodes sur un ensemble standard d'images. Ces bases de données doivent contenir, en plus des images, des informations de référence (dites de *vérité terrain*) expriment les résultats idéals attendus de certaines étapes de traitement, afin de pouvoir évaluer objectivement les résultats obtenus. Dans la plupart du temps, ces informations de vérité terrain sont préparées manuellement.

Pour les anciens documents Arabes, qui sont le cœur de notre travail, le problème reste toujours posé. Presqu'aucune base de données, actuellement, n'est disponible au grand public et peu de travaux de recherche ce sont attaqué à la construction d'une base d'anciens documents. En fait, plusieurs difficultés persistent, parmi eux, (*a)* une grande partie des documents arabes et islamiques sont éparpillés à travers le monde (chez des familles spécifiques ou dans des mosquées ou des *Zawaia*) et ne sont pas conservés dans les institutions spécialisées, *(b)* la difficulté de l'établissement manuel des informations de vérité terrain, à cause des couts élevés, *(c)* les caractéristiques de dégradation et la complexité de structure des documents anciens en plus de l'absence des informations suffisantes sur ces images, compliquent de plus la construction d'une base d'anciens documents Arabes.

Dans ce projet, nous nous intéressons à la construction d'une base d'images de documents arabes dégradés synthétiques mais dans le sens inverse c'est à dire à partir de textes et d'images de vérité terrain jusqu'aux images de documents dégradés synthétiques. Nous proposons ainsi un outil permettant la création automatique des images de documents synthétiques par la combinaison des images de texte avec des images de fond anciens, en plus des images bruitées avec différents types de dégradations liées documents anciens en s'inspirant de certains travaux de modélisation de bruit et de dégradation. La base construite contient également d'autres informations de vérité terrain (nombre de lignes de texte, leur position, nombre de mots dans chaque ligne, etc.) obtenues après plusieurs étapes d'analyse et de segmentation. Ainsi, la base construite contient trois niveaux d'informations de vérité terrain: textes de vérité terrain, images binaire de vérité terrain, et fichiers d'annotation de vérité terrain, ce qui rend notre base appropriée pour plusieurs applications d'analyse et de reconnaissance de documents.

Table de matières

Table de figures

Liste des tableaux

Introduction générale

Le développement des civilisations et de leurs cultures permet de conserver et de transmettre les connaissances d'une génération à l'autre. Cependant, les documents joue un rôle important pour ça; ils ont été passés par différentes étapes de développement, depuis la documentation sur des supports physiques solides (le bois, la pierre, le marbre, les galets, les tablettes de cire ou d'argile) jusqu'aux supports flexibles (le papyrus, le parchemin, le cuir, le papier). En effet, le papier est considérer comme le support le moins coûteux et le plus commode, et pour cette raison il s'est généralisé dans le monde comme le principal support de sauvegarde et de transmission de l'information. Mais, le papier est un support fragile dont la conservation est difficile. Aujourd'hui dans le monde du numérique le document électronique est le vecteur de diffusion le plus utilisé dans le monde car il possède de nombreux avantages par rapport aux documents papier [DRI 07].

Dans le monde des documents numériques, l'existence d'une bases de données pour les chercheurs dans tous les domaines est importante et surtout dans les domaines qui traitent des quantités massives de données comme *le traitement et l'analyse d'images de document.* La détection de la ligne de base, la segmentation, la localisation des zones d'intérêts, la reconnaissance de l'écriture,... sont des applications qui nécessitent la disponibilité de grandes bases de données pour l'apprentissage et le test. Ses bases doivent répondre, au maximum, aux problèmes existants dans le domaine étudié afin de donner un espace d'évaluation des résultats des travaux. Ainsi, lors de la création d'une base de données, vous devez prendre en compte certains critères importants. La base doit être riche en qualité et en quantité; elle doit contenir un grand nombre d'informations et en même temps une grande variation de ces informations. Elle doit également inclure diverses informations sur le document (taille, type, domaine/contenu, scripteur, origine, nombre des lignes, période, etc.). Un autre critère important est qu'elle doit contenir des informations de référence (dites informations *de vérité terrain*). Ces informations, stockées sous diverses formes (image, fichiers XML,TXT,...), expriment les résultats finaux espérés, ou les résultats idéals attendus de certaines étapes de traitement (comme la binarisation, la segmentation, la localisation, etc.).

Cependant, avec le traitement des documents Arabes anciens, qui sont le cœur de notre travail, le problème posé c'est que presqu'aucune base de données, actuellement, n'est disponible au grand public et, une grande partie des documents Arabes et islamiques sont éparpillés à travers le monde (chez des familles spécifiques ou dans des mosquées ou *Zawaya*) et ne sont pas conservés dans les institutions spécialisées. La difficulté de l'établissement manuel des images de vérité terrain, à cause des caractéristiques de

dégradation et la complexité de structure des documents anciens en plus de l'absence des informations suffisantes sur ces images, compliquent de plus la construction d'une base d'anciens documents Arabes.

Dans ce mémoire, nous nous intéressons à la construction d'une base d'images de documents arabes dégradés synthétique mais dans le sens inverse (des images de vérité terrain jusqu'aux images de documents bruitées). Nous proposons un outil permettant la création automatique des images de documents bruités avec différents types de bruit sur les images (bruit local et global de Kanungo, l'effet de transparence, l'effet de rotation, etc.) en plus des images par combinaison texte/fond des documents anciens.

Le mémoire est structuré autour de quatre chapitres organisés de la manière suivante :

Chapitre 1 : Les bases d'images de document

Ce chapitre présente une vue globale sur le développement et la conception de bases de données pour l'analyse et la reconnaissance de documents.

Chapitre 2 : Les bruits et Segmentation d'images de documents

Le deuxième chapitre sera consacré aux différents types de bruit et segmentation qui peuvent être trouvé dans les images de documents.

Chapitre 3 : Contribution

Le troisième chapitre détaille les différentes étapes réunies pour la réalisation de notre allure suivie pour la construction de la base.

Chapitre 4 : Implémentation et résultats

Ce chapitre illustre l'implémentation de notre application, ainsi que les résultats obtenus.

Chapitre 1. Les bases d'images de document

1. Introduction

La numérisation de documents et la reconnaissance automatique de l'écriture manuscrite ont fait un grand pas en avant au vu du travail intensif dans ce domaine et de la publication de grandes bases de données liées à ce domaine, qui ont permis aux chercheurs de fournir des rapports fiables et de comparer leurs résultats avec de nombreux autres résultats de recherche.

Nous essayons à travers ce chapitre de mettre l'accent sur la conception et le développement des bases d'images de documents publiques tout en survolant les principales bases existantes.

La suite de ce chapitre est organisée comme suit. Une présentation de la classification des bases de données. Ensuite, une vue globale sur les caractéristiques générales des bases de données de textes arabes. Finalement, nous mentionnons quelques bases de données publiques disponibles et utilisées par une partie de la communauté.

2. Catégories des bases de données pour l'analyse de documents

D'après [ELA 08], les bases de données dans le domaine de l'analyse de documents et la reconnaissance de l'écriture sont souvent classées en deux catégories, en se basant sur la nature et l'origine des données, comme suit:

2.1. Bases de données réelles

Dans ces bases de données, les données sont directement recuites de l'application réelle. Une caractéristique très importante pour ces bases de données est le fait qu'on a un à faire avec un acteur "ignorant", par exemple pour la reconnaissance de l'écriture manuscrite le scripteur n'est pas au courant de la procédure de la reconnaissance au moment où il a rédigé son texte. Cette caractéristique a aussi une facette négative, pour des raisons morales et juridiques la publication de ces données, parfois personnelles, est très restreinte. Ces restrictions sont l'une des raisons pour lesquels on constate une manque de présence de telles bases de données [ELA 08].

2.2. Bases de données synthétiques ou artificielles

La deuxième catégorie est celle des données synthétiques ou artificielles. En fait, ce type de bases de données est apparu à cause des coûts élevés des bases de données commerciales et des phases de collections de données. Ces coûts ont obligé les chercheurs à chercher d'autres moyens de collecter les données pour l'apprentissage et l'évaluation.

Cependant plusieurs groupes de recherche ont développé des méthodes de création de base de données synthétiques [KAN 99]. Ces données ont eu pour buts de faciliter les taches de

reconnaissance et d'évaluation des systèmes de reconnaissance, et surtout les systèmes basés sur les méthodes statistiques, qui ont besoin d'une grande quantité de données pour atteindre de bonnes performances[ELA 08].

Un exemple de ces travaux est celui de Märgner et Pechwitz [MÄR 01] qui ont présenté un système de génération de données synthétiques a partir d'un texte arabe imprimé. Ce système est basé sur la génération automatique de données en différents formats, ainsi que les images et les annotations correspondantes (utilisant LATEX). Une variété de méthodes de dégradation d'images est aussi présentée.

3. Caractéristiques générales des bases de données de textes arabes

Les caractéristiques présentées dans cette section doivent être souvent prises en compte dans le processus de développement des bases de données. On peut classifier ces caractéristiques en trois catégories [ELA 08]:

- ***La classification des données.*** Il peut s'agir de données liées directement à des applications industrielles ou de données synthétiques. Dans ce dernier cas, les données sont généralement collectées en utilisant des formulaires bien spécifiques, qui permettent d'accélérer et de faciliter les étapes ultérieures de la collection des données.
- ***La présence d'annotations*** (données de vérité terrain, anglais : label ou Ground Truth). C'est l'une des caractéristiques les plus importantes (si elle n'est pas la plus importante) des bases de données utilisées pour l'apprentissage et l'évaluation des systèmes d'analyse et de reconnaissance. Ces informations sont toujours présentes dans une base de données (sinon on ne pourrait pas effectuer les phases d'apprentissage et d'évaluation d'un système de reconnaissance). Elles peuvent être enregistrées sous formes de fichiers XML ou autres formats. Elles peuvent être classifiées en plusieurs types :
  - Dans le cas d'annotation de données contenant des lignes de texte, il est important d'avoir l'information sur la position du mot dans le texte.
  - Dans le cas des bases de données composées de mots, c'est important de connaître la séquence des caractères dans le mot. Ce point est important surtout pour les données de l'écriture arabe, car les caractères prennent des formes différentes en fonction de leurs positions dans un mot.
  - Images représentant le résultat d'un traitement dans le cas où la base de données est conçue pour l'évaluation d'une des étapes d'un système d'analyse et de reconnaissance de texte.

- ***Caractéristiques optionnelles.*** Elles peuvent enrichir une base de données et élargir ces champs d'applications. Parmi ces caractéristiques on peut lister :
  - Des informations sur le scripteur, par exemple âge, profession, genre.
  - Des informations sur les sources des images et des données
  - Des informations sur l'organisation générale des données, par exemple l'information sur l'appartenance d'une image à un ensemble d'apprentissage ou un ensemble d'évaluation.
  - Des informations de qualité de l'écriture, de l'image, etc.
  - Des informations sur la position des lignes de base, des ligatures, etc.

4. Principales bases de données existantes

Les bases de données annotées utilisées dans les travaux de recherche sur les documents et l'écriture sont nombreux et variées. Aussi nombreux que le nombre de ces travaux et variées d'autant que le type des documents traités et l'objectif visé du travail. Malheureusement, la plupart de ces bases de données ne sont plus accessibles, elles étaient développées pour un travail de recherche bien défini. Nous présentons certaines de ces bases de données.

4.1. Les bases d'écriture arabe

4.1.1. base IFN/ENIT

Construite en 2002 par l'institut des technologies de communications (IFN) en coopération avec l'école nationale d'ingénieurs de Tunis (ENIT). Cette base de données contient des noms manuscrits de villes / villages arabes (tunisiens) [PEC 02].

La base de données IFN / ENIT contient 26 759 mots manuscrits représentant 956 classes et 210 000 caractères écrits par 411 scripteurs. Chaque mot est livré avec une image et des informations de vérité terrain (d'annotation). Ainsi, pour chaque nom, les informations de base, par exemple (l'ordre des formes de caractère, les informations sur le style de l'écriture, la présence des signes diacritiques secondaires, et une approximation de la ligne de base) sont codées. Le tableau 1.1 présente un exemple.

Image	حمّام بياضة	رؤاد
Ground truth:		
Postcode	6132	2056
Global word	حمّام بياضة	رؤاد
Character shape sequence	|A_م |E_ا |M_L_م |B_ح |E_ة |B_ض |E_ا |M_ي |B_ب	|A_د |A_ا |A_L_و |A_ر
Baseline y1,y2	70,50	46,39
Baseline quality	B1 (B1=OK; B2=bad)	B1
Quantity of words	2	1
Quantity of PAWs	4	4
Quantity of characters	9	4
Writing quality	W1 (W1=OK; W2=bad)	W1

Tableau 1.1. *Exemple d'annotation de deux images de la base IFN/ENIT [PEC 02].*

Les données de cette base sont automatiquement extraites à partir du nom imprimé correspondant de chaque image et elles sont manuellement vérifiées et modifiées.

Dans sa version définitive, l'IFN / ENIT est divisé en neuf ensembles. Depuis la présentation de cette base de données gratuite, plusieurs recherches des équipes ont été évaluées sur leurs résultats [ELA 08].

4.1.2. La base AHCR (Arabic Handwritten Character Recognition)

Cette base a été proposée par Alkhateeb en 2015 [ALK 15].Pour être éligible pour la recherche sur la reconnaissance manuscrite arabe. Elle contient des images numérisées des lettres arabes écrites par 100 différents scripteurs arabes natifs. Chaque scripteur a été interrogé pour remplir un formulaire, où le scripteur doit écrire chaque lettre des 28 lettres de l'alphabet arabe 10 fois, ce qui donne 280 dans chaque formulaire. Les scripteurs garantissent une grande variété de styles d'écriture.

Après, 28000 images de lettres arabes valides ont été recadrées et extraites manuellement des formulaires. Sur ces images, un filtrage Médian, a été appliqué afin d'améliorer la qualité d'images numérisées. Finalement les images sont séparées en deux dossiers: *Apprentissage*, contenant 80% des images de la base, et *Test* contenant les 20% restantes.

4.1.3. La base AHTID / MW (Arabic Handwritten Text Images Database written by Multiple Writers

La base de données d'images de textes manuscrits en arabe écrite par plusieurs scripteurs (AHTID / MW) a été construite au MIRACL Lab, ISIMS, Université de Sfax - Tunisie en

collaboration avec l'Institut des technologies de la communication (IfN), Braunschweig– Allemagne [SLI 14].

Cette basse contient 3710 lignes de texte et 22896 mots écrits par 53 scripteurs natifs arabe. Pour chaque élément (image de ligne ou mot) de la base de données, un le fichier de vérité terrain correspondant sous forme XML est accompagné. La base de données peut s'avérer utile pour diverses applications de recherche telles que les systèmes de reconnaissance de texte manuscrit arabe, la segmentation en mots, le repérage de mots, et l'identification de scripteurs. Notons que la base de données AHTID / MW est disponible gratuitement à la disposition des chercheurs intéressés.

4.1.4. La base AHD / AMSH (Arabic Handwritten Database/ Amer-Shubair)

La base AHD / AMSH contient trois types d'images: des images de mots, de caractères isolés et de chiffres. Pour construire cet base, des formulaires contenant chacun 150 mots, 350 montant et 20 chiffres ont été préparés. Après vérification et approbation des formulaires par un linguiste, ils ont été distribués aux des scripteurs. En fait, 82 scripteurs répartis en 5 groupes d'âge ont participé à ce travail. Le choix des mots se fait avec prudence et précision afin de garantir que toutes les formes des caractères arabes sont couvertes [ALN 07].

Comme résultat, la base de données AHD / AMSH comprend 12300 mots manuscrits Arabes, 29028 sous-mots, 56170 caractères, 2870 montants de courtoisie, 820 chiffres indiens et 820 chiffres arabe. Les auteurs assument que cette base est utile pour les expérimentations sur la détection de ligne de base, segmentation de caractères, normalisation, amincissement, et pour toute les tâches d'apprentissage et test.

4.1.5. La base LMCA (Lettres, Mots et Chiffres Arabes)

La base LMCA est une double base de données, elle a été développée au sein du Groupe de recherche sur les machines intelligentes (REGIM), et elle est composée de 100 000 lettres, 500 mots et 30 000 chiffres. Les caractères et les mots manuscrits en ligne / hors ligne sont pris en compte. 55 participants ont été invités à contribuer au développement de la base de données. Les mots et les caractères sont stockés dans un fichier au format JPEG [ELM 08].

4.1.6. Base du CENPARMI(Center for Pattern Recognition and Machine Intelligence)

Cette base de chèques arabes a été construite en collaboration avec la Banque Al Radjhi (une des banques les plus populaires en Arabie Saoudite) dans le but de supporter les recherches sur la reconnaissance des chèques arabes manuscrites.

7000 images de chèques réels en niveau de gris ont été collectées. Après le processus de numérisation, les informations personnelles, y compris les noms, numéros de compte et signatures sont enlevées des images.

L'ensemble de données ainsi obtenu est subdivisé en plusieurs sous-bases. La première sous-base contient 2499 montants légaux ; la seconde comprend 2499 montants de courtoisie écrits en chiffres indiens. La troisième sous-base est composée de 29498 sous-mots dans le domaine du montant légal, et la dernière sous-base contient 15175 chiffres indiens. Chacune de ces sous-bases est divisée en ensemble d'apprentissage (66% à 75%) et ensemble de test, et stockée au format TIFF [ALO 03].

4.1.7. La base AHDB (Arabic handwriting database)

Cette base contient des mots et textes arabes écrits par une centaine de personnes différentes.

D'abord, les auteurs sélectionnent les 20 mots les plus répétés à partir de plusieurs textes extraits d'Internet. Ensuite, le formulaire a été conçu en six pages. Les trois premières pages étaient remplies de 96 mots, dont 67 mots de nombre manuscrits qui peuvent être utilisés dans la rédaction de chèques manuscrits. Les 29 autres mots étaient extraits des mots les plus populaires de l'écriture arabe. La quatrième page est conçue à être remplie avec trois phrases de mots, nombres et quantités manuscrits qui peuvent être écrits sur les chèques. La cinquième page est lignée, à remplir à main par le scripteur sur un sujet de son choix. 105 formulaires ont été numérisés, et chaque image de mot enregistrée au format TIFF [ALM 02].

4.1.8. La base ARABASE

Contient une large quantité de données issue de documents imprimés et manuscrite. La collection de hors ligne des données d'écriture manuscrite ont été exploitées en particulier dans des environnements académiques éloignés. Plus de 400 scripteurs ont participé à la collecte des données. La plupart des scripteurs étaient Tunisiens. Les auteurs ont utilisé des formulaires A4 avec des cases prédéfinies. Le contenu de ces formulaires a été inspiré de ceux utilisés dans le cas de l'ensemble de données UNIPEN. Le scripteur a été invité à remplir les cases vides, en copiant le texte juste au-dessus (chiffre, ville, noms, montant).Un espace a été laissé pour la signature. La vérité terrain du texte à écrire est fourni en imprimé. Chaque formulaire rempli est ensuite vérifié par un opérateur humain pour s'assurer qu'il a été correctement rempli. Les formulaires sont ensuite numérisés en niveaux de gris et en binaire à l'aide de deux scanners. ARABASE est subdivisée en plusieurs sous bases représentant

chacune une entité spécifique des données (noms de villes extraits de la base de données IFN / ENIT, montants littéraux, chiffres et caractères isolés) [AMA 05].

4.1.9. La base AHTD (Arabic handwriting text recognition database)

C'est une base de données de reconnaissance du texte en écriture arabe se composant de 3000 pages représentant 6000 paragraphes écrits par 300 scripteurs d'Arabie saoudite. Chaque formulaire de cet ensemble de données comprend quatre pages. La première page comprend des champs pour les informations de l'auteur: le nom, la tranche d'âge, la résidence, la qualification, sexe, droitier/ gaucher, et une section à des fins de gestion. Les trois pages restantes contiennent six paragraphes, le premier paragraphe est un paragraphe résumé qui couvre tous les caractères et formes arabes (début, milieu, fin et isolé). Le deuxième paragraphe contient du texte sélectionné au hasard dans un corpus arabe. La troisième page comporte deux paragraphes; le premier est sélectionné au hasard où le second est une répétition du premier paragraphe de la deuxième page. Sur la quatrième page, l'écrivain a été invité à écrire deux paragraphes sur n'importe quel sujet qu'il aime. Une fois les formulaires remplis par des scripteurs et collectés, ils sont numérisés à résolution 200 dpi, 300 dpi et 600 dpi en niveaux de gris. La vérité terrain des paragraphes est également stockée [MAH 11].

4.1.10. La base Alamri

La base de données Alamri est composée de : chiffres indiens isolés, des chaînes numériques, des lettres arabes isolées et une collection de 70 mots arabes, cette base de données comprend un échantillon de format libre d'une date arabe. Le formulaire de l'ensemble de données est composé de deux pages. La première page comprend : un échantillon d'une date arabe, 20 chiffres isolés, 38 chaînes numériques de différentes longueurs. La deuxième page contient le reste des mots candidats. Le premier formulaire a été rempli par 100 scripteurs arabes sélectionnés au hasard de sexe, d'âge, de niveau d'instruction et de nationalité différents au Canada. Le deuxième formulaire a été rempli en Arabie saoudite par 228 participants. La base de données est divisée en trois séries : la série 1 contenait les échantillons des 100 premiers scripteurs ; la série 2 est composée des échantillons des 228 derniers scripteurs, et la série 3 comprenait les combinaisons de tous les échantillons. Au total, il y a 656 pages de formulaires remplis. Les formulaires étaient numérisés en couleurs. Des prétraitements sont ensuite appliqués sur les différentes séries d'images. Tout d'abord, un filtre spécial a été appliqué pour supprimer toutes les bordures rouges des cases contenant les images cibles. Puis les images en vraies couleurs ont été converties en niveaux de gris [ALA 08].

4.1.11. La base Colombie-Britannique

La base de données de la Colombie-Britannique développée dans le département de l'informatique et du génie électrique de l'Université de la Colombie-Britannique. Cette base est composée de phrases, de mots, de chiffres et de signatures. 500 étudiants sélectionnés au hasard de l'Université Al-Isra ont participé à la collecte de données. Chaque étudiant a été invité à signer cinq fois et à copier une liste présélectionnée de mots et de chiffres, ainsi qu'une phrase. Les mots de base ont été scrupuleusement choisis de manière à assurer la présence de toutes les lettres arabes sous diverses formes (début, milieu, fin et isolé). Cette base contient 37000 mots arabes, 10000 chiffres indiens et arabes, 2500 signatures et500 phrases arabes de forme libre; tous enregistrés en images en niveaux de gris et en binaire au format BMP [AMA 05].

4.1.12. La base DBAHCL (database for Arabic handwritten characters and ligatures)

Cette base de données a été construite pour remédier de l'absence des ligatures (qui sont fréquentes dans l'écriture manuscrite arabe) dans les différentes bases de données existantes à cet époque. Cependant, DBHCL est implémentée de façon à contenir des lettres arabes et des ligatures sans points diacritiques, écrites par 50 déférents scripteurs. Afin de couvrir toutes les formes possibles, chaque scripteur écrit deux fois un caractère sous toutes ses formes (isolé, au début du mot, au milieu et attaché à la fin du mot). Chaque l'auteur remplit quatre formulaires : un formulaire de caractères simples et trois formulaires de ligature. Le formulaire de caractères simples contient 55 caractères, alors que les trois autres formulaires contiennent 99 caractères en chevauchement (ligatures). Ainsi, l'ensemble était 110 caractères simples et198 ligatures, ce qui donne 308 caractères écrits par chaque scripteur, et 15400 caractères au total.

La base de données est subdivisée en deux dossiers : «DB CHARACTERS» et «DB LIGATURES». Le premier dossier «DB CHARACTERS» contient les caractères arabes simples stockés dans des dossiers dont le nom indique le nom et la position du caractère. Le deuxième dossier regroupe les formes possibles des ligatures arabes [LAM 17].

4.2. Bases de données pour d'autres langues

4.2.1. Base du CEDAR (Center of Excellence for Document Analysis and Recognition)

C'est une base payante qui contient presque 28000 images de noms de villes et d'états, de codes postaux, de chiffres et de caractères isolés, en niveaux de gris ou binaires. Elle est divisée en plusieurs parties décrites dans le tableau suivant :

Nom	Nombre d'images	Contenu
BB	300	traiter un problème particulier (Caractères qui se touchent, mauvais formats de codes postaux...)
BC	211	images d'un même code postal
BD	2636	Adresses
BL	973	images d'un même nom de ville avec les codes postaux
BS	500	tests pour la base BD
BR	3962	Codes postaux
BU	1072	codes postaux avec chiffres connectés

Tableau 1.2. *Parties de la base CEDAR [CHE 08].*

Le site du CEDAR propose aussi une autre base des caractères japonais [HUL 94].

4.2.2. Bases du SRTP (Service de Recherches Techniques de la Poste)

Le SRTP utilise ses propres bases de données pour ses tests, mais celles-ci ne sont pas disponibles. De nombreux travaux utilisent cependant ces bases d'images de chèques ou d'adresses pour l'évaluation de leurs algorithmes [CHE 08].

4.2.3. Base MNIST

La base de données MNIST est composée d'images de chiffres manuscrits extraits de la base NIST et déjà prétraités et normalisés (60.000 exemples d'apprentissage, et 10.000 exemples de test) [W11].Cette base gratuite comporte 10 classes de chiffres (de 0 à 9) homogènes. Les caractères se présentent sous forme d'images en niveaux de gris de 28×28 pixels.

4.2.4. Base du CENPARMI

La base de données du CENPARMI (*Center for Pattern Recognition and Machine Intelligence*) contient des images de chèques canadiens en anglais et en français. Elle contient 17000 chiffres isolés extraites à partir de 3400 codes postaux [COT 97].

4.2.5. Bases ETL

Les bases de données de caractères ETL ont été construites au laboratoire d'électrotechnique (ETL) de l'AIST (*Advanced Industrial Science and Technology*) en collaboration avec des universités et industriels. Les bases de données ETL1 – ETL9 contiennent environ 1,2 million de caractères manuscrits et imprimés répartis en caractères japonais, chinois, latins et numériques. Elles sont gratuites pour des applications de recherche [W5]. Les contenus des différentes bases sont synthétisés dans le tableau 1.3.

Nom	Types de caractères	Nombres d'échantillons	Manuscrit/ Imprimé
ETL1	Numériques Latins Spéciaux Katakana	141 319	M
ETL2	Kanji Hiragana Katakana Alphanumériques Spéciaux	52 796	I
ETL3	Numériques Latins Spéciaux	9600	M
ETL4	Hiragana	6120	M
ETL5	Katakana	10608	M
ETL6	Katakana Numériques Latins Spéciaux	157 662	M
ETL7L	Hiragana	16800	M (grandes caractères)
ETL7S	Hiragana	16800	Petits caractères
ETL8G	Kangi Hiragana	152 960	M
ETL8B2	Kangi Hiragana	152 960	M (binaires)
ETL9G	Kangi Hiragana	607 200	M
ETL9B	Kangi Hiragana	607 200	M (binaires)

Tableau 1.3. *Caractéristiques des bases ETL [CHE 08].*

4.2.6. La base UW de l'université de Washington

Ce corpus est composé de publications scientifiques numérisées. Les régions d'intérêt et leurs labels fonctionnels sont fournis. Le défaut majeur de cette base est que la segmentation s'arrête au niveau du mot, et en plus elle n'est pas encore indexée avec des attributs typographiques suffisamment précis [PHI 93].

4.2.7. Base UNIPEN

C'est une base de données conçue pour la reconnaissance de l'écriture manuscrite enligne [GUY 94]. Elle est composée des caractères, chiffres, mots et des phrases (plus que 5 millions caractères) écrites dans plusieurs styles et de plus de 2200 scripteurs. Malheureusement cette base de données n'est pas disponible au public.

4.2.8. Farsi-City

Cette base a été présentée par Dehghani en 2001. Elle est composée de 17000 images de 198 noms de villes iraniennes [ELA 08].

4.2.9. La base GrCor (Grecque Coror)

La base possède 47 classes pour un total de 18236 caractères grecs manuscrits écrits par quatre scripteurs dans le cadre de COROC (Base grecque de caractères a été réalisée au sein de RC-Soft pour les tests internes). Elle est relativement propre et constitue la base la moins complexe [ARR 06].

4.2.10. La base GrAnc

La base possède 38 classes pour un total de 1445 caractères qui sont issus de quatre pages de manuscrits grecs anciens différents et d'époques différentes. L'acquisition a été effectuée avec un scanner grand public à 300 dpi. Les images ont été binarisées puis ont subi une fermeture morphologique pour supprimer les pixels isolés. La segmentation a été réalisée par une analyse en composantes connexes avec calcul de boîtes englobantes [ARR 06].

4.2.11. La base de hiéroglyphes manuscrits : HrMan

La base possède 94 caractères. sélectionnés pour leur bonne segmentation et non pour leur contenu et plusieurs caractères peuvent posséder le même contenu sémantique. L'acquisition a été effectuée avec un scanner grand public à 300 dpi. Les images ont été binarisées puis ont subi une fermeture morphologique pour supprimer les pixels isolés. La segmentation a été réalisée par une analyse en composantes connexes avec calcul de boîtes englobantes. Les boîtes ont été triées à la main pour réaliser la base d'apprentissage [ARR 06].

4.3. Base de documents historiques

Pour les documents historiques, il n'existe pas des bases de données à proprement parler mais des collections des documents anciens numérisées, indexés manuellement et disponibles enligne à partir des sites des bibliothèques. Parmi ces collections, nous citons :

4.3.1. Collection de George Washington

Les papiers de l'officier de l'armée et du premier président américain George Washington (1732-1799) constituent la plus grande collection de documents originaux de Washington au monde. Elle se compose d'environ 77000 articles accumulés par Washington entre 1745 et 1799.La collection a été produite par un seul auteur (George Washington), et elle est conservé à la *division des manuscrits de la Bibliothèque du Congrès* [W3].

4.3.2. Collection de la Bibliothèques Virtuelles Humanistes

Les Bibliothèques Virtuelles Humanistes (BVH), créées en 2002 au Centre d'Études Supérieures de la Renaissance, et élaborées avec la collaboration de l'IRHT (CNRS, section de l'humanisme), diffusent des documents patrimoniaux et poursuit des recherches associant des compétences en sciences humaines et en informatique en 2019 la bibliothèque représentant un total de 26 485 vues (éléments de reliures, pages blanches et vues de contrôle – mires, chartes colorimétriques, échelles – inclus) [W1].

4.3.3. Collection de la Bibliothèque Nationale de Tunis

Contient aussi un grand nombre de livres et documents arabes et islamique rares et elle a commencé à numériser des fonds historiques important pour les rendre accessible au grand public à distance. La collection numérique peut être consultée via l'adresse web de la bibliothèque [W2].

4.3.4. Gallica, la bibliothèque numérique de la BNF et de ses partenaires

Gallica est la bibliothèque numérique de la Bibliothèque Nationale de France (BNF) et de ses Partenaires. En libre accès depuis 1997, elle est l'une des plus importantes bibliothèques numériques accessibles gratuitement sur l'internet. Elle offre l'accès à tous types de documents : imprimés (livres, presse et revues) en mode image et texte, manuscrits, documents sonores, documents iconographiques, cartes et plans, vidéos [W7].

Elle a 6573228 documents au 1er janvier 2020, dont 702538 livres, 3591983 fascicules de presse et revues, 1410638 images, 134087 manuscrits, 173039 cartes, 50291 partitions, 51150 enregistrements sonores, 457839 objets et 1663 vidéos [W8].

4.3.5. La Base de Sulaiman et al.

D'après la recherche bibliographique que nous avons fait, nous avons trouvé un seul travail proche de notre présent travail. Ce travail est celui de Sulaiman et al. [SUL 17]. Dans ce travail les auteurs ont proposé une base pour les anciens manuscrits arabes historiques. La

base construite couvre, d'après les auteurs, différents types de dégradations (voir le chapitre 2). Elle contient trois images pour chaque type de dégradation accompagnée chacune de l'image binaire de vérité terrain correspondante. Aucune autre information de vérité terrain n'est disponible pour ces images. Cependant, l'extraction des images de vérité terrain a été obtenue grâce à l'utilisation manuelle d'outils graphiques. Les auteurs ont utilisé les deux programmes de traitements d'image GIMP et Photoshop, comme ces programmes sont efficaces pour différents traitements d'images. Les auteurs procèdent ainsi en plusieurs étapes. Dans la première étape, l'image a été nettoyée par l'application de différents filtres comme étape de prétraitement. La deuxième étape consiste à appliquer un seuillage pour extraire l'image binaire. La troisième étape consistait à améliorer la qualité de l'image de sortie en utilisant certains filtres et outils pour obtenir le meilleur résultat. En dernière étape, l'image résultat de vérité terrain est examinée et vérifiée par un expert de traitement d'image pour vérification. La figure suivante illustre un exemple:

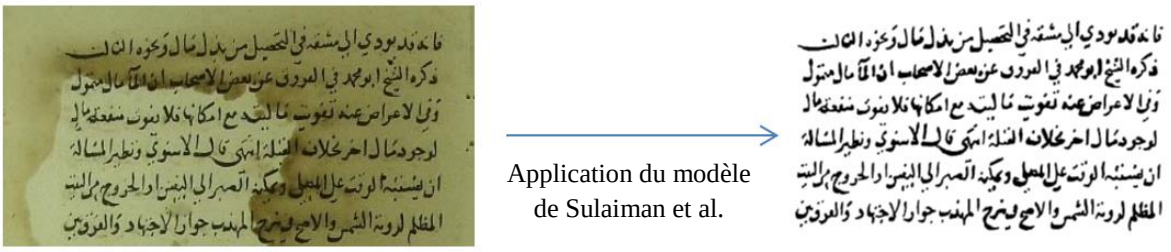


Image d'entrée Image de vérité terrain

Figure 1.1. *Modèle de Sulaimane et al. [SUL 17].*

5. Conclusion

Dans ce chapitre nous avons exposé un aperçu sur les bases de données pour l'analyse et la reconnaissance de documents. En effet, la construction d'une base de données est une priorité pour les chercheurs dans le domaine de l'analyse et la reconnaissance de documents. Ceci est évident car les bases de données présentent un outil indispensable dans le processus de développement de systèmes de d'analyse et de reconnaissance. Surtout dans les phases d'apprentissage et de test où on a besoin de travailler avec des bases de données standards.

Chapitre 2. Modélisation de bruit et Segmentation d'images de document

1. Introduction

Dans ce chapitre nous allons passer en revue deux traitements en relation avec les images de documents, et plus particulièrement les images de documents dégradés, à savoir l'étude et la modélisation de bruits, et la segmentation. Ces deux traitements seront utilisés ensemble pour la constructions de notre base de données (voir le chapitre 3).

Le présent chapitre est subdivisé en deux section, la première section traite la notion de bruit et dégradation ainsi que les différents travaux de modélisation de bruit. La deuxième section adresse un aperçu sur la segmentation d'images de documents, ses différentes approches, etc.

2. Modélisation de bruit

A la différence des bases d'images de documents réels qui se composent des images résultantes de la numérisation d'un grand nombre de documents papiers présentant du bruit et dégradations réels, et accompagnés des informations d'annotation ou de vérité terrain, les bases de documents synthétiques sont construites dans le sens inverse. Ainsi, une telle base commence des informations et d'images de vérité terrain auxquels sont ajoutés différents types de dégradations et bruits. Comme nous nous intéressons dans ce présent travail à la construction d'une base d'images de documents historiques synthétiques, nous jugeons nécessaire de présenter même brièvement une étude sur le bruit et la modélisation de bruit (dégradations des documents anciens plus précisément), ce champ de recherche qui n'est pas suscité suffisamment de travaux de recherches dans la littérature.

2.1. Notion de bruit et dégradation

2.1.1. Bruit numérique

On appelle bruit numérique toute fluctuation ou dégradation que subit l'image de l'instant de son acquisition jusqu'à son enregistrement. Le bruit numérique est une notion générale à tout type d'image numérique, et ce quel que soit le type du capteur à l'origine de son acquisition (appareil photo numérique, scanner, caméra thermique... etc.) [RAF 02].

2.1.2. Dégradation

Selon [DRI 07], une dégradation est tous les effets indésirables cumulés nuisant à la lisibilité, au traitement ou à la conservation des images. Les dégradations ont plusieurs origines et l'accumulation des défauts rend difficile leurs séparations. Les dégradations ont des effets plus destructeurs sur les images de documents que sur toutes autres images dites "naturelles".

2.2. Différents types de dégradation

De bonnes études sur les dégradations et leur catégorisation ont été menées dans [DRI 07] et [RAB 13]. La catégorisation la usuelle des dégradations les sépare en deux classes : *dégradations relatives aux documents originaux* et *dégradations résultantes d'une mauvaise numérisation*. Dans cette section nous présentons les altérations couramment rencontrées dans les images de documents numérisés.

2.2.1. Dégradations des documents originaux

Cette catégorie de dégradations regroupe tous les défauts qui proviennent essentiellement de l'état du document original. Parmi les dégradations les plus usuelles nous trouvons : les défauts touchant la forme des caractères, l'effet de transparence, les taches, les pliures et ondulations, les déchirures et trous : qui peuvent se produisent à cause d'une mauvaise utilisation, de l'usure, des insectes, etc., les annotations et les ajouts indésirables [RAB 13].

2.2.2. Dégradations dues à la numérisation

Les défauts causés par la numérisation peuvent avoir plusieurs formes, citons: l'effet du flou, le bruit marginal, les défauts géométriques: tels que l'inclinaison du document, la mauvaise colorimétrie d'un document, les défauts chromatiques, l'illumination non uniforme, l'apparition des éléments inutiles : comme les doigts tenant le document, le marque page, etc., le bruit des images numériques: causé par le capteur du scanner, le bruit de fouillis (clutter noise) [RAB 13].

2.3. Travaux sur la modélisation de bruit

Selon [KIE 14], une partie significative des travaux relatifs à la génération d'images de documents synthétiques s'est intéressée à la modélisation des dégradations les plus couramment observés. Ces différents travaux classent les différents défauts observés sur les images de documents selon leurs spécificités visuelles en trois classes : bruit local (e.g. caractères), bruit global (e.g. illumination), ou bruit diffus(e.g. transparence). Dans cette section nous présentons quelques travaux sur la modélisation de bruit qui nous semblent intéressants.

2.3.1. Bruit local

a) Bruit local de Kanungo

Les auteurs de [KAN 93] présentent un modèle de bruit local qui permet de dégrader le contour de caractères dans les images binaires. Le bruit local de Kanungo est basé sur la

transformée de distance des données de vérité terrain et certains post-traitements morphologiques. Dans ce bruit, les auteurs modélisent la probabilité de changement d'un pixel de sa valeur idéale comme une fonction de la distance de ce pixel à la limite d'un caractère.

Soit d la distance d'un pixel de premier plan ou d'arrière-plan à partir de la limite du caractère. α et β sont des paramètres d'échelle. Soit $P(1/d, \beta, f)$ et $P(0/d, \beta, f)$ la probabilité d'un pixel de premier plan à une distance d, de rester 1 et de passer à 0, respectivement. De même manière, soit $P(1/d, \alpha, f)$ et $P(0/d, \alpha, f))$ la probabilité d'un pixel d'arrière-plan à une distance d changeant en 1 et restant 0, respectivement. Les fonctions $P(1/d, \alpha, f)$ et $P(1/d, \alpha, b)$ pourrait être différents. Le processus de perturbation aléatoire procède ensuite au changement des valeurs de pixel de manière indépendante pixel par pixel.

Cela est suivi par une opération de fermeture morphologique afin de tenir compte de la corrélation introduite par la fonction d'étalement de point optique précédant l'opération de seuillage, ce qui produit l'image bruitée. L'opération de fermeture a été réalisée avec une un élément structuré binaire 2 × 2.

Les formes suivantes pour les probabilités conditionnelles de l'arrière-plan et du premier plan ont été utilisées:

$$P(1/d, \alpha, b) = 1 - P(0/d, \alpha, b) = e^{-\alpha d^2}$$

$$P(0/d, \beta, f) = 1 - P(0/d, \beta, f) = e^{-\beta d^2}$$

La figure 2.1 présente un exemple d'un caractère affecté par le bruit local de Kanungo avec déférent pourcentage.

Figure 2.1. *Bruit local Kanungo [MAR 14].*

b) Bruit "hard pencil"

Le modèle "hard pencil noise" proposé par Zhai et al. [ZHA 03] est également dédié aux images binaires, et il permet de reproduire les lignes blanches apparaissant près des formes.

Ce bruit répond au problème de texture du papier, lorsque le crayon dessine la surface du papier, certains endroits restent intacts, laissant de minuscules espaces blancs sur les lignes noires (voir la figure 2.2).

L'algorithme se déroule comme suit: Pour chaque point *g(i,j)* qui est sur la ligne de balayage et n'est pas un point dans une couleur de fond (blanc), on génère une variable aléatoire *R* où:

$$R = \frac{\text{rand ( )} \times L\_MAX}{\text{RAND \_MAX}} \times 13 \quad \text{.....(1)}$$

En utilisant un seuil *L*, allant de 0 à *L_MAX* , si *R* < *L*, un nombre aléatoire *w* compris entre 0 et (*L* + 5) / 3 est généré. Une ligne blanche de longueur *w* pixels est dessinée dans la même direction que la ligne de balayage. Si cette ligne blanche sort de la zone noir, sa queue est coupée, effectivement ignorée [ZHA 03].

Figure 2.2. *Bruit hard pencil [ZHA 03].*

c) Bruit local de flou de mouvement

Le bruit de flou de mouvement se produit lorsque le document est perturbé et déplacé pendant sa numérisation ou son accrochage [ZHA 03]. La direction du mouvement est choisie au hasard. Chaque pixel du document original est divisé par *L* pour obtenir un document *M*. *M* est ensuite déplacé le long de la direction sélectionnée de *i* pixels, où *i* est un entier allant de –*L* / 2 à *L* / 2. Chaque fois, *Mi* est ajouté à *M*. Toute valeur de pixel supérieure à 255 est définie sur 255. Le résultat est le document flou de mouvement *M'*. *L* représente le niveau de bruit.

Figure 2.3. *Bruit local de flou de mouvement [ZHA 03].*

2.3.2. Bruit global

a) Bruit global de Baird

L'auteur de [BAI 90] présente un modèle de dégradation pouvant simuler des bruits globaux apparaissant lors du processus de numérisation d'une image (*Figure 2.4*). Ce modèle se compose de dix paramètres qui permettent de modéliser quatre types de défauts : la rotation,

la mise à l'échelle, la translation et l'erreur d'échantillonnage des couleurs induite par l'acquisition utilisant certains scanners, en particulier les plus anciens.

***Figure 2.4**. Bruit global de Baird [MAR 14].*

b) Bruit global de Kanungo

Le modèle global de Kanungo et al. [KAN 93] permet de simuler la distorsion du papier et l'effet de l'illumination sur la page apparaissant lors de la numérisation d'un ouvrage avec une reliure épaisse [MAR 14]. Kanungo et al. ont modélisé quatre sources de dégradation.

b.1) Modèle de déformation pour processus de pliage de la page physique

La page subit une déformation physique où la page du document passe par un processus de pli. La page n'est plus une surface plane mais une surface incurvée. Les auteurs modélisent cette portion courbée de la page sous forme d'un segment d'arc circulaire le long de l'axe x et supposent qu'il n'y a pas une telle déformation le long de l'axe y [KAN 93]. Voir la figure 2.5.

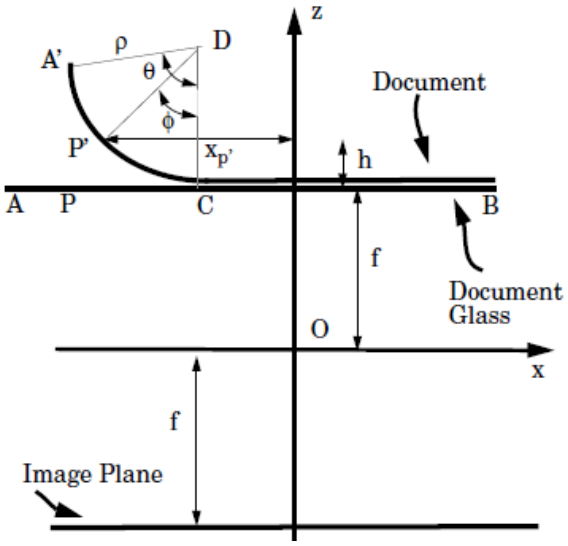


***Figure 2.5**. Déformation de pliage de pages de documents [KAN 93].*

Soit $A = (x_a, y_a, f)'$, $B = (x_b , y_b, f)'$. De plus, soit ρ le rayon du cercle de déformation et soit le segment plié sous-tend un angle θ au centre du cercle D. Soit le point A mappé sur le point $A' = (x_{a'}, y_{a'}, z_{a'})'$ après déformation. Puis les coordonnées de A' sont données par :

$$\left.\begin{aligned} xa' &= xa + \rho(\theta - sin\,\theta) \\ ya' &= ya \\ za' &= f + \rho(1 - \cos\theta) \end{aligned}\right\} \quad \text{.....(2)}$$

Soit le point $P = (x_{p'}, y_{p'}, f)'$ tel que $x_a \leq x_p \leq x_b$, et soit P mappé au point $P' = (x_{p'}, y_{p'}, z_{p'})'$ après déformation. Laissez l'angle sous-tendu par l'arc $P'C$ au centre D être ϕ tel que:

$$\emptyset = \frac{xa + \rho\theta - xp}{\rho} = \theta - \left(\frac{xp - xa}{\rho}\right) \quad \text{.....(3)}$$

Maintenant, les coordonnées de P' peuvent être calculées comme dessous:

$$\begin{aligned} x_p' &= x_p + \rho(\phi - \sin\phi) \\ y_p' &= y_p \\ z_p' &= f + \rho(1 - \cos\phi) \end{aligned} \quad \text{.....(4)}$$

Remarque : si $x_p > x_b$; nous n'avons pas de déformation et donc $P' = P$.

b.2) Modèle de distorsion en perspective

La déformation de pliage est suivie d'une distorsion en perspective où le point P' sur le document correspond au point P'' de l'image (voir la figure 2.6). Soit f la distance focale du système optique et soit le centre de perspective O être à l'origine. Supposons que le plan de l'image est au plan focal à $-f$. Soit $P'' = (x_{p''}, y'', z_{p''})'$ être la projection en perspective du point P'' sur la page du document.

Les cordonnées de P'' sont données par les équations suivantes [HAR 92, HOR 86]

$$\begin{aligned} x_p'' &= -f\left(\frac{x_p'}{f + \rho(1 - \cos\phi)}\right) \\ &= -f\left(\frac{x_p + \rho(\phi - \sin\phi)}{f + \rho(1 - cos\ \phi)}\right) \\ y_p'' &= -f\left(\frac{y_p'}{f + \rho(1 - \cos\phi)}\right) \\ &= -f\left(\frac{y_p}{f + \rho(1 - \cos\phi)}\right) \\ z_p'' &= -f \end{aligned} \quad \text{.....(5)}$$

Remarque : les points P du document original avec0 $x_p > x_b$; nous n'avons aucune déformation et donc $-P'' = P' = P$.

b.3) Modèle d'éclairage non linéaire

Puisque la page du document n'est plus plane mais une surface courbe, l'éclairage sur le document n'est pas constant. L'éclairement en un point P' sur le page du document est inversement proportionnel à la distance du point P' à partir de la source lumineuse L.

La source lumineuse L se déplace sous la vitre d'exposition d'une extrémité à l'autre. Laissez la distance entre la vitre d'exposition et la source lumineuse L être l_0. voir figure 2.6.

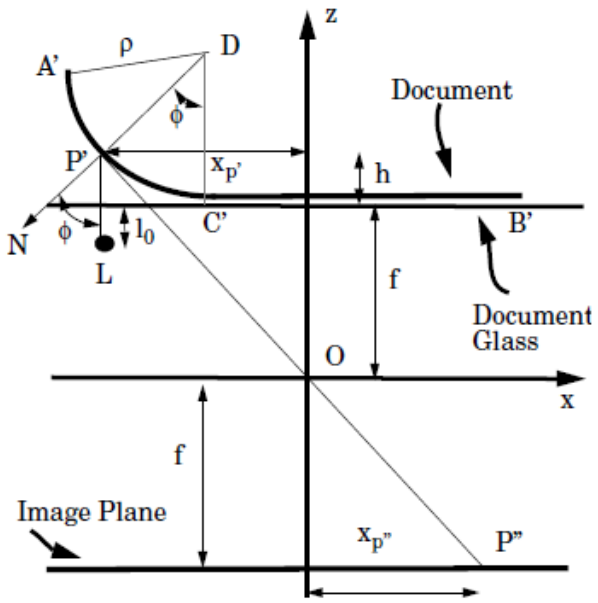


Figure 2.6. *Distorsion en perspective [KAN 93].*

Aux endroits où la page est incurvée, le distance entre la source lumineuse et le document pages est $l = l_0 + \rho(1 - \cos\phi)$ où ϕ est l'angle l'arc $P'B$ sous-tend en B. Notez ϕ l'angle entre la normale à P' et la direction négative z, l'intensité de la lumière est proportionnelle à la cosinus de l'angle ϕ. De plus, le modèle d'utilisation suppose que l'intensité de la lumière est la même dans toutes les directions. Soit I_0 l'intensité à un point où le document n'est pas courbé, c'est-à-dire la distance entre la lumière et le point considéré est l_0 . Ainsi $l_0 \, \alpha \, \frac{1}{l0^2}$

Ensuite, l'intensité en I_p' un point sur la partie courbe est proportionnel à $\cos\phi$ et inversement proportionnel à $(l_0 + \rho(1 - \cos\phi)^2)$ donc:

$$I_p' \propto \frac{\cos\phi}{(l_0 + \rho(1 - \cos\phi))^2} \quad \text{.....(6)}$$

Prenant ainsi un rapport des deux équations ci-dessus, nous avons

$$\boldsymbol{I_p' = I_0\left(\frac{l_0}{l_0 + \rho(1 - \cos\phi)}\right)^2} \quad \text{.....(7)}$$

sous l'hypothèse d'un éclairage diffus, nous avons $l_p' = l_p''$.

b.4) Fonction d'étalement de point optique non linéaire

En fait, l'image d'un point est géométriquement un disque si le plan d'image n'est pas net (figure 2.7). Si Δ est le diamètre de la lentille, et h est la distance du plan d'image à partir du plan focal, puis le diamètre du disque est donné par:

$$\boldsymbol{d = \Delta\left(\frac{h}{f}\right)} \quad \text{.....(8)}$$

Mais en raison d'irrégularités optiques, en réalité on obtient un disque comme image mais une version floue d'un disque.

En fait, ce disque flou peut être modélisé comme un gaussien avec un écart type $\sigma = k * d$; où k est une caméra constant.

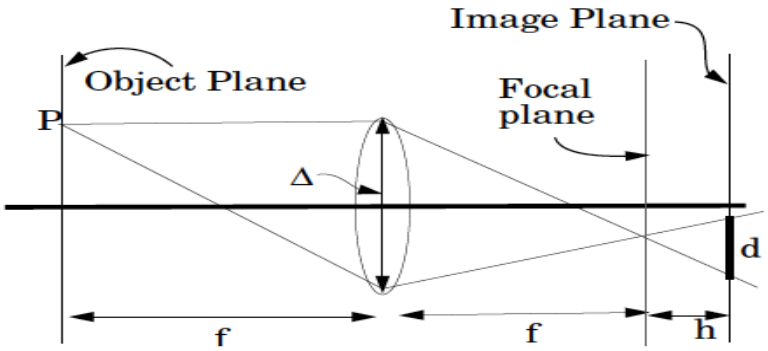


Figure 2.7. *Fonction d'étalement de point optique non linéaire [KAN 93].*

Après avoir effectué la transformation de pliage, la distorsion en perspective et l'éclairage non-linéaire, une autre étape est nécessaire où l'image est convolée avec un noyau gaussien variant dans l'espace.

Le noyau a un écart type σ donné par la formule suivante dans les régions courbes et par une constante σ_0 ailleurs:

$$\sigma = k \,.\, \rho(1 - \cos \emptyset) \quad \text{.....(9)}$$

Le résultat final de bruitage par le modèle global de Kanungo est illustré par la figure 2.8.

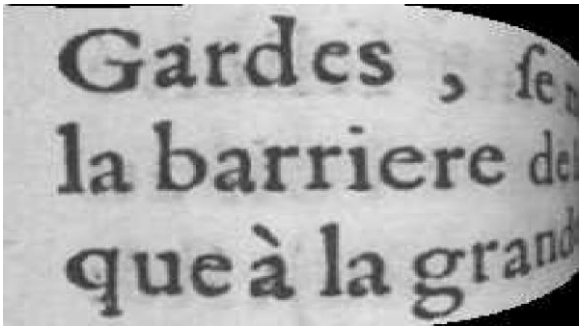


Figure 2.8. *Bruit global de Kanungo [MAR 14].*

c) Bruit global de Liang

Liang et al. [LIA 08] ont également modélisé ce défaut (***Figure 2.9***). Dans ce modèle, la forme de la page est considérée comme une surface courbe en 3D pouvant se déplier sur un plan 2D [MAR 14]. Ainsi, ils ont présenté un cadre de rectification qui extrait la forme du document 3D d'une seule image 2D et effectue une rectification géométrique basée sur la forme pour restaurer le front vue à plat du document [LIA 08].

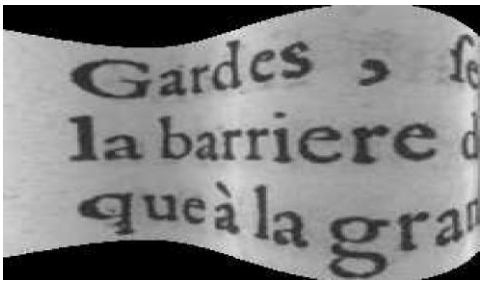


Figure 2.9. *Bruit global de Liang et al. [MAR 14].*

d) Bruit global de Kieu

Ce modèle de bruit propose également une déformation d'une page de document [KIE 13]. Il s'adapte particulièrement bien au contexte des documents anciens, puisqu'il reproduit les distorsions locales telles que les plis, les trous, les abrasions présentes dans les vieux documents [MAR 14]. Voir la Figure 2.10.

Figure 2.10. *Bruit global de Kieu et al. [MAR 14].*

2.3.3. Bruit diffuse

a) Bruit transparence

Ce modèle proposé par Moghaddam [MOG 09] permet de simuler l'apparition de l'encre du recto sur le verso (figure 2.11). Ce modèle se base sur un processus de diffusion dont l'idée principale est d'exécuter itérativement des opérateurs de diffusion. Un opérateur représente un processus de diffusion d'encre de la source (verso) à la cible (recto) en niveau de gris [MAR 14].

Figure 2.11. *Bruit transparence de Moghaddam [MAR 14].*

b) Bruit diffuse de Curtis

Ce modèle de bruit est un modèle de diffusion d'encre simulant la diffusion d'un liquide tombant sur une surface plane [CUR 97]. Voir la Figure 2.12. Il est basé sur la re-création,

synthétiquement, des traits artistiques les plus saillants de aquarelle d'une manière à la fois prévisible et contrôlable [CUR 97].

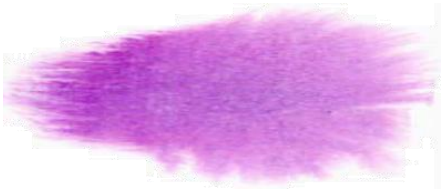


Figure 2.12. *Bruit diffuse de Curtis [CUR 97].*

3. Segmentation d'images de documents

La segmentation est le cœur de tout système de vision et étape primordiale en traitement d'image. Elle s'inspire du système de perception visuel humain pour générer une interprétation d'une image. La segmentation cherche à extraire, de façon aussi exacte que possible, des régions homogènes et disjointes les unes des autres dont chacune caractérise un des objets présents dans cette image.

3.1. Notion de document

3.1.1. Définition

Le document est un ensemble formé par un support et une information. Celle-ci est enregistrée de manière persistante [W4]. Il a une valeur explicative, descriptive ou de preuve. C'est un vecteur matériel de la pensée humaine, il joue un rôle essentiel dans la plupart des sociétés contemporaines. On peut distinguer quelque exemples de documents: visuelle (vidéo, image, texte, document ancien, …), document audio(musique, les cours vocal ….), document textuel (manuscrit, imprimé). Ainsi, le document pose toujours le problème de sa véracité. On s'intéresse beaucoup plus dans notre travail aux documents manuscrits.

3.1.2. Catégories de documents manuscrits

De manière générale, on peut classer les documents manuscrits dans quatre catégories [OUW 10] en fonction de leur structure, et ceci quel que soit la langue. Ces classes sont :

- **Documents mono-orientés :** les lignes dans cette classe sont orientées suivant (en adoptant) une seule direction. Voir la figure 2.13.a.
- **Documents multi-orientés**: les lignes dans ces documents occupent ou non toute la largeur du document où elles sont rangées par blocs. Elles ont des orientations différentes. Voir la figure 2.13.b.

- **Documents multi-scripts :** il s'agit ici de documents écrits par plusieurs personnes différentes, conduisant à des écritures ou scripts différents. Cela arrivait souvent dans le passé, les personnes se succédaient pour compléter un même document, ou collaboraient à l'écriture d'un même ouvrage. De nos jours, certains documents modernes peuvent aussi être multi-scripts, comme par exemple les cartes de vœux, les écrits destinés à différentes nationalités,… etc. Voir la figure 2.13.c.
- **Documents hétérogènes**: ce type de document contient à la fois du texte et des images ou des illustrations. Beaucoup de documents cartographiques, mécaniques ou architecturaux sont de ce type. Les documents manuscrits de ce type peuvent contenir plusieurs orientations correspondant à des lignes de cotation, ou d'illustration des dessins. Voir la figure 2.13.d.

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

***Figure 2.13**. Les catégories des documents manuscrits [OUW 10].*

3.2. Segmentation

En traitement et analyse de documents, la segmentation d'image est une opération de traitement d'images qui a pour but de rassembler des pixels entre eux suivant des critères prédéfinis [W10]. Les pixels sont ainsi regroupés en régions, qui constituent un pavage ou une partition de l'image. Il peut s'agir par exemple de séparer les objets du fond ou des photos du texte ou des mots.

3.3. Approches de segmentation

Un document numérisé est à la base, une image de pixels. Les approche existantes dans la littérature sont réparties en trois: descendantes, ascendantes, et mixtes [OUW 10].

3.3.1. Approche descendante

Les approches descendantes considèrent le document dans sa globalité et une seule orientation est en principe recherchée. Dans cette classe, le document est décomposé en allant du plus large (blocs) au plus fin (pixel ou composantes connexes).

a) L'analyse de profils de projections

Dans cette méthode, pour extraire les lignes, on suit en général les maxima du profil de projection du document entier ou d'une bande de ce document en respectant l'orientation trouvée.

Le profil de projection est obtenu en additionnant les valeurs des pixels sur l'axe de l'orientation du document binaire. Ensuite, les maxima et les minima du profil sont déterminés et les lignes entre deux minima consécutifs sont recherchées. Ces composantes connexes forment une ligne si la taille du pic engendré correspond à la largeur de la ligne.

il y a plusieurs travaux utilisant cette technique comme ceux de Nicolaou et Gatos, Shapiro et al., Antonacopoulos et Karatzas [OUW 10]. Pour la segmentation de documents arabes manuscrits nous citons la méthode proposée par Bennasri et al. [BEN 99]. Cette méthode permet d'extraire les lignes d'un texte manuscrit arabe, en utilisant la projection. D'abord le document est divisé en plusieurs colonnes. Puis, les points de départ de toutes les lignes sont détectés en utilisant les minima du profil de projection partielle puis, un suivi de contour partiel de chaque ligne est effectué, d'abord dans le sens de l'écriture, puis dans le sens opposé. Puis, les lignes adjacentes qui se chevauchent sont séparées. Les points diacritiques, sur lesquels demeure un doute sont marqués et leur affectation à l'une ou l'autre des lignes est effectuée.

b) L'algorithme de découpage XY

L'algorithme de découpage XY introduit par Nagy et al. [NAG 00], 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 satisfaite. 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. Nagy et al. donnent quelques résultats d'extraction de lignes sans toutefois mentionner le taux d'extraction. L'algorithme de

découpage XY a été utilisé pour des documents qui ne contiennent pas beaucoup des variations [NAG 00].

3.3.2. Approche ascendante

En effet, les approches descendantes ne s'adapte pas avec les variations locales de l'image [TOR 03]. Pour éviter cette difficulté et afin d'améliorer la qualité de la segmentation, il convient d'introduire l'information contextuelle autour de chaque point de l'image, alors que les approches ascendantes sont basées sur les éléments de bas niveau de l'image comme les pixels ou les composantes connexes. Dans cette catégorie, nous trouvons la classification par k-plus proches voisins, la transformée de Hough, la technique de lissage RLSA (Run Length Smoothing Algorithm), la technique de réseau répulsif-attractif, le regroupement par l'arbre recouvrant minimal, et le regroupement par la déformation de contour localisé et propagation.

a) L'algorithme RLS

La technique de lissage consiste à noircir les petits espaces entre les pixels noirs consécutifs sur la direction horizontale ou verticale, c.-à-d. les connecter. Les boîtes qui englobent les composantes connexes successives horizontalement dans l'image lissée forment les lignes.

Les auteurs Shi Zhi xin et Venu Govindaraju ont proposé une méthode en utilisant l'algorithme de "fuzzy run length". Ce lissage crée des alignements qui vont être analysés selon plusieurs heuristiques pour former les lignes de texte. L'alignement dont la taille n'excède pas une valeur prédéfinie est supprimé [CHI 04].

3.3.3. Approche mixte

Les approches ascendant et descendante s'appuient sur des informations différentes et complémentaires [DRI 07], pour ça les chercheur utilisent des approches hybrides avec une combinaison séquentielle en série des méthodes d'analyse ascendantes/ descendantes, généralement une approche descendante est suivie d'une approche ascendante pour affiner les résultats de la première analyse.

En exemple d'approche mixte : la relaxation probabiliste [BON 01] est une méthode utilisant le résultat d'une analyse descendante et qui cherche ensuite à affiner le résultat par une approche ascendante. L'estimation des probabilités initiales basées sur la distribution des couleurs de l'image s'apparente à une segmentation globale. Ces probabilités initiales sont ensuite affinées progressivement en analysant les distributions locales dans un voisinage de chaque point de l'image.

4. Conclusion

Dans ce chapitre nous avons fait une étude théorique sur deux notions en relation au traitement et à l'analyse d'images de documents. Cette étude est répartie en deux parties.

La première partie passe en revue l'étude du bruit et défauts présents dans les images de documents anciens, ainsi que leur modélisation. La partie commence par la présentation des notions de bruit et dégradation. Nous avons vu que les dégradations sont de plusieurs types, peuvent provenir de causes multiples, et elles affectent la qualité visuelle des images et peuvent engendrer une perte significative de l'information. Le reste de cette partie est consacré à la synthèse de certains travaux sur la modélisation de bruit, les plus connus de la littérature.

Dans la deuxième partie de ce chapitre nous avons présenté les notions de documents et segmentation et les déférents catégories de documents manuscrits. Nous avons également exposé les différentes approches de segmentation en citons quelques exemple de chaque approche.

Chapitre 3. Contribution

1. Introduction

Nous avons vu dans le premier chapitre une vue globale sur le développement et la conception de bases de données pour l'analyse et la reconnaissance de documents. Dans le deuxième chapitre, nous avons vu quelques types bien connus de bruits dans les images de document, ainsi que la segmentation dans le domaine de documents.

Dans le présent chapitre, nous présentons notre contribution dans la conception d'un outil de généralisation d'images de documents bruités tout en détaillant les différentes étapes impliquées.

2. Objectif du travail

La construction des bases de données pour fournir des grandes collections d'apprentissage et de test a une importance énorme au chercheur dans le domaine du traitement d'images et de l'écriture. Cela les permet de tester et d'évaluer leurs méthodes sur un ensemble standard d'images.

Notre objectif principal dans ce mémoire de master est la construction d'une base d'images de documents arabes dégradés synthétique. Dans le but d'avoir des images bruitées à partir des images de documents propres (images de vérité terrain), on propose de créer un outil permettant de générer différents types de bruits, qui se trouvent dans les images de documents historiques, sur les images de texte (vérité terrain).

La base de documents créée permettra aux chercheurs dans le domaine de traitement et d'analyse d'images de documents dégradés de développer leurs méthodes de traitement, de les tester sur une collection standard d'images et de comparer les résultats avec les informations de vérité terrain.

3. Description de l'approche proposée

Comme nous avons avancé dans le chapitre 1, la difficulté principale lors de la construction d'une base de documents est l'établissement des informations de vérité terrain (images binaires, lignes de bases, étiquettes des différents éléments, etc.) à partir des images collectées. Ces informations de vérité terrain sont établies dans la plupart de temps manuellement par des experts et après des analyses approfondies, ce qui est couteux et très difficile, voire impossible dans quelque cas comme le cas des documents anciens.

L'idée de base de notre proposition est de partir d'un ensemble de textes arabes et d'images de vérité terrain (images en noir et blanc nettoyées épurés de bruit) et de produire des images altérées par différents types de bruits. Nous détaillons dans la suite la démarche suivie.

3.1. Préparation de textes

Nous sommes intéressé à construire une base de documents de différentes classes de bruit, comportant plusieurs types d'écriture imprimée, en couleur, en niveaux de gris et en noir et blanc, de différents styles, différentes polices. Comme première étape nous avons préparé un ensemble de fichiers texte écrits en Microsoft Word. Chaque fichier contient une ou plusieurs pages. Les textes sont écrits de façon à couvrir une grande variété de styles, polices, et de tailles d'écriture afin d'enrichir la diversité de la base qui va être construite.

En effet, il est intéressant de partir des fichiers texte et puis de les transformer en images au lieu de partir des images directement. Ces textes font une partie importante des informations de vérité terrain de notre base. La figure 3.1 illustre un exemple d'un tel fichier texte.

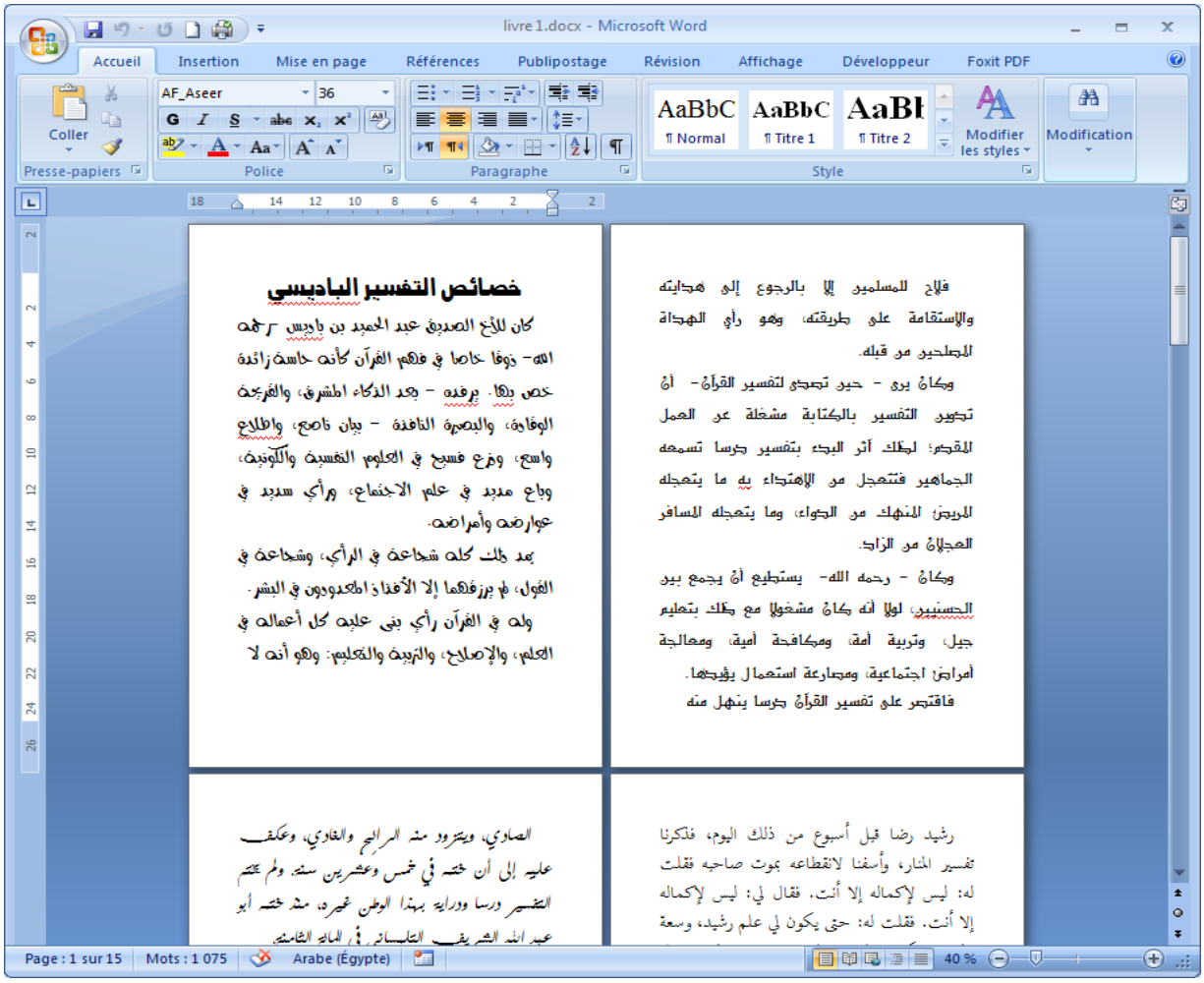


Figure 3.1. *Exemple d'un fichier texte de départ.*

3.2. Images de vérité terrain

La deuxième étape est le capture d'mages à partir des fichiers texte préparés sous format Microsoft Word, tel que chaque page du document texte est capturée en une image. Pour ce faire nous avons passé par un format intermédiaire: le format PDF. Ainsi, tous les fichiers Word sont enregistrés en PDF. La conversion des fichiers PDF en images a été effectuée à l'aide d'un convertisseur en ligne[1] et le format d'images choisi est le format PNG (figure 3.2)

Figure 3.2. *Exemples d'images de vérité terrain*

3.3. Génération d'images bruitées

Dans notre outil, nous proposons de générer quatre types d'images de documents dégradés. Les bruits et dégradation affectant les images de départ sont des dégradations très rencontrées dans les documents anciens, et ont été modélisées par les chercheurs ce qui permet leur intégration dans notre outil. Nous détaillons dans cette partie du chapitre les étapes suivies.

3.3.1. Obtention d'images de documents bruités par combinaison texte/fond

Comme nous avons dit précédemment, on va générer automatiquement des images de documents historiques synthétiques, en ajoutant aux images propres de départ (considérées comme images de vérité terrain) du bruit aléatoire relatif aux documents dégradés, dans le but de construire une base synthétiques de documents historiques pour lesquelles des informations de vérité terrain existent. Une image de document historique artificiel (synthétique) est obtenue en superposant (en combinant) une image de départ avec un masque. Les masques sont des images de fonds anciens vierges (ne contiennent aucun texte ou graphiques), collectées à partir du web et couvrent la plupart des problèmes qu'on peut les trouver dans les

[1]https://pdf2png.com/fr/

documents anciens (présence des tâches, illumination inégale, plis et déchirure, etc.). La figure 3.3 illustre quelques exemples de fonds anciens de notre collection.

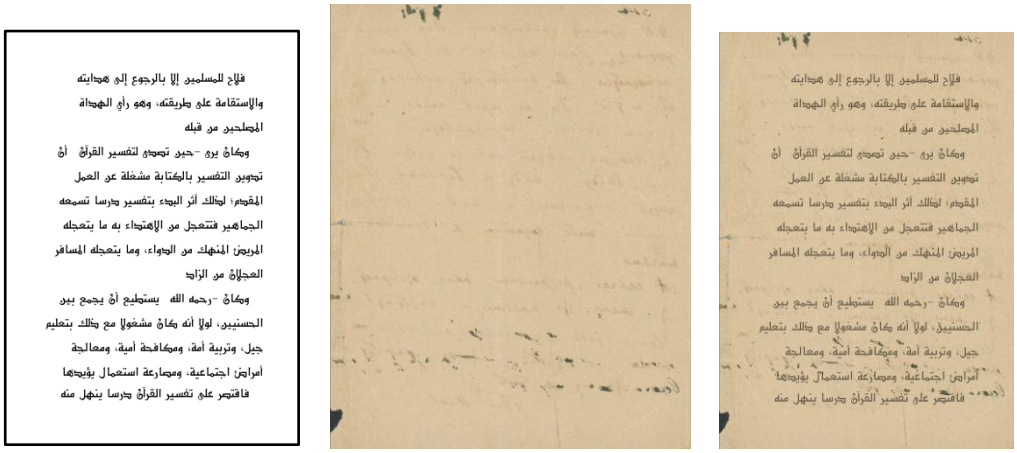
Figure 3.3. *Quelques images de fonds anciens [W6].*

La combinaison est effectuée en utilisant la technique de Mosaicing d'images proposée par Stathis et al. [STA 08]. Ainsi dans [STA 08], deux techniques de combinaison ont été proposées : l'intensité maximale, et l'intensité moyenne. Nous avons adopté la technique d'intensité maximale qui consiste à privilégier les pixels les plus foncés des deux images, avec une modification lorsque le pixel dans l'image de texte est plus foncé. De telle façon, pour les pixels de l'avant plan, l'image du texte aura dans la plupart des cas l'avantage sur l'image du masque et pour les pixels de l'arrière-plan, l'image du masque aura l'avantage car elle est généralement plus foncée. Cette technique peut être résumée par le pseudo code suivant :

Entrées : *GT*: l'image de texte
BG: l'image de fond ancien
Sortie: *R*: l'image résultat
Pour chaque pixel *(i, j)* de *GT* faire
Si *BG(i,j)* est plus foncé que *GT(i,j)* alors *R(i,j)* ← *BG (i,j)*
Sinon *R(i,j)* ← *(GT(i,j) + BG(i,j))/2*;
Fin Pour;
Fin.

En effet, la fusion en utilisant cet algorithme est possible lorsque les deux images à fusionner ont la même taille ou l'image du masque est plus grande. Dans ce dernier cas, la fusion ne prend pas l'image du fond dans sa totalité mais une partie de l'image ayant la même taille que l'image du texte. Lorsque l'image du texte soit la plus grande, la fusion en utilisant l'algorithme précédent ne peut être effectuée correctement. La dernière partie de l'image du

texte reste sans couverture par le masque. Pour surmonter ce problème, nous avons proposé d'étirer l'image du masque en la redimensionner automatiquement à la taille de l'image du texte. La figure suivante illustre la fusion et la redimensionne de l'image de fond.



(a) ***(b)*** ***(c)***

Figure 3.4. *Fusion par Mosaicing, (a) image du texte, (b) image du masque, (c) image résultante*

3.3.2. Génération d'images bruitées par le bruit local de Kanungo

Le bruit local de Kanungo ajoute du bruit poivre et sel sur les contours des caractères dans une image binaires, donc il s'applique dans notre proposition sur les images de texte de départ. La méthode de Kanungo consiste en suite à appliquer une opération morphologique de fermeture pour lisser le contour [KIE 14].

On définit un contour interne l'ensemble de pixels noirs ayant au moins un voisins blanc. De même, un contour externe est l'ensemble de pixels blancs ayant au moins un voisin noir. La figure 3.5 présente un exemple. Le pixel *p1* est un pixel de contour interne et le pixels *p2* est un pixel de contour externe.

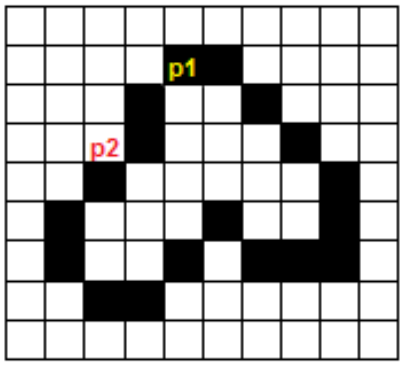


Figure 3.5. *Exemple de pixels de contour*

Le bruitage par le bruit local de Kanungo peut être résumé par le pseudo code suivant. *random()* est une fonction retournant un nombre aléatoirement.

Entrées : *GT*: l'image de texte d'hauteur *Haut* et de largeur *Larg*
P : pourcentage de points bruités.
Sortie : *R*: image bruitée
Variables utilisées:
VPB[]: Vecteur de points bruités de taille *TV* = (*Haut* * *Larg*) * *P*
Nbp: Nombre de pixels déjà bruités;
Nbp ← 0;
Tantque *nbp* < *TV* Faire
i = random() * *Haut*; // générer un numéro de ligne aléatoirement;
j = random() * *Larg*; // générer un numéro de colonne aléatoirement;
VPB[Npb] ←*(i, j); //* Ajouter l'élément (*i, j*) dans le vecteur *VPB ;*
Nbp ← *Nbp* + 1 ;
Fin Tantque;
Pour chaque pixel *(i, j)* de *VPB* faire
Si *(i, j)* est un pixel de contour interne alors *R(i,j)* ← blanc;
Sinon si *(i, j)* est un pixel de contour externe alors *R(i,j)* ← noir;
Sinon *R(i,j)* ← *GT(i,j)*;
Fin si;
Fin Pour;
Fin.

3.3.3. Génération d'images présentant un effet de transparence

L'effet de transparence est communément présent dans les images de documents historiques et c'est pour cette raison qu'on a choisi de l'intégrer dans notre outil. En effet, pour la génération d'un effet de transparence synthétique nous faisons recours à une deuxième image de texte représentant le verso qui apparait à travers le recto. Cette image sera combinée avec l'image initiale de recto. Ainsi, l'image de verso subit d'abord certaines prétraitements avant la combinaison, notamment la symétrie horizontale, comme le texte du verso apparait à l'envers (le sens de l'écriture est inversé) sur l'image du recto, et la diminution de la densité de l'image de verso. La combinaison de l'image de recto avec celle de verso (après sa prétraitement) donne l'impression d'un effet de transparence.

Cette étape peut être résumée par le pseudo-code suivant:

Entrées : *GT*: l'image de texte
V: l'image de texte verso d'hauteur *Haut* et de largeur *Larg*
P: valeur de transparence du verso (entre 0 et 255)
Sortie: *R*: l'image résultat
Variables utilisées:
rgb: valeur de pixel ;
//Symétrie horizontale de *V*
Pour *i* ←*0 à Haut* faire
Pour *j*←*0 à Larg/2* faire
rgb ← *V(i, j) ;*
V(i, j) ← *V(i, Larg* - 1 - *j) ;*
V(i, Larg - 1 - *j)* ← *rgb ;*
Fin Pour ;
Fin Pour ;
//Combinaison recto/ verso en tenant compte de la valeur de transparence
Pour *i* ← *0 à Haut* faire
Pour *j* ← *0 à Larg* faire
Si *V*(*i, j*) < *GT(i,j)* alors *R(i ,j)* ← *P;*
Sinon *R(i ,j)* ← *GT*(*i, j*)*;*
Fin Pour ;
Fin Pour ;
Fin.

3.3.4. Génération d'images présentant un effet de rotation

Un autre défaut fréquent dans le cas des images de documents anciens est l'inclinaison. Cette dernière peut se produire lors de l'écriture du document comme elle peut résulter d'une mauvaise numérisation. Dans les deux cas, le document numérique semble être incliné à un certain angle. Nous avons choisi d'ajouter l'effet de rotation à notre outil. Le principe de la rotation est que, les nouvelles coordonnées de chaque pixel notées (x', y') sont calculées à partir des anciennes coordonnées (x, y) et de l'angle de rotation θ comme suit [BAN 09]:

$$\begin{pmatrix} x' \\ y' \end{pmatrix} = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}$$

3.3.5. Génération d'images présentant un effet de courbure

La courbure est également un bruit fréquemment rencontré surtout dans les images issues de la numérisation des pages d'un livre. Pour générer un effet de courbure affectant une image de

départ nous utilisons le modèle de bruit global de Kanungo, présenté en détails dans le chapitre 2. Ce modèle permet de simuler la distorsion du papier et l'effet de l'illumination sur la page apparaissant lors de la numérisation d'un ouvrage avec une reliure épaisse.

Pour ce faire, on applique premièrement une transformation de pliage pour modéliser une portion de la page de document sous forme d'un segment d'arc circulaire le long de l'axe *x*. Puis on fait la distorsion en perspective, suivie par l'éclairage non linéaire. Les détails de calcul sont présentés en détails dans le chapitre précédent. Notons que ce modèle utilise plusieurs paramètres qui doivent être fixés. L'ajout d'un effet de courbure en utilisant le modèle de Kanungo peut être résumé par le pseudo-code suivant:

Entrées : *I*: l'image de texte ou de combinaison d'hauteur *Haut* et de largeur *Larg*
p: point limite de document
θ: l'angle au segment plié
f: distance
lo : la distance entre la vitre d'exposition et la source lumineuse
neg : booléean (vrais /faux)
Sortie: *R*: l'image résultat
Variables utilisées:
xa, xb, ya,yb, x, y, x2, y2,i2, j2, c: nombres entiers
Inisialiser *R* à une couleur unique (noire)
Si *larg* mod 2 = 0 alors *xa* ← - *larg/2* et *xb* ← *larg/2*- 1 *;*
Sinon *xa* ← - *larg/2* et *xb* ← *larg/2 ;*
Si *Haut* mod 2 = 0 alors *ya* ← - *Haut*/2 + 1 et *yb* ← *Haut*/2;
Sinon *ya* ← *Haut*/2 et *yb* ← *Haut*/2 ;
Pour chaque pixel *(i, j)* de *I* ayant des coordonnées entre (*xa* à *xb, ya* à *yb*) faire
(*x, y*) ← transformation de pliage de *(i. j)*
(*x2, y2*) ← distorsion en perspective de *(x. y)*
(i2, j2) ← nouvelles coordonnées calculées à partir de *x2*, et *y2*
c ← nouvelle couleur obtenue de l'é1clairage non linéaire utilisant *I0*, θ, *p,* et *I*(*i, j*);
R(i2, j2) ← *c*;
Fin Pour
Pour chaque pixel *(i, j)* de *I* non transformé
R(i,j) ← *R(i+1,j)*;
Fin Pour
Fin.

3.4. Segmentation du document et préparation des informations de vérité terrain

La segmentation est la deuxième phase principale de notre projet. Elle a pour objectif de séparer les différentes entités composant le document et qui constituent des informations importantes sur les documents (information de vérité terrain). Dans notre proposition on considère la segmentation en lignes, en mots, en sous-mots, et les composantes connexes. Ces différentes entités séparées constituent des informations importantes (information de vérité terrain) sur le document et le texte contenu dans ce document.

3.4.1. Segmentation en lignes

Dans cette étape on segmente l'image en lignes. Pour faire ça nous appliquons la célèbre méthode d'analyse des profils de projection. Cette méthode procède comme suit. Nous parcourons l'image et nous calculons le nombre des pixels noirs dans chaque ligne de l'image. Si ce nombre est supérieur à zéro on enregistre le début de la ligne et on continue le parcours jusqu'à un nombre égal à zéro et on enregistre la fin de ligne. La zone de ligne de texte est entre les deux et le nombre max des pixels de la zone et la ligne de base.

Entrées : *GT*: l'image de texte d'hauteur *Haut* et de largeur *Larg*
Sortie: *VZ []*: Vecteur contient les zones de lignes de textes ;
Variables utilisées:
VP[]: Vecteur exprimant l'histogramme de projections horizontales de taille *TV = Haut*
VL []: Vecteur contient les débuts et les fin des lignes de texte;
Nb: Nombre
//Calcul de l'histogramme de projections horizontales
Pour *i* ←*0 à Haut* faire
Pour *j* ←*0 à Larg* faire
Si GT(i,j) = noire alors *VP [i]* ← *VP [i]* +1 *;*
Fin Pour ;
Fin Pour ;
//Extraire les zones des lignes de texte
Nb ← 0;
Pour *i* ←*0 à Haut* faire
Si VP[i] > 0 alors *VL[*Nb*]* ← *i;*
Nb ← *Nb* + 1 *;*
Fin si
Fin Pour ;
Nb← 0;

Pour *i* ←*0 à Haut* faire
VZ[Nb] ← (*VL [*i*], VL [*i+1*]);*
i ← i + 2;
Calculer la rectangle englobante de la ligne de texte *VZ [Nb] ;*
Nb ← *Nb+1 ;*
Fin Pour ;
Fin.

3.4.2. Segmentation en mots

Dans cette étape on segmente les lignes en mots et la même manière procédée lors de la segmentation en lignes mais cette fois-ci le parcours se fait verticalement. Nous parcourons ainsi chaque ligne et nous calculons son histogramme de projections verticales. A partir de cet histogramme on extrait les valeurs nuls qui correspondent au début et à la fin d'un mot probable. On enregistre alors le début et la fin de chaque mot probable. Ensuite un filtrage est appliqué afin de ne laisser que les mots réels. Ce filtrage est effectué à l'aide d'un seuil prédéfini. Cependant, si la distance entre deux mots voisins est inférieur au seuil, ces deux mots sont regroupés ensemble car ils forment un seul mot. Le seuil est établi basé sur la variance et la distance moyen entre les mots et fixé après des expérimentations égale au minimum entre la largeur minimale des mot * 3 et la largeur de l'image /50.

Entrées : *GT*: l'image de texte d'hauteur *Haut* et de largeur *Larg*
VL []: Vecteur des lignes
Seuil: Seuil de regroupement
Sortie: *VZM []*: Vecteur contient les rectangles englobantes des mots ;
Variables utilisées:
Vpv []: Vecteur de projection vertical ;
VM[]: Vecteur contient les débuts et les fins des mots
Pour chaque ligne de *VL* faire
Pour *i* ← *xmin* à *xmax* faire //*xmin, xmax, ymin,y max sont les limite de la ligne ;*
Pour *j* ← *ymin* à *ymax* faire
Si GT(i,j) = noire alors *Vpv[i]* ← *Vpv [i]* + 1 *;*
Fin Pour ;
Fin Pour ;
Nb ← *0;*
Pour *i* ← *0 à Haut* faire
Si *Vpv [i]* > 0 alors *VM [Nb]* ← *i*;

Nb ← *Nb* + 1 *;*
Fin si
Fin Pour ;
Pour *i* ← *0* au *nombre d'éléments de VM* - 1 faire
Si *VM [i+1] - VM[i] < seuil* alors *Regrouper VM [i+1] et VM[i] dans un seul mot*
Fin Pour ;
Nb ← *0;*
Pour *i* ← *0* au *nombre d'éléments de VM* - 1 faire
VZM[*Nb*] ← (*VM [i], VM [*i+1*]);*
i ← i + 2;
Nb ← *Nb+1 ;*
Fin Pour;
Fin Pour ;

3.4.3. Segmentation en sous-mots

Dans cette étape on segmente les mots en sous mots. La méthode est la même que la précédente. Cependant on calcule l'histogramme de projections verticales pour chaque mot. Les entrées nuls dans l'histogramme correspondent aux zones de segmentation.

Entrées : *GT*: l'image de texte d'hauteur *Haut* et de largeur *Larg*
VM []: Vecteur de mots
Sortie: *VZSM []*: Vecteur contient les rectangles englobantes des sous-mots ;
Variables utilisées:
Vpv []: Vecteur de projection vertical ;
VSM[]: Vecteur contient les débuts et les fins des mots
Pour chaque mot de *VM* faire
Pour *i* ← *xmin* à *xmax* faire //*xmin, xmax, ymin,y max sont les limite du mot*
Pour *j* ← *ymin* à *ymax* faire
Si GT(i,j) = noire alors *Vpv[i]* ← *Vpv [i]* + 1 *;*
Fin Pour ;
Fin Pour ;
Nb ← *0;*
Pour *i* ← *0 à Haut* faire
Si *Vpv [i]* > 0 alors *VSM [Nb]* ← *i*;
Nb ← *Nb* + 1 *;*
Fin si
Fin Pour ;

```
    Nb ← 0;
    Pour i ← 0 au nombre d'éléments de VSM - 1 faire
        VZSM[Nb] ← (VSM [i], VSM [i+1]);
        i ← i + 2;
        Nb ← Nb+1 ;
    Fin Pour;
Fin Pour ;
```

3.4.4. Segmentation en composantes connexes

Dans cette étape on segmente les sous-mots en composantes connexes. Par définition, une composante connexe est un ensemble de pixels interconnectés. Pour le regroupement des pixels en composantes connexes on utilise la méthode d'agrégation de pixel en version récursive. Ainsi, pour chaque sous-mot, on commence d'un pixel noir et on essaye de regrouper tous ses voisins ensemble. On voit alors s'il a un voisin noir alors on l'ajoute avec lui dans la même composante connexe et on réitère la même chose pour ce voisin, puis le voisin du voisin, et ainsi de suite. Le regroupement s'arrête lorsqu'aucun voisin noir n'est resté et on passe à un autre pixel noir non visité dans le sous-mot s'il y on a.

```
Entrées : GT: l'image de texte d'hauteur Haut et de largeur Larg
          VSM []: Vecteur de sous-mots
Sortie:   VCC []: Vecteur contient les composantes connexes ;
Variables utilisées:
     CC[]: vecteur contient les pixels d'une composante connexe
Nb ← 0;
Pour chaque sous-mot de VSM faire
    Pour i ← xmin à xmax faire  //xmin, xmax, ymin,y max sont les limite du sous-mot
        Pour j ← ymin à ymax faire
            Si GT(i, j) = noire alors
                CC ← regrouper les pixels noirs voisins de (i, j) ;
                VCC[Nb] ← CC;
                Nb ← Nb + 1 ;
        Fin Pour ;
    Fin Pour ;
Fin Pour ;
Fin.
```

3.5. Extraction des informations de vérité terrain

Comme avancé dans les chapitres précédents, ce sont les informations de vérité terrain qui donnent valeur aux bases de données publiques, et de ce fait leur présence dans la base de données est nécessaire et critique. Ces informations annotent les documents et elles représentent les caractéristiques essentielles des documents (nombre de lignes, nombre de mots, positions de lettres, etc.). Les informations de vérité terrain peuvent être sous formes d'images, fichiers XML, ou autres formats.

Cependant, dans notre proposition les informations de vérité terrain sont en trois niveaux: niveau de caractères, niveau d'image binaire, et niveau des autres informations d'annotation.

3.5.1. Textes de vérité terrain

Ce sont les mêmes que les textes de départ. Ainsi, à chaque image bruité généré automatiquement est associé un fichier texte, celui du texte d'origine à partir duquel l'image a été créée. Les fichiers textes (et plus précisément les caractères de ces textes) présentent les résultats attendus de la reconnaissance des images.

Cette caractéristique rend notre base de données adéquate pour l'évaluation des algorithmes de reconnaissance de l'écriture, que ce soit au niveau de caractères ou de mots.

Dans notre outil, nous proposons de générer quatre types d'images de documents dégradés. Les bruits et dégradation affectant les images de départ sont des dégradations très rencontrées dans les documents anciens, et ont été modélisées par les chercheurs ce qui permet leur intégration dans notre outil. Nous détaillons dans cette partie du chapitre les étapes suivies.

3.5.2. Images binaires de vérité terrain

Bien que les textes de départ sont écrits tous en noir sur un fond blanc, leur conversion en image produit des pixels en nuance de gris. Même si le nombre de ces pixels est négligeable par rapport au nombre total de pixels de l'image, la transformation en une image binaire comportant deux valeurs uniquement: noir pour le texte et blanc pour le fond est nécessaire. Ces images binaires font partie des informations de vérité terrain et elles expriment les résultats attendus de la binarisation. A cause de ces informations (images binaires de vérité terrain), notre base de données est adaptée à l'apprentissage et l'évaluation des méthodes de binarisation.

Nous avons utilisé une méthode de binarisation à seuillage globale à savoir la méthode d'Otsu qui essaye de trouver le seuil T qui sépare l'histogramme de façon optimale en deux segments

(qui maximise la variance inter-segments ou bien qui minimise une mesure de variance intra-segments). Le calcul de la variance inter-classes ou intra-classes est basée sur l'histogramme normalisé *H2*= [*h0...h255*] de l'image d'où Σ*hi*=1.

La variance inter-classes pour chaque seuil *t* est donnée par :

$$\boldsymbol{\sigma}^2_{inter} = q1(t) \times q2(t) \times [\mu 1(t) - \mu 2(t)]^2.$$

Tel que:

$$q_1(S) = \sum_{i=0}^{S-1} H2(i) \; ; q_2(S) = \sum_{i=S}^{255} H2(i) \; ;$$

$$\mu_1(S) = \frac{1}{q_1(S)} \sum_{i=0}^{S-1} H2(i) \times i \; ; \quad \mu_2(S) = \frac{1}{q_2(S)} \sum_{i=S}^{255} H2(i) \times i \; ;$$

3.5.3. Autres informations d'annotation - Génération du fichier XML

La segmentation a permis d'extraire les différentes entités existantes. Elle nous indique du nombre de lignes dans le texte, la position exacte de chaque ligne dans l'image et le nombre de mots dans la ligne, la position du mot et le nombre de sous-mots qui le composent, la position du sous-mot et le nombre de ses composantes connexes, et enfin la position de chaque composante connexe. Toutes ces informations sont enregistrées finalement dans un fichier d'annotation correspondant à l'image de document traitée. Ces informations d'annotation permettent de rendre notre base de données appropriée pour l'apprentissage et l'évaluation des algorithmes de segmentation sur plusieurs niveaux.

Dans notre projet, on a choisi d'utiliser le format XML pour enregistrer les informations de vérité terrain. Le langage XML (eXtended Markup Language) c'est un langage orienté texte et formé de balises qui permettent d'organiser les données de manière structurée [W9]. Les informations de vérité terrain représentent les caractéristiques essentielles du document: nombre de lignes et leur position dans l'image, les mots, les sous mots, les composant connexes, et la transcription textuelle de chaque mot. Toutes les informations sauf la transcription textuelle sont obtenues à partir des étapes de segmentation précédentes. Pour la transcription textuelle, il est nécessaire de charger le fichier texte initial correspondant à l'image traitée, et de l'analyser. Les lignes de texte sont extraites et puis séparées en mots. Chacun de ces mots en ASCII exprime la transcription textuelle d'un mot détecté dans l'image. Le mot en ASCII est alors ajouté comme attribut à chaque mot détecté de l'image.

Chaque fichier d'annotation XML contient les informations suivantes sur l'image de document :

- Nom, hauteur et largeur de l'image, la langue avec laquelle le texte est écrit

- Nombre de lignes dans le document
- Position de chaque ligne dans l'image (coordonnées de leur rectangle englobante)
- Nombre et position des mots dans chaque ligne
- La transcription en ASCII correspondante à chaque mot
- Nombre et position de sous-mots dans chaque mot
- Composantes connexes principales (corps), et secondaires (diacritiques) dans chaque sous-mot

Un fichier d'annotation XML prend la forme suivante:

```
<?xml version="1.0" encoding="UTF-8" ?>
- <DOCUMENT imageName="livre 1_001.png" height="3508" width="2481" nbTextLines="12" language="Arabic">
  - <TEXTLINE id="0" nbWords="3" boundingBox="337,476,513,2000">
    - <WORD id="2" nbPAWs="2" boundingBox="337,1528,496,2000" ocr="خصائص">
      - <PAW id="1" nbDiacritics="1" boundingBox="337,1749,458,2000">
          <CC id="0" type="Primary" nbPixels="14179" boundingBox="337,1749,458,2000" />
          <CC id="1" type="Secondary" nbPixels="235" boundingBox="354,1943,369,1960" />
        </PAW>
      + <PAW id="0" nbDiacritics="1" boundingBox="337,1528,496,1741">
      </WORD>
    + <WORD id="1" nbPAWs="2" boundingBox="337,1058,490,1487" ocr="التفسير">
    + <WORD id="0" nbPAWs="4" boundingBox="337,476,513,1016" ocr="اليابيسي">
    </TEXTLINE>
  - <TEXTLINE id="1" nbWords="8" boundingBox="583,346,725,2072">
    - <WORD id="7" nbPAWs="2" boundingBox="583,1934,704,2072" ocr="كان">
      + <PAW id="1" nbDiacritics="0" boundingBox="583,1994,687,2072">
      + <PAW id="0" nbDiacritics="1" boundingBox="643,1934,704,1987">
      </WORD>
    + <WORD id="6" nbPAWs="1" boundingBox="590,1745,725,1901" ocr="للأخ">
    + <WORD id="5" nbPAWs="3" boundingBox="599,1434,720,1710" ocr="الصديق">
    + <WORD id="4" nbPAWs="1" boundingBox="626,1256,709,1401" ocr="عبد">
    + <WORD id="3" nbPAWs="2" boundingBox="592,1005,720,1219" ocr="الحميد">
    + <WORD id="2" nbPAWs="1" boundingBox="643,881,709,968" ocr="بن">
    + <WORD id="1" nbPAWs="3" boundingBox="602,632,720,847" ocr="باديس">
    + <WORD id="0" nbPAWs="3" boundingBox="598,346,725,589" ocr="-رحمه">
    </TEXTLINE>
  + <TEXTLINE id="2" nbWords="9" boundingBox="787,324,935,2185">
  + <TEXTLINE id="3" nbWords="8" boundingBox="1006,323,1145,2191">
  + <TEXTLINE id="4" nbWords="7" boundingBox="1232,368,1359,2185">
  + <TEXTLINE id="5" nbWords="7" boundingBox="1407,444,1571,2188">
  + <TEXTLINE id="6" nbWords="8" boundingBox="1636,586,1782,2188">
  + <TEXTLINE id="7" nbWords="2" boundingBox="1848,1442,1992,2187">
  + <TEXTLINE id="8" nbWords="8" boundingBox="2060,373,2206,2074">
  + <TEXTLINE id="9" nbWords="8" boundingBox="2265,354,2417,2185">
  + <TEXTLINE id="10" nbWords="9" boundingBox="2481,446,2629,2070">
  + <TEXTLINE id="11" nbWords="7" boundingBox="2695,359,2844,2185">
  </DOCUMENT>
```

Figure 3.6. *Exemple d'un fichier d'annotation XML.*

4. Structure de la base de documents à créer

La base de documents construite regroupera dix répertoires comme suit :

- ***Dossier « Fichiers_TXT »:*** contient les fichiers textes de départ mais en format TXT.

- ***Dossier « Images_Bin »:*** contient l'ensemble d'images de texte en binaire (images de vérité terrain).
- ***Dossier « Fond_Im »:*** contient les images de fond utilisées pour la génération de documents historiques artificiels.
- ***Dossier « Images_Comb»:*** contient les images résultantes de la combinaison.
- ***Dossier « Bruit_Kanungo» :*** contient les images bruitées par un bruit local de Kanungo
- ***Dossier « Bruit_Transparence» :*** contient les images avec transparence verso
- ***Dossier « Bruit_Rotation » :*** contient les images bruitées par un bruit rotation
- ***Dossier « Bruit_Courbure » :*** contient les images bruitées par un bruit courbure
- ***Dossier « Tous_Bruits » :*** contient les images bruitées successivement par tous les bruits décrits précédemment.
- ***Dossier « Informations_VT »:*** contient les fichiers d'annotation XML correspondants aux images de vérité terrain.

5. Conclusion

Dans ce chapitre, nous avons présenté les différentes étapes de la création d'une base d'images de documents arabes bruités. Les détails d'implémentation, les interfaces de l'application, et les résultats obtenus seront l'objet du prochain chapitre.

Chapitre 4. Implémentation et résultats

1. Introduction

Nous allons présenter dans ce chapitre l'application que nous avons développée pour la construction d'une base de documents arabe dégradés synthétique, tout en exposant son fonctionnement par des illustrations.

Dans un premier temps nous donnons un aperçu sur l'environnement de développement, ensuite nous décrivons les étapes de fonctionnement de l'application et ses déférentes fenêtres. Dans la dernière section, nous présentons les détails de la base construite, avant de conclure.

2. Environnement de développement

Notre application a été développée en langage de programmation Java, avec l'environnement de développement NetBeans IDE 8.2 et la Version: Build201609300101.

JAVA est un langage de programmation orienté objet développé par Sun Micro system. Il n'a que quelque années de vie (la première version date de 1995). Le système java comporte plusieurs parties: un environnement, le langage, les interfaces de programmation d'application, et diverses bibliothèques de classes.

NetBeans IDE est un environnement de développement intégré gratuit et open source pour le développement d'applications sur les systèmes d'exploitation Windows, Mac, Linux et Solaris.

L'EDI simplifie le développement d'applications Web, d'entreprise, de bureau et mobiles qui utilisent les plates-formes Java et HTML5. L'EDI offre également un support pour le développement d'applications PHP et C / C ++. Voir la figure 4.1.

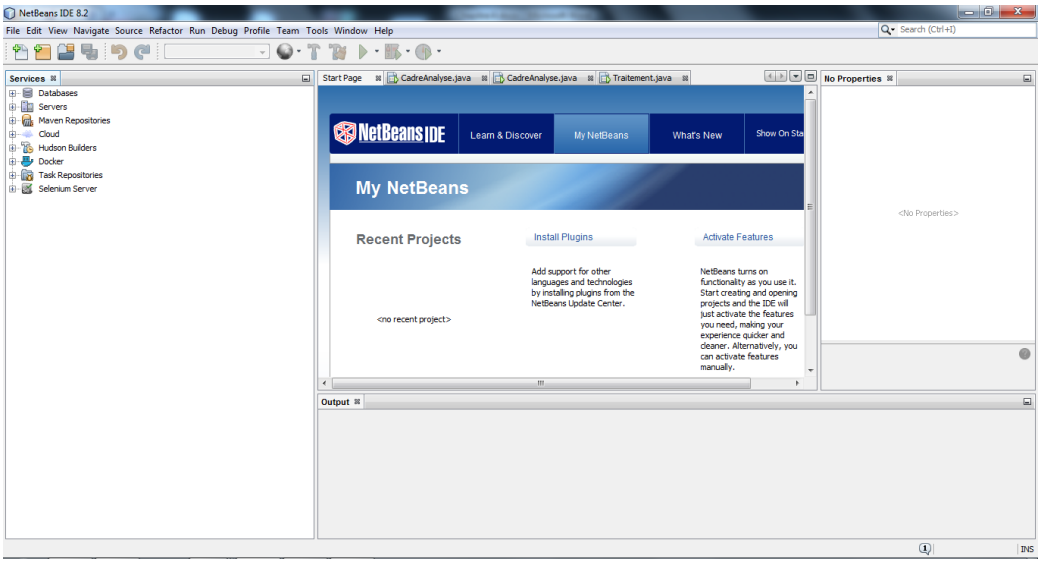


Figure 4.1. *Interface graphique de Netbeans IDE.*

3. Présentation de l'application

Notre application se compose de deux fenêtres principales: l'une d'analyse et d'extraction des informations de vérité terrain, et l'autre celle de génération de documents bruités.

3.1. Interface d'analyse et d'extraction des informations de vérité terrain

C'est l'interface principale de notre application (figure 4.2). La partie gauche de cette interface contient des onglets créés à chaque étape de traitement où nous affichons les résultats; au dessous des onglets, on a une glissière de zoome. Tandis que la partie droite contient les boutons de traitements. On laisse un seul bouton activé à la fois. L'appuie sur un bouton permet d'exécuter l'action de ce bouton et d'activer un autre bouton (le bouton du traitement suivant). Ainsi les boutons de cette interface sont: *Ouvrir image*, *Binarisation*, *Extraire lignes*, *Extraire mots*, *Extraire sous-mots*, *Composantes connexes*, *Informations de vérité terrain*, *Analyser un dossier auto*, et finalement *Fenêtre de bruitage*, et *Fenêtre de bruitage auto*, qui nous passent à l'interface de génération de bruit.

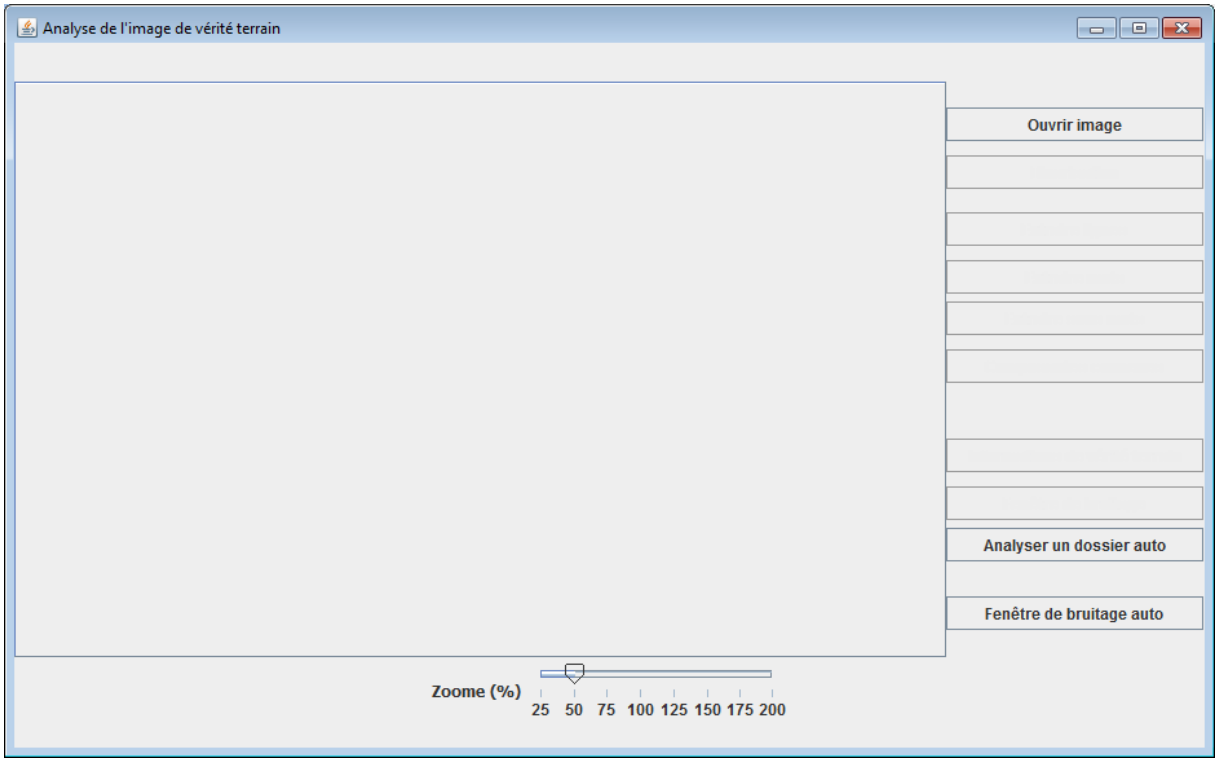
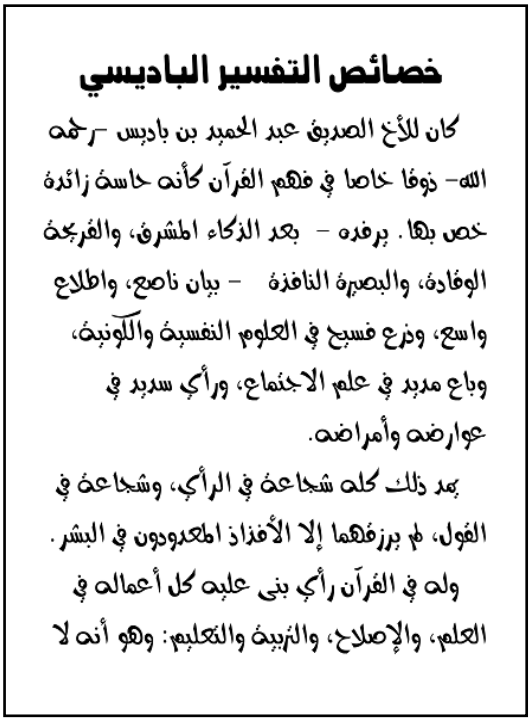


Figure 4.2. *Interface d'analyse et d'extraction des informations de vérité terrain.*

Dans la suite de notre démonstration nous allons utiliser l'image suivante comme un exemple de test.


خصائص التفسير الباديسي

كان للأخ الصديق عبد الحميد بن باديس –رحمه الله– ذوقا خاصا في فهم القرآن كأنه حاسة زائدة خص بها. يرفده – بعد الذكاء المشرق، والقريحة الوقادة، والبصيرة النافذة – بيان ناصع، واطلاع واسع، وزرع فسيح في العلوم النفسية والكونية، وباع مديد في علم الاجتماع، ورأي سديد في عوارضه وأمراضه.

بعد ذلك كله شجاعة في الرأي، وشجاعة في القول، لم يرزقهما إلا الأفذاذ المحدودون في البشر. وله في القرآن رأي بنى عليه كل أعماله في التعليم، والإصلاح، والتربية والتعليم: وهو أنه لا

Figure 4.3. *Exemple d'un image de test.*

3.1.1. Chargement d'images

L'outil débute par l'importation d'une image de texte. Pour ce faire nous cliquons sur le bouton « Ouvrir image ». Une boite de dialogue s'affiche permettant de choisir l'image de texte à traiter, voir la figure suivante:

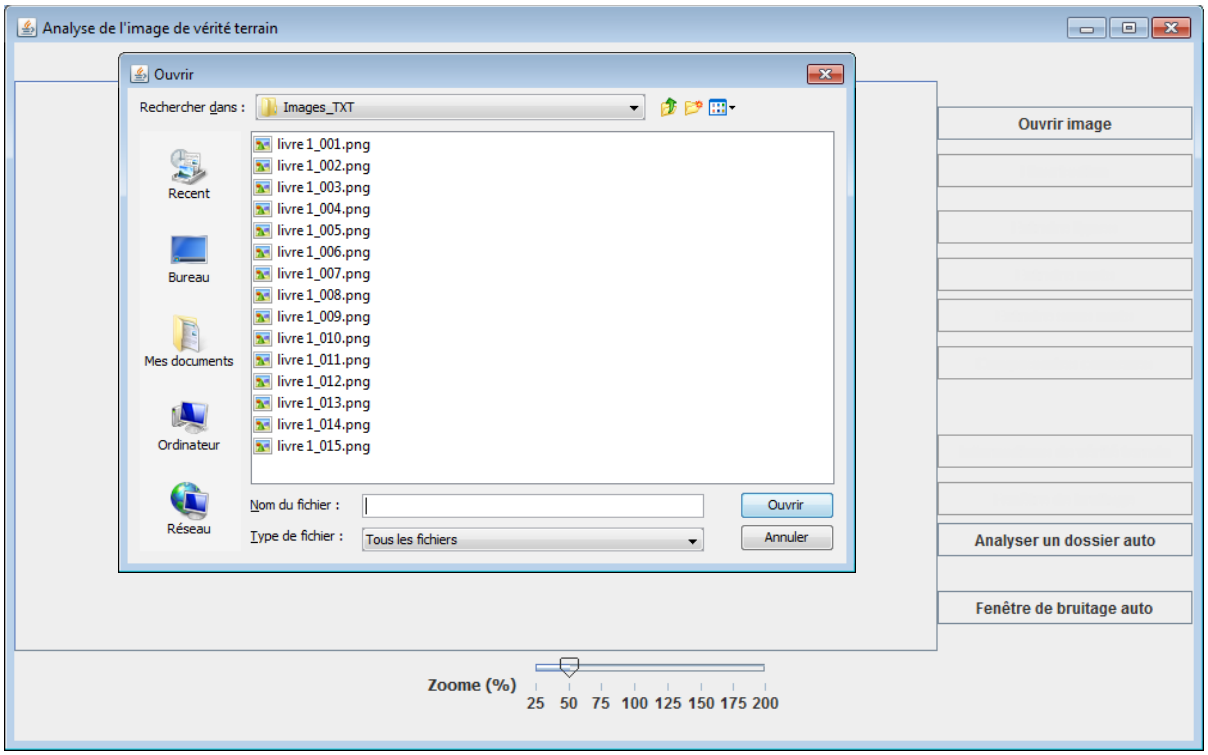


Figure 4.4. *Chargement d'une image de texte.*

En cliquant sur « ouvrir l'image», l'image s'affiche donc sur le panneau du premier onglet comme suit :

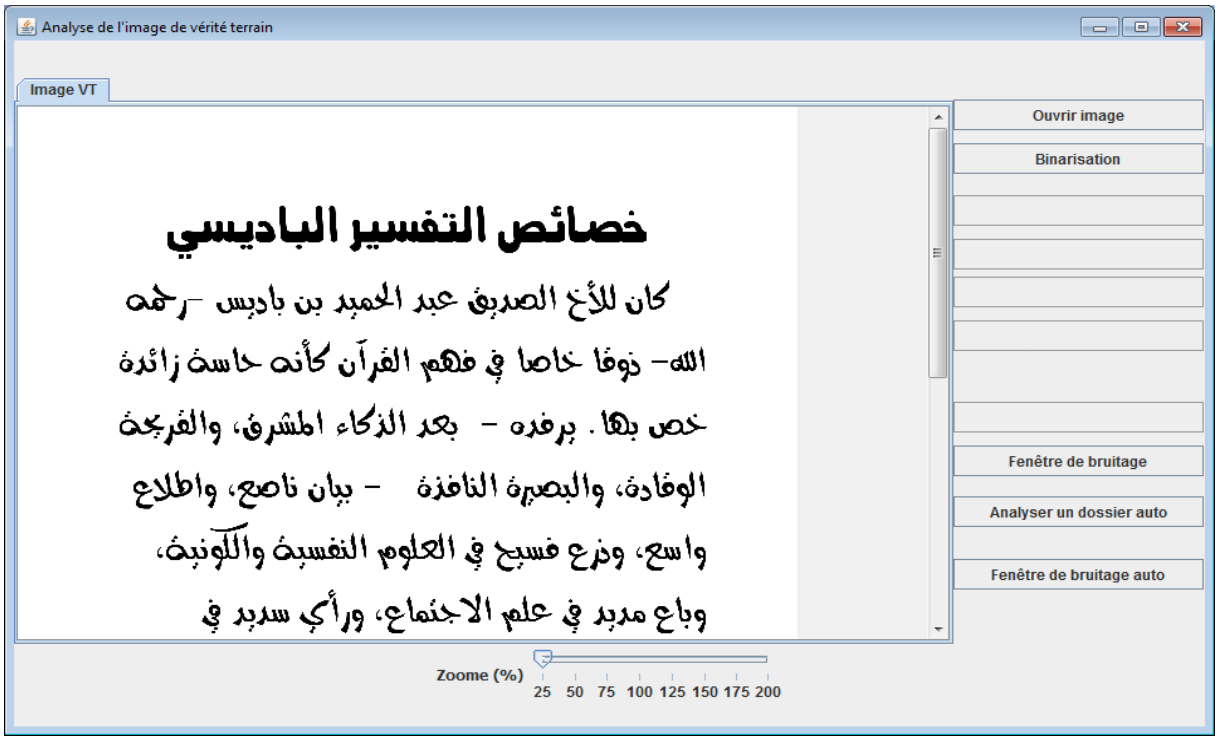


Figure 4.5. *Affichage de l'image chargée sur le panneau d'affichage.*

3.1.2. Binarisation

Cette étape donne une image binaire. L'image résultante s'affiche sur la fenêtre principale dans un nouvel onglet comme suit :

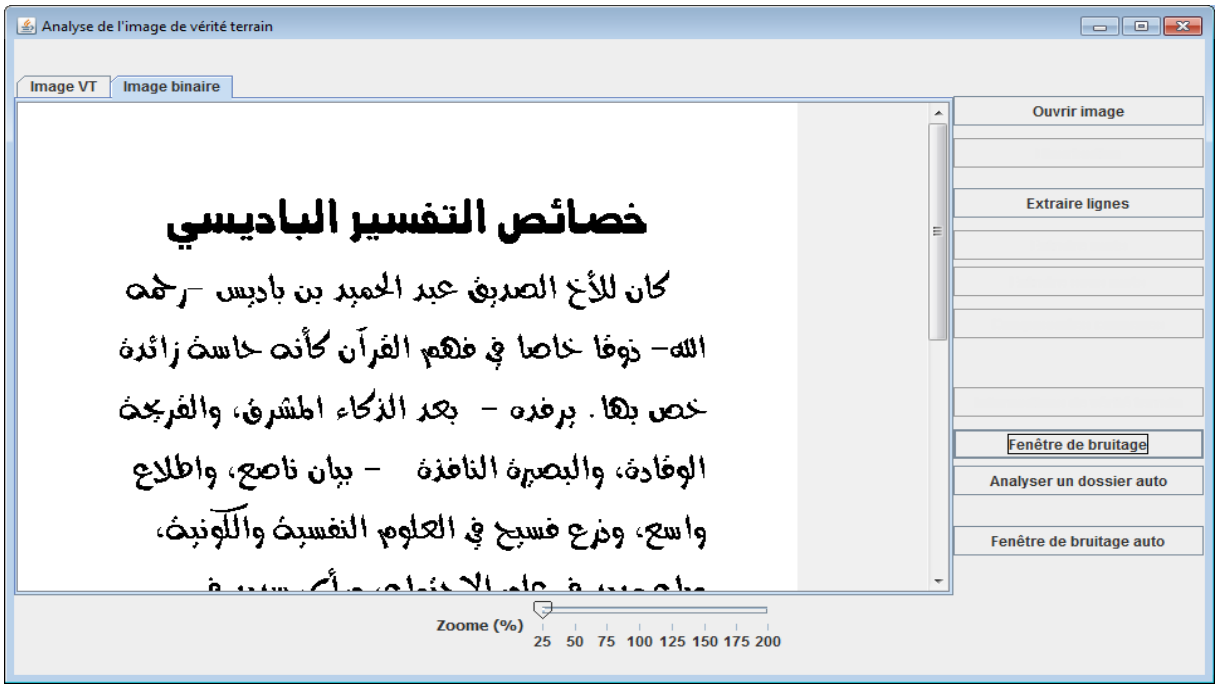


Figure 4.6. *Affichage de l'image binaire dans un nouvel onglet.*

3.1.3. Extraction des lignes

Lorsqu'on clique sur le bouton "Extraire linges", un nouvel onglet affiche limage de texte avec trompeuse l'écriture sur chaque ligne. Voir la figure suivante:

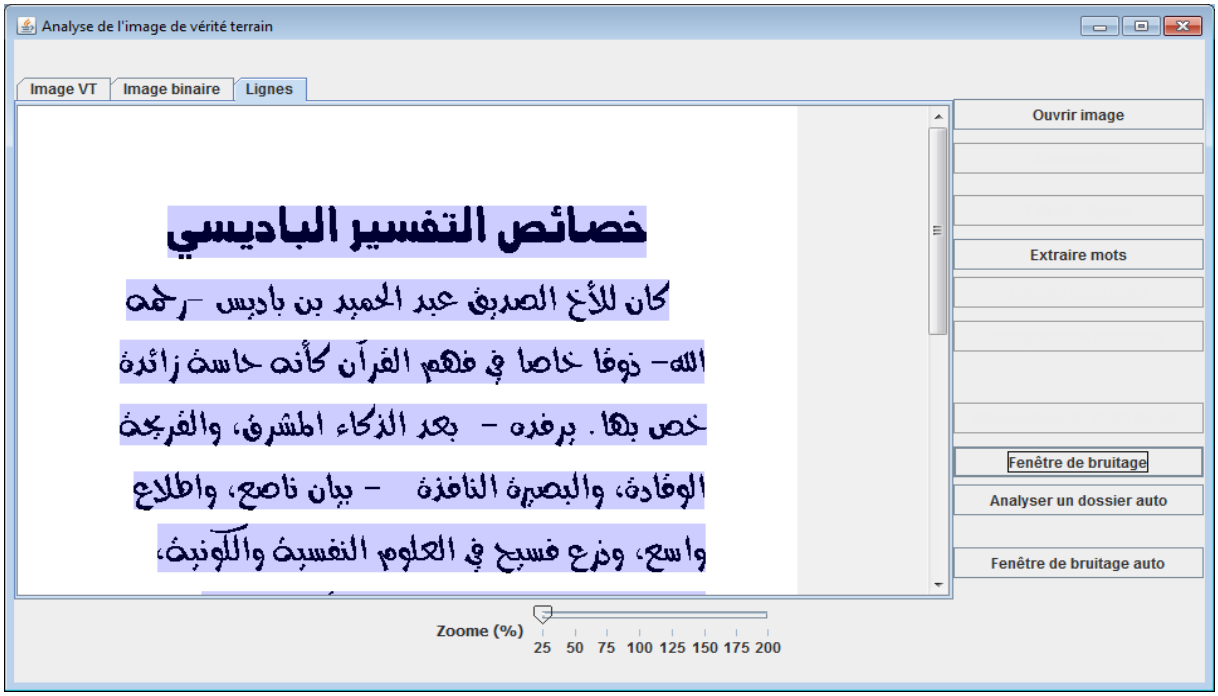


Figure 4.7. *Résultat de séparation en lignes affiché dans un nouvel onglet.*

3.1.4. Extraction des mots

Lorsqu'on clique sur le bouton "Extraire mots", un nouvel onglet affiche limage de texte avec trompeuse l'écriture sur chaque mot (figure suivante).

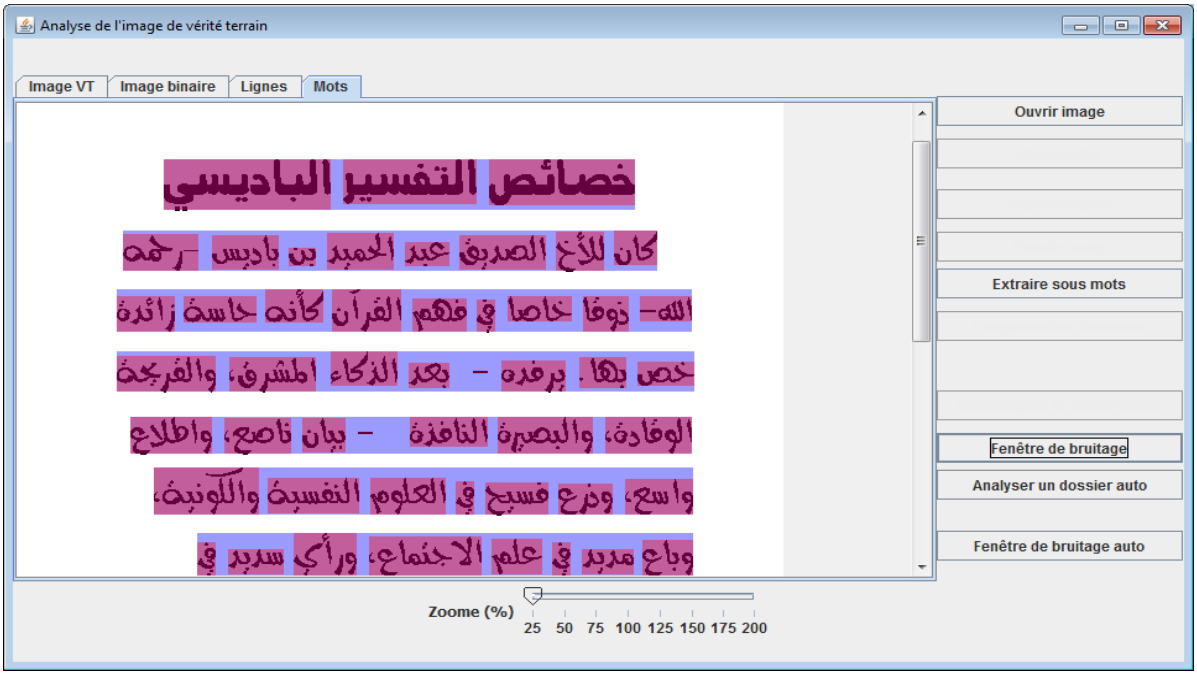


Figure 4.8. *Résultat d'extraction de mots affiché dans un nouvel onglet.*

3.1.5. Extraction des sous-mots

Lorsqu'on clique sur le bouton "Extraire sous mots", un nouvel onglet affiche limage de texte avec trompeuse l'écriture sur chaque sou-mot. Voir la figure suivante:

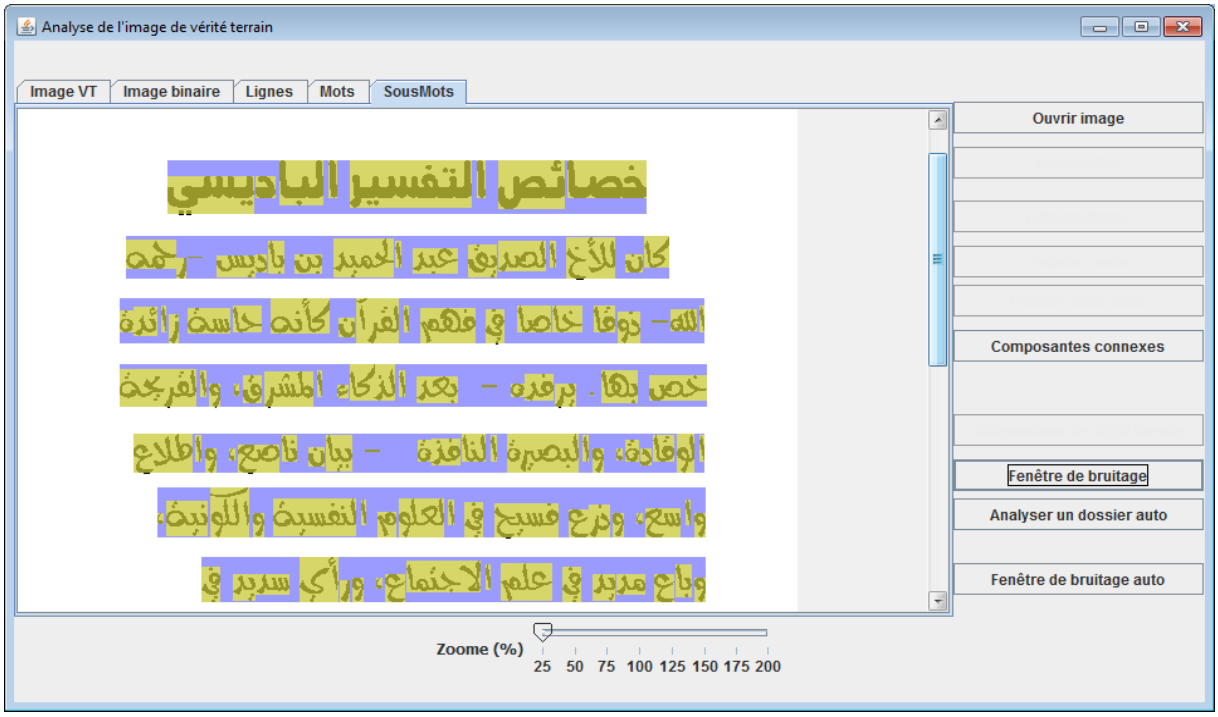


Figure 4.9. *Résultat d'extraction de sous-mots affiché dans un nouvel onglet.*

3.1.6. Extraction des Composantes connexes

Lorsqu'on clique sur le bouton "composantes connexes", un nouvel onglet affiche limage de texte avec trompeuse l'écriture sur chaque composante connexe.

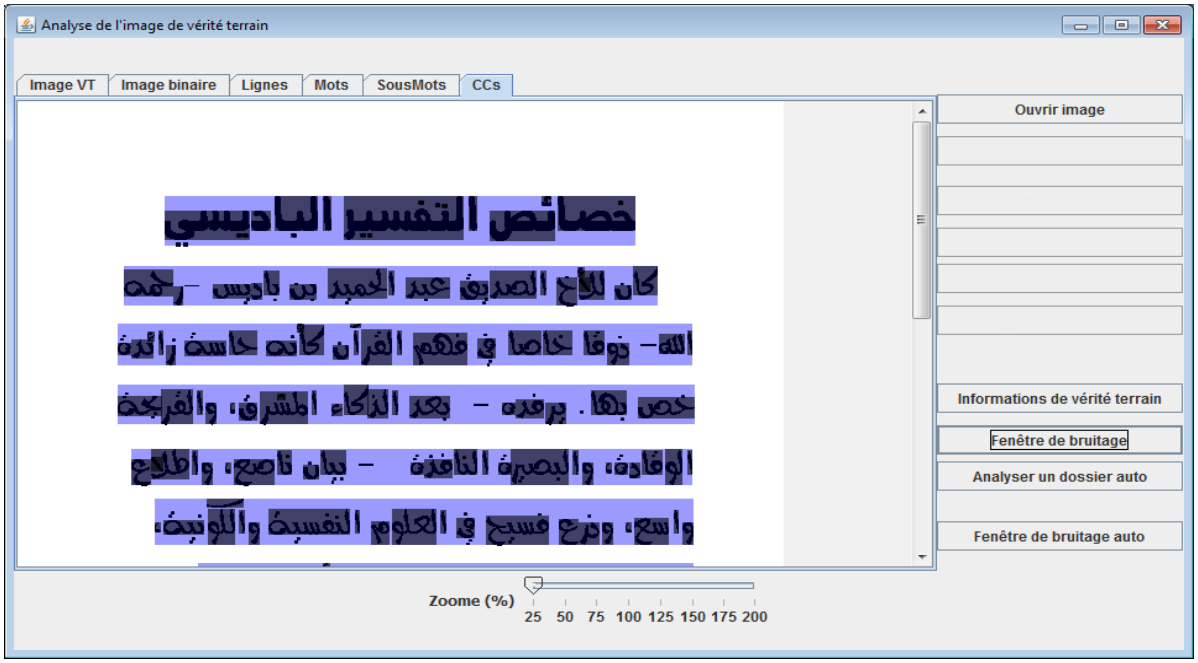


Figure 4.10. *Résultat d'extraction de composantes connexes.*

3.1.7. Génération d'un fichier XML

Nous cliquons sur le bouton «Informations de vérité terrain». Une boite de dialogue s'affiche permettant de choisir le ficher de texte correspondant à l'image choisie (le fichier texte de départ). Voir la figure suivante:

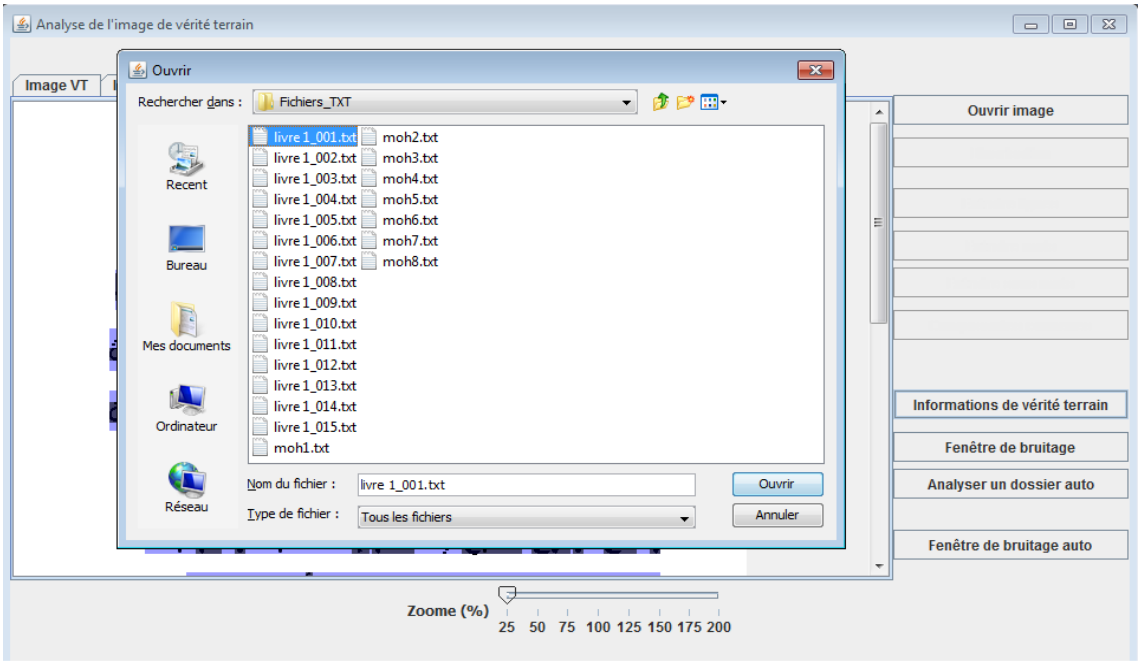


Figure 4.11. *Sélection du fichier texte correspondant à l'image traitée.*

En cliquant sur « ouvrir », le code XML qui contient les informations d'annotation de l'image s'affiche sur un nouvel onglet comme suit :

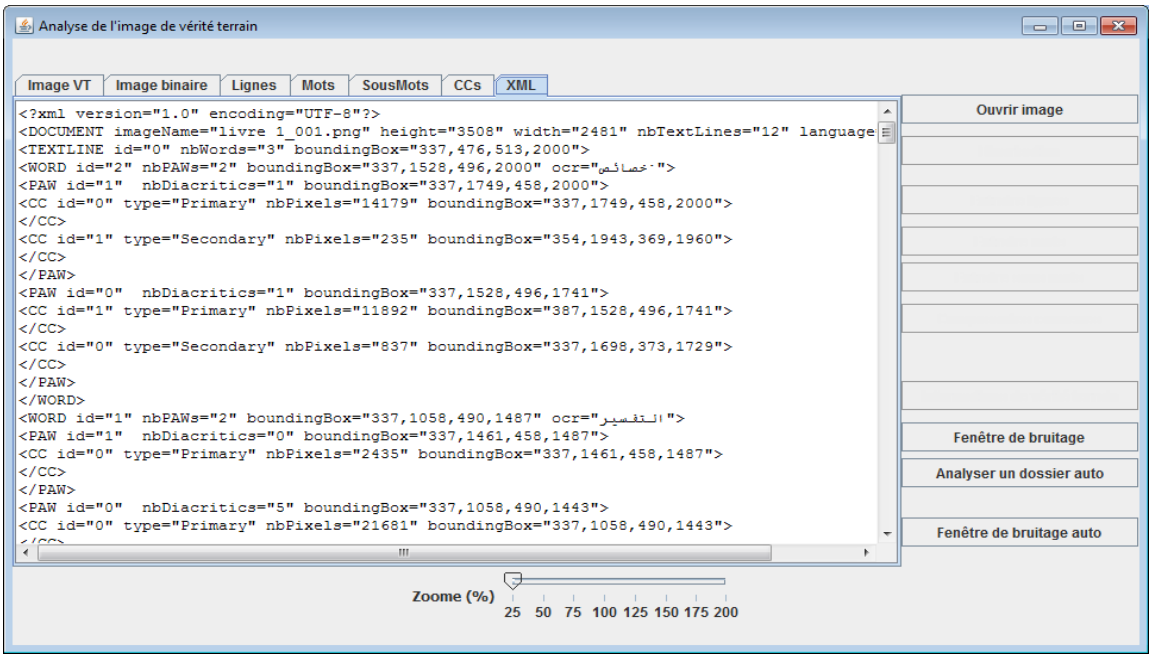


Figure 4.12. *Code XML généré.*

En cliquant sur le bouton "Fenêtre de bruitage", l'interface de génération de bruit s'affiche.

3.2. Interface de génération de documents bruités

La partie gauche de notre interface contient des onglets créés à chaque étape de génération de bruit. Au dessous des onglets on a une glissière de zoome. Tandis que la partie droite contient les boutons des bruit qui vous voulez ajouter: *Bruit local de Kanungo*, *Transparence*, *Combiner texte/fond*, *Rotation d'image*, *Courbure*., en plus de deux boutons permettant le bruitage automatique d'un dossier complet d'images. Voir la figure suivante:

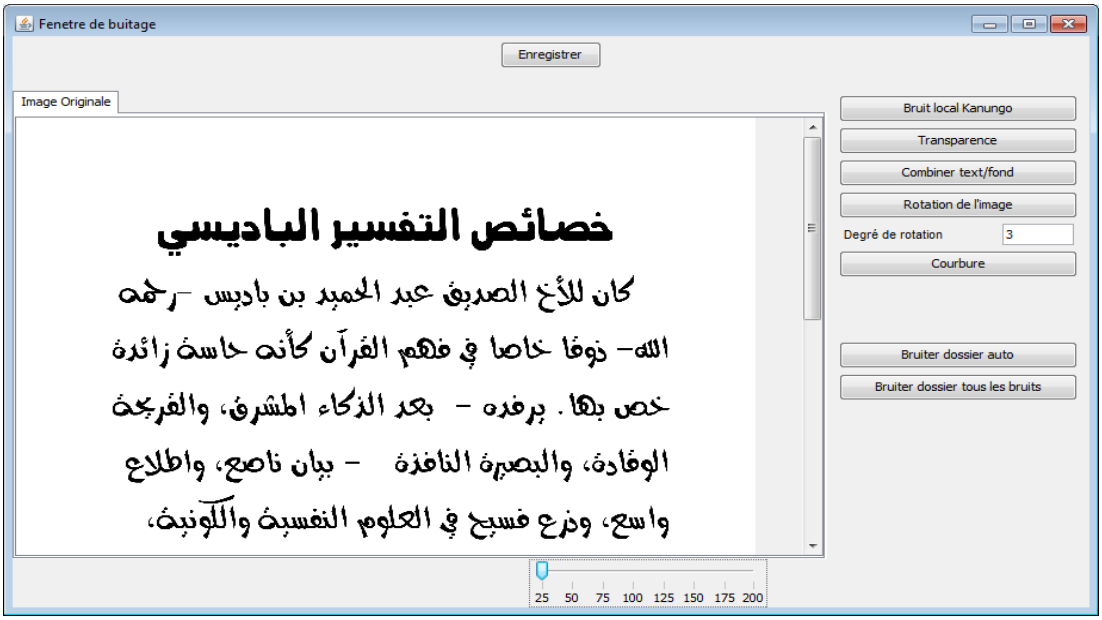


Figure 4.13. *Interface de génération de documents bruités.*

3.2.1. Ajout du bruit local de Kanungo

En cliquant sur « bruit local Kanungo », une glissière s'affiche permettant à l'utilisateur de choisir le pourcentage de bruit qu'il veut appliquer. Par défaut la valeur est 25%.

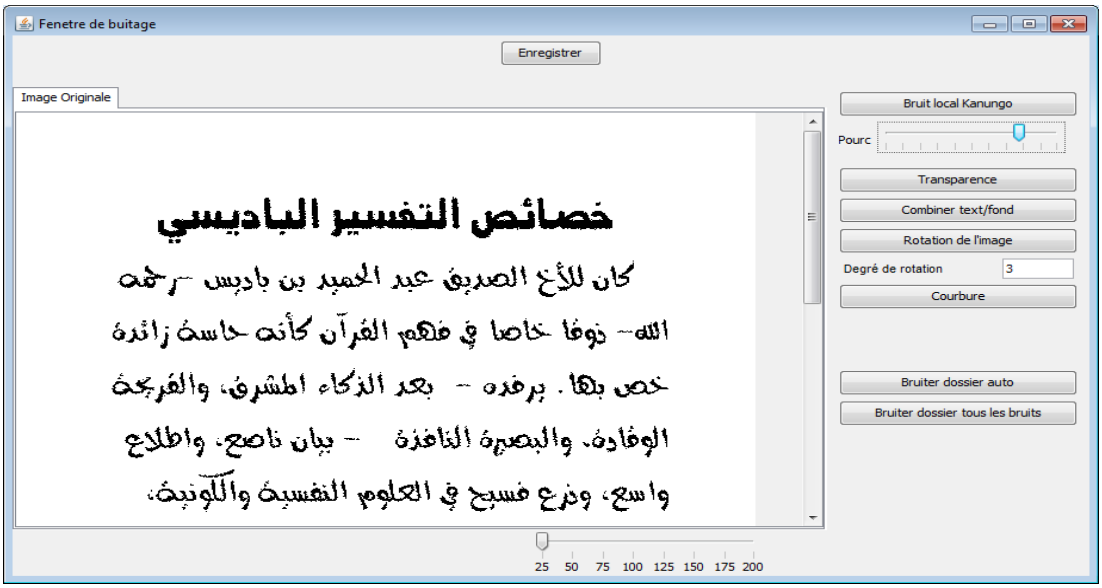


Figure 4.14. *Résultat de bruitage par le bruit local de Kanungo.*

3.2.2. Ajout de l'effet de transparence

En cliquant sur « Transparence», une boite de dialogue s'affiche permettant de choisir l'image de verso. En validant, le résultat s'affiche. Une glissière pour choisir le nivaux de gris de limage de verso de 0 a 255 est apparue.

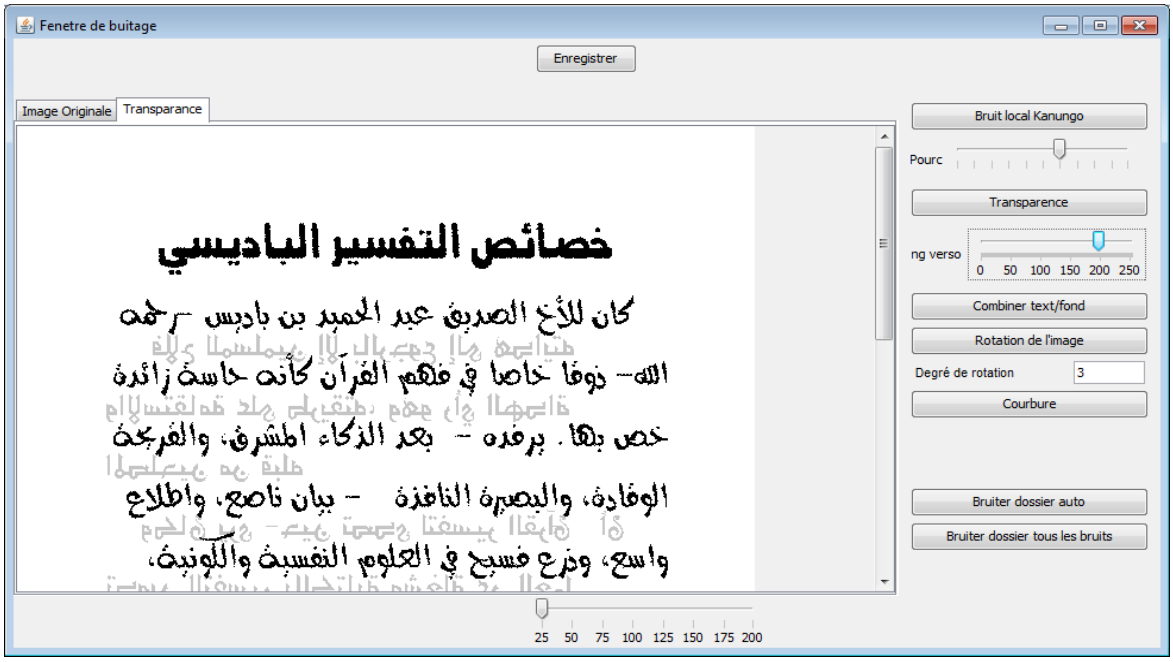


Figure 4.15. *Résultat de l'ajout de l'effet de transparence.*

3.2.3. Création d'image par combinaison texte/fond

En cliquant sur le bouton « Combiner texte/fond», une boite de dialogue s'affiche permettant de choisir l'image de fond à appliquer. Voir la figure suivante:

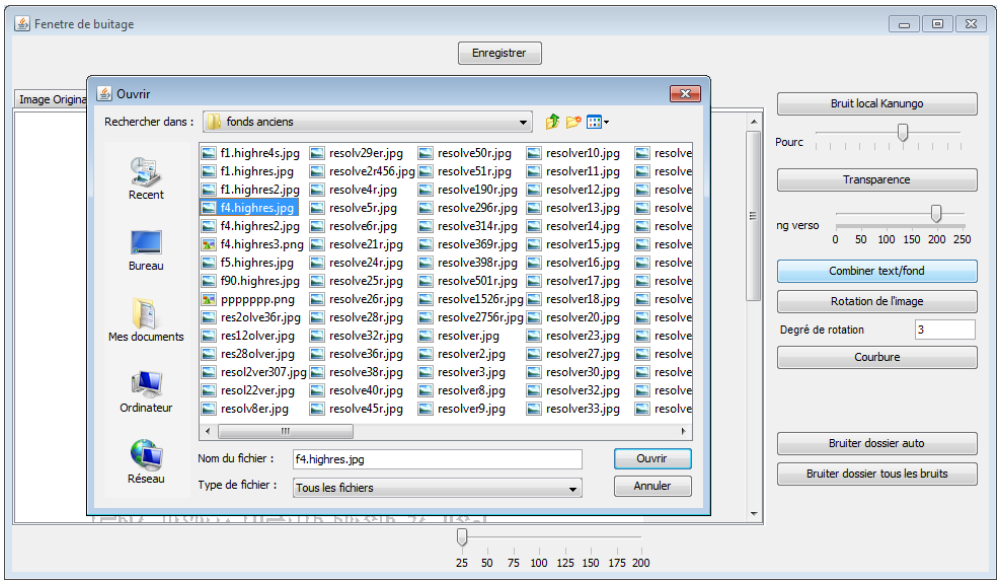


Figure 4.16. *Sélection d'une image de fond à combiner avec.*

Par le clique sur « ouvrir », l'image de combinaison s'affiche comme dans la figure suivante:

Figure 4.17. *Affichage de l'image résultante de la combinaison texte/ fond .*

3.2.4. Ajout de l'effet de rotation

On écrit dans la zone de texte le degré de rotation voulu puis on clique sur le bouton « Rotation de l'image», et le résultat s'affiche comme dans la figure suivante:

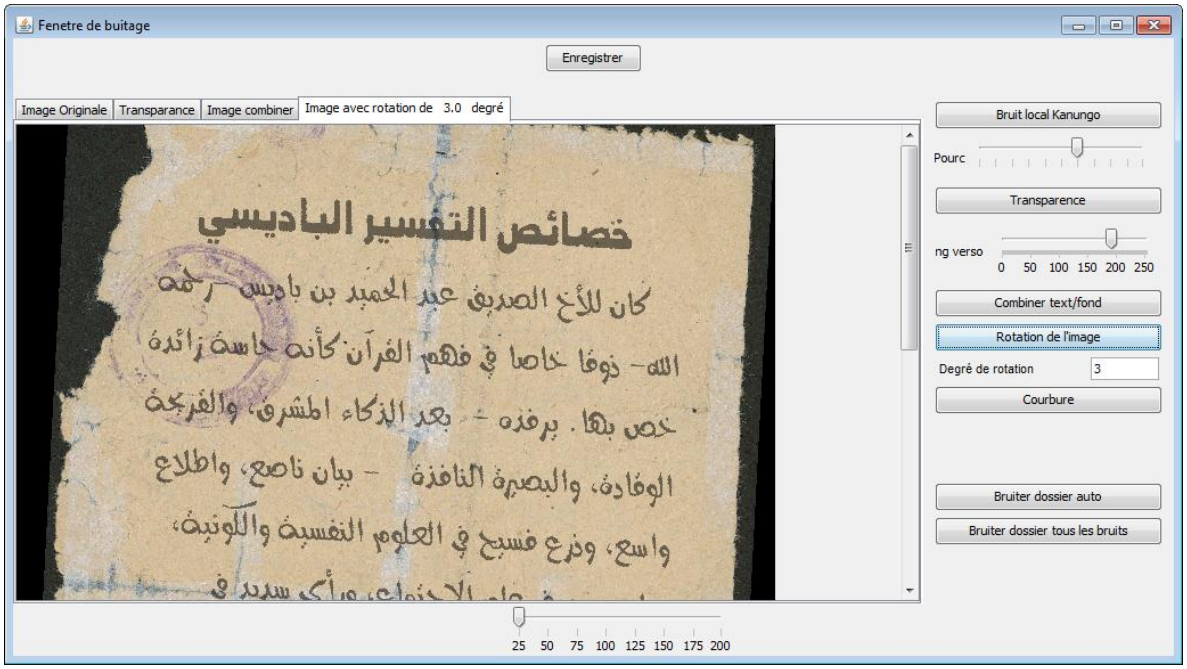


Figure 4.18. *Résultat de rotation de l'image de combinaison précédente.*

3.2.5. Ajout du bruit courbure

En cliquant sur « Courbure», s'affiche des glissières pour modifier et choisir le degré de courbure ainsi que d'autres paramètres. Le résultat s'affiche comme dans la figure suivante:

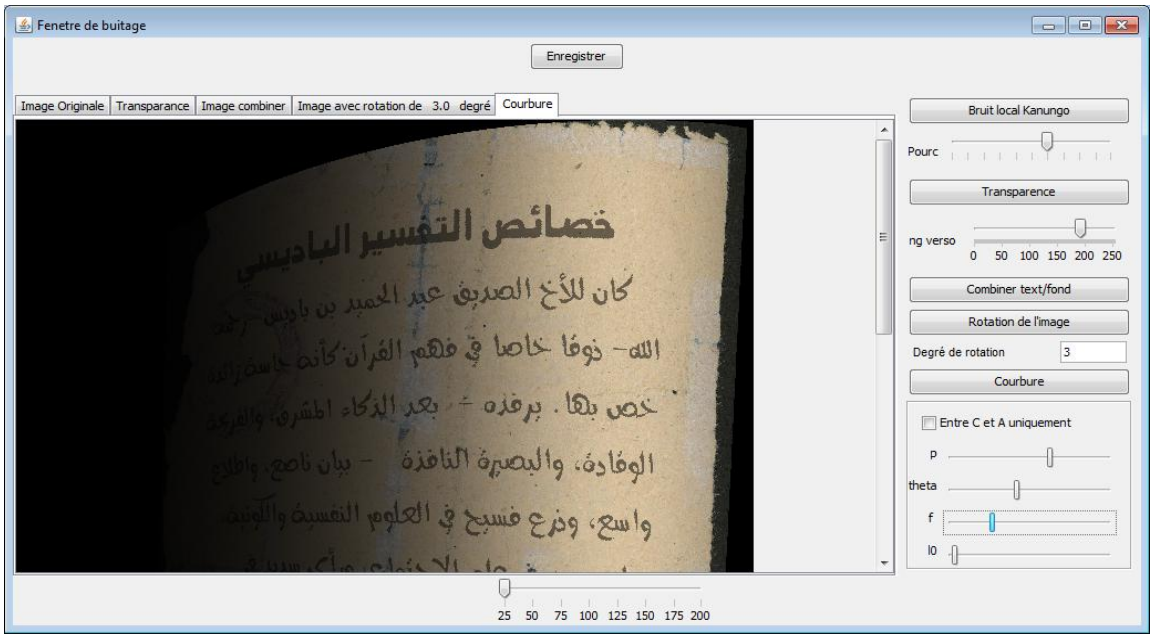


Figure 4.19. *Résultat de courbure de l'image de combinaison précédente.*

4. Base construite

La base construite jusqu'à présent est organisée en dossiers comme suit :

Dossier	**Nb d'images**	**Détails**
Fichiers_TXT	30	Fichier *.txt* de départ contenant les textes en ASCII
Images_Bin	30	Images de texte imprimé arabe en noir et blanc.
Fond_Im	46	Images de fonds anciens vierge
Images_Comb	1380	Résultants de la combinaison texte/ fond
Bruit_Kanungo	30	Résultants de bruitage des images par le ***Bruit local de Kanungo***
Bruit_Transparence	30	Résultants de bruitage des images par le ***Bruit de transparence***
Bruit_Rotation	30	Résultants de bruitage des images par le ***Bruit de rotation***
Bruit_Courbure	30	Résultants de bruitage des images par le ***Bruit de courbure***
Tous_Bruits	1380	Les images de ce dossier sont bruitées par **tous les bruits** successivement.
Informations_VT	30	Fichiers d'annotation en XML

Tableau 4.1*. Détails de la base construite.*

5. Conclusion

A travers ce dernier chapitre nous avons présenté un guide d'utilisateur de notre application de construction d'une base de documents bruités.

Conclusion générale et perspectives

Conclusion générale:

Aujourd'hui, la recherche scientifique se développe rapidement et tout le monde travaille dur pour suivre le rythme des progrès scientifiques. Le travail adressé dans ce mémoire a pour but la création d'une base de données pour l'analyse et la reconnaissance des documents arabes dégradés, et le développement d'un outil permettant de la construire. En effet, la construction d'une base de données est un projet très important dans le domaine de traitement et d'analyse de l'image en général et plus précisément dans le traitement et reconnaissance de documents. C'est surtout dans les phases d'apprentissage et de test qu'on a besoin de travailler avec des bases de données standardisées. Pour ce problème, nous avons essayé dans ce projet de contribuer à la création d'une base de données de documents arabes dégradés. Un outil est proposé pour la génération automatique, à partir des textes et d'images de vérité terrain, des images bruités en suivant deux stratégies différentes: combinaison par *mosaiciing* des images de texte avec des images de fonds anciens afin de produire des images synthétiques imitant les dégradations fréquemment rencontrées dans les documents historique, et génération de quelques types de bruit en s'inspirant de certains travaux de modélisation de bruit et de dégradation. Cela nous a permis de construire une base de données synthétique des anciens documents arabes. Des fichiers XML d'annotation ont été aussi créés comportant des informations importantes sur les documents de vérité terrain.

Ainsi, la base construite contient trois niveaux d'informations de vérité terrain: textes de vérité terrain, images binaire de vérité terrain, et fichiers d'annotation de vérité terrain, ce qui rend notre base appropriée pour plusieurs applications d'analyse et de reconnaissance de documents. Elle est adéquate aux application de reconnaissance de l'écriture à cause de l'existence des fichiers texte de vérité terrain qui peuvent être utilisés pour l'apprentissage des systèmes de reconnaissance et pour leur évaluation. Elle est adéquate également pour l'apprentissage ou l'évaluation des systèmes de segmentation (en lignes, en mots, en sous-mots, et en composantes connexes) vue la disponibilité des informations sur ces entités dans les fichiers d'annotation XML. Finalement la présence des images binaire de vérité terrain rend notre base adéquate à l'apprentissage, le test, et l'évaluation des méthodes de binarisation.

Perspectives:

Comme tout travail non complet, plusieurs perspectives sont envisageables:

- Continuer à enrichir la base de données en termes de quantité et de qualité.
- Inclure d'autres types de bruits et dégradation.
- Ajouter d'autres informations utiles aux fichiers d'annotations XML.
- Traiter d'autres images de documents plus complexes.

Références

Bibliographie:

[ALN 07] AL-NASSIRI, A. M. E. R., and SHUBAIR A. ABDULLA."A New Arabic (AHD/AMSH) Handwritten Database."

[ALK 15] AlKhateeb, Jawad H. "A database for Arabic handwritten character recognition." Procedia Computer Science 65 (2015): 556-561.

[ALM 02] Al-Ma'adeed, Somaya, Dave Elliman, and Colin A. Higgins."A data base for Arabic handwritten text recognition research." Proceedings eighth international workshop on frontiers in handwriting recognition. IEEE, 2002.

[ALO 03] Al-Ohali, Yousef, Mohamed Cheriet, and Ching Suen."Databases for recognition of handwritten Arabic cheques." Pattern Recognition 36.1 (2003): 111-121.

[AMA 05] Amara, Najoua Essoukri Ben, et al. "ARABASE: A Relational Database for Arabic OCR Systems." Int. Arab J. Inf. Technol. 2.4 (2005): 259-266.

[ALA 08] Alamri, Huda, et al. "A novel comprehensive database for Arabic off-line handwriting recognition." Proceedings of 11th International Conference on Frontiers in Handwriting Recognition, ICFHR. Vol. 8. 2008.

[ARR 06] Arrivault, Denis. Apport des graphes dans la reconnaissance non-contrainte de caractères manuscrits anciens. Diss. 2006.

[BAI 90] Baird, Henry S. "Document image defect models." Structured Document Image Analysis. Springer, Berlin, Heidelberg, 1992. 546-556.

[BAN 09] Banerjee, Serene, and Anjaneyulu Kuchibhotla. "Real-time optimal-memory image rotation for embedded systems." *2009 16th IEEE International Conference on Image Processing (ICIP)*. IEEE, 2009.

[BEN 99] Bennasri, A., A. Zahour, and B. Taconet. "Extraction des lignes d'un texte manuscrit arabe." Vision interface. Vol. 99. 1999.

[BON 01] Bonnet, Noël, and Jérôme Cutrona. "Improvement of Unsupervised Multi-component Image Segmentation through Fuzzy Relaxation." VIIP. 2001.

[CUR 97] Curtis, Cassidy J., et al. "Computer-generated watercolor." Proceedings of the 24th annual conference on Computer graphics and interactive techniques. 1997.

[CHE 08] Sylvain Chevalier, "Reconnaissance d'écriture manuscrite par des techniques markoviennes : une approche bidimensionnelle et générique", Thèse de doctorat, Université René Descartes - Paris 5, 2008.

[CHI 04] Shi, Zhixin, and Venu Govindaraju. "Line separation for complex document images using fuzzy runlength." First International Workshop on Document Image Analysis for Libraries, 2004. Proceedings.. IEEE, 2004.

[COT 97] Cote, Myriam. Utilisation d'un modèle d'accès lexical et de concepts perceptifs pour la reconnaissance d'images de mots cursifs. Diss. 1997.

[DRI 07] Drira, Fadoua. "Contribution à la restauration des images de documents anciens." Doctorat thesis (2007).

[ELA 08] El Abed, H., and V. Märgner. "Base de données et compétitions-outils de développement et d'évaluation de systèmes de reconnaissance de mots manuscrits arabes." 2008.

[ELM 08] El, Monji Kherallah Abdelkarim Elbaati Haikal, and Abed Adel M. Alimi."The On/Off (LMCA) Dual Arabic Handwriting Database." In : 11th International conference on frontiers in handwriting recognition (ICFHR). 2008.

[GUY 94] Guyon, Isabelle, et al. "UNIPEN project of on-line data exchange and recognizer benchmarks." Proceedings of the 12th IAPR International Conference on Pattern Recognition, Vol. 3-Conference C: Signal Processing (Cat. No. 94CH3440-5). Vol. 2. IEEE, 1994.

[HAR 92] R.M. Haralick and L.G. Shapiro. Machine Vision. Addison-Wesley Publishing Co., Inc., Reading, Massachusetts, 1992.

[HOR 86] Horn, Berthold, Berthold Klaus, and Paul Horn. Robot vision. MIT press, Cambridge, MA, 1986.

[HUL 94] Hull, Jonathan J. "A database for handwritten text recognition research." IEEE Transactions on pattern analysis and machine intelligence 16.5 (1994): 550-554.

[KIE 14] Van Cuong Kieu, et al. "Génération d'images semi-synthétiques de documents anciens à des fins d'évaluation de performances et d'apprentissage." *CORIA-CIFED*. 2014.

[KAN 99] Kanungo, Tapas, and Robert M. Haralick. "An automatic closed-loop methodology for generating character ground truth for scanned documents." IEEE Transactions on Pattern Analysis and Machine Intelligence 21.2 (1999): 179-183.

[KAN 93] Kanungo, Tapas, Robert M. Haralick, and Ihsin Phillips. "Global and local document degradation models." Proceedings of 2nd International Conference on Document Analysis and Recognition (ICDAR'93). IEEE, 1993.

[KIE 13] Kieu, V. C., et al. "Semi-synthetic document image generation using texture mapping on scanned 3d document shapes." 2013 12th International Conference on Document Analysis and Recognition. IEEE, 2013.

[LAM 17] Lamghari, Nidal, and Said Raghay. "DBAHCL: database for Arabic handwritten characters and ligatures." International Journal of Multimedia Information Retrieval 6.3 (2017): 263-269.

[LIA 08] Liang, Jian, Daniel De Menthon, and David Doermann. "Geometric rectification of camera-captured document images." IEEE Transactions on Pattern Analysis and Machine Intelligence 30.4 (2008): 591-605.

[MÄR 01] Margner, V., and Mario Pechwitz. "Synthetic data for Arabic OCR system development." Proceedings of Sixth International Conference on Document Analysis and Recognition. IEEE, 2001.

[MAR 14] Maroua, Mehri, et al. "Génération d'images semi-synthétiques de documents anciens à des fins d'évaluation de performances et d'apprentissage." 2014.

[MAH 11] Mahmoud, Sabri A., et al. "A database for offline Arabic handwritten text recognition." International Conference Image Analysis and Recognition. Springer, Berlin, Heidelberg, 2011.

[MOG 09] Moghaddam, Reza Farrahi, and Mohamed Cheriet. "Low quality document image modeling and enhancement." International Journal of Document Analysis and Recognition (IJDAR) 11.4 (2009): 183-201.

[NAG 00] Nagy, George, et al. "Two complementary techniques for digitized document analysis." Proceedings of the ACM conference on Document processing systems. 2000.

[OUW 10] Ouwayed, Nazih. Segmentation en lignes de documents anciens: application aux documents arabes. Diss. 2010.

[PEC 02] Pechwitz, Mario, et al. "IFN/ENIT-database of handwritten Arabic words." Proc. of CIFED. Vol. 2. Citeseer, 2002.

[PHI 93] Phillips, Ihsin T., et al. "English document database design and implementation methodology." Proceeding of the 2nd Annual Symposium on Document Analysis and Retrieval. 1993.

[RAB 13] Rabeux, Vincent. Évaluation de la qualité des documents anciens numérisés. Diss. Bordeaux 1, 2013.

[RAF 02] Rafael, C. "Gonzalez, Richard E. Woods, Digital Image Processing." (2002).

[STA 08] Stathis, Pavlos, Ergina Kavallieratou, and Nikos Papamarkos. "An evaluation survey of binarization algorithms on historical documents." 2008 19th International Conference on Pattern Recognition. IEEE, 2008.

[SLI 14] Slimane, Fouad, et al. "Icfhr2014 competition on arabic writer identification using AHTID/MW and KHATT databases." 2014 14th International Conference on Frontiers in Handwriting Recognition. IEEE, 2014.

[SUL 17] Sulaiman, Alaa, Khairuddin Omar, and Mohammad F. Nasrudin. "A database for degraded Arabic historical manuscripts." 2017 6th International Conference on Electrical Engineering and Informatics (ICEEI). IEEE, 2017.

[TOR 03] Torralba, Antonio. "Contextual priming for object detection." International journal of computer vision 53.2 (2003): 169-191.

[ZHA 03] Zhai, Jian, et al. "A line drawings degradation model for performance characterization." Seventh International Conference on Document Analysis and Recognition, 2003. Proceedings.. Vol. 3. IEEE Computer Society, 2003.

Webographie:

[W1] https://bvh.hypotheses.org consulté le 10/09/2020

[W2] https://www.bibliotheque.nat.tn/?lang=fr consulté le 10/09/2020

[W3] https://www.loc.gov/collections/george-washington-papers/about-this-collection/ consulté le 04/08/2020.

[W4] https://fr.wikipedia.org/wiki/Document consulté le 05/09/2020

[W5] http://etlcdb.db.aist.go.jp/ consulté le 03/08/2020

[W6] https://pxhere.com/fr/photo/1057404 consulté le 07/08/2020

[W7] https://gallica.bnf.fr/accueil/fr/content/accueil-fr?mode=desktop consulté le 07/06/2020

[W8] https://www.bnf.fr/fr/gallica-la-bibliotheque-numerique-de-la-bnf-et-de-ses-partenaires consulté le 07/06/2020

[W9] https://www.irif.fr/~carton/Enseignement/XML/Cours/Annexes/introduction.html consulté le 03/05/2020

[W10] https://fr.wikipedia.org/wiki/Segmentation_d%27image consulté le 05/09/20

[W11] http://yann.lecun.com/exdb/mnist/index.html consulté le 09/08/2020