RÉPUBLIQUE ALGÉRIENNE DÉMOCRATIQUE ET POPULAIRE
MINISTÈRE DE L'ENSEIGNEMENT SUPÉRIEUR
ET DE LA RECHERCHE SCIENTIFIQUE

UNIVERSITÉ SAAD DAHLEB – BLIDA 1

FACULTÉ DES SCIENCES
DÉPARTEMENT D'INFORMATIQUE

Mémoire de fin d'études en vue de l'obtention du diplôme de master II

Spécialité : Traitement Automatique de la Langue

Thème :

Amélioration du taux de la reconnaissance des caractères arabes

Présenté par :

M^elle BENCHABANE Katia **M^elle AICHOUNI Chaimaa**

Soutenu le 29/09/2019 devant le jury:

M^eme MEZZI M.	MCB	USDB	**Présidente**
Mr CHERIF-ZAHAR S.A.	MAA	USDB	**Promoteur**
M^eme KARAOUI F.	DR	AALA	**Co-Promotrice**
Mr NEHAL D.	MAA	USDB	**Examinateur**

Promotion : 2018 / 2019

Résumé

La reconnaissance des textes cursifs reste toujours un problème ouvert aussi bien dans sa forme imprimée que manuscrite. Ceci à cause des difficultés auxquelles sont confrontés les chercheurs et les développeurs, telles que la variabilité de la forme, du style, et de l'inclinaison de l'écriture. L'écriture manuscrite arabe est naturellement cursive, difficile à traiter, et présente une grande variabilité.

Dans le cadre de notre travail, qui consiste à contribuer à l'amélioration du taux de reconnaissance des caractères arabes, nous avons tenté d'étudier les techniques de bases utilisées de nos jours dans ce domaine, pour être capable par la suite, de développé un système de reconnaissance complet de la langue arabe.

Tout au long du processus de développement de notre système, nous avons contribué à plusieurs niveaux à l'amélioration de la qualité des résultats. Notre approche est basée sur une nouvelle méthode de reconnaissance qui considère à la fois les mots et les caractères isolés, la reconnaissance se fait grâce à un réseau de neurones à convolution.

Mots clés : OCR, reconnaissance, réseaux de neurones à convolution, taux de reconnaissance

Remerciements

En tout premier lieu, nous remercions le bon Dieu, tout puissant, de nous avoir donné la foi, la force et le courage d'accomplir ce modeste travail.
Nos sincères gratitudes vont à :

Notre promoteur **Monsieur Cherif-Zahar Sid-Ahmed Amine,** pour sa confiance, sa rigueur, sa patience et son exigence dans le travail.

Notre Co-promotrice **Madame Karaoui Fazia** pour son orientation, sa gentillesse, sa patience et sa disponibilité.

Nous remercions aussi , notre enseignant **Monsieur Bala Mahfoud** pour ses orientations et ses conseils qui nous ont été très précieux.

Nous exprimons toute notre sincère reconnaissance à **Madame Mezzi Melyara** pour avoir bien voulu accepter de présider le jury de ce mémoire, pour sa qualité d'enseignement et pour la motivation qu'elle a pu nous fournir durant ces deux dernières années.

Que **Monsieur Nehal Djillali**, trouve ici l'expression de nos vifs remerciements pour avoir bien voulu examiner ce travail.

Nous tenons à remercier l'ensemble du personnel de l'académie Algérienne de la Langue Arabe pour leur patience, leurs conseils pleins de sens et pour le suivi et l'intérêt qu'ils ont porté à nos travaux.

Enfin, nous n'oserons oublier de remercier tout le corps professoral de l'Université de Saad Dahlab – Blida 1, pour le travail énorme qu'il effectue afin de créer les conditions les plus favorables pour le déroulement de nos études.

Dans l'impossibilité de citer tous les noms, nos sincères remerciements vont à tous ceux et celles, qui de près ou de loin, ont permis par leurs conseils et leurs compétences la réalisation de ce mémoire.

Dédicaces

Je dédie ce travail à mes chers parents pour leur soutien indéfectible et sans limites, merci de m'avoir donné avec amour le nécessaire pour que je puisse arriver à ce que je suis aujourd'hui. Que dieux vous protège et que la réussite soit toujours à ma portée pour que je puisse vous combler de bonheur.

A ma chère sœur Kenza pour son encouragement permanent, son soutien moral et sa présence.

A ma meilleure amie Meriem qui a su être là lors des moments difficiles et à tous mes amis.

A toute ma famille pour leur soutien tout au long de mon parcours universitaire.

A ma chère binôme Chaimaa et à toute sa famille.

Que ce travail soit l'accomplissement de vos vœux tant allégués, et le fruit de votre soutien infaillible,

Merci d'être toujours là pour moi.

Dédicaces

A mes chers parents Mohammed et Razika, pour tous leurs sacrifices, leur amour, leur tendresse, leur soutien et leurs prières tout au long de mes études.

A mes chères sœurs Nour El Houda et Manel pour leur soutien.

A ma chère binôme Katia et à toute sa famille.

Puisse dieu vous donné santé, bonheur, courage et surtout réussite

Enfin, je souhaiter adresser mes remerciements à tous ceux qui m'ont aidé de près ou de loin pour la réalisation de ce travail.

Sommaire

Liste des abréviations

AOCR : Arabic Optical Character Recognition (Reconnaissance optique des caractères arabes : en français)

CNN : Convolutional neural network (Réseau de neurones à convolution : en français)

DPI : Dots Per Inch (points par pouce : en français)

OCR : Optical Character Recognition (reconnaissance optique de caractères : en français)

SRCAA : Système de reconnaissance des caractères académie arabe

Liste des tableaux

Liste des figures

Introduction générale

Contexte

La reconnaissance optique de l'écriture manuscrite relève du domaine de la reconnaissance des formes qui s'intéresse aux formes de caractères. Le but est d'attribuer à une forme un identifiant des prototypes de référence déterminés préalablement. Elle fait objet de l'avenir de la communication entre l'homme et la machine. En effet, depuis la fin des années soixante, les travaux intensifs accomplis dans l'OCR (reconnaissance optique de caractères) de l'écriture latine ont permis de l'intégrer dans plusieurs secteurs où le texte est la base de travail, principalement en bureautique, pour des buts d'indexation et d'archivage automatique de documents, en publication assistée par ordinateur (PAO) pour faciliter la composition à partir d'une sélection de plusieurs documents, dans la poste pour le tri automatique du courrier, dans une banque pour faciliter la lecture des montants de chèques, …

Problématique et objectifs :

Pour l'écriture arabe, la situation est totalement différente. Sa nature cursive et sa grande variabilité la rendent difficile à traiter bien dans sa forme imprimée que manuscrite. Ceci à cause des difficultés auxquelles sont confrontés les chercheurs et les développeurs, telles que la variabilité de la forme, du style, et de l'inclinaison de l'écriture. Ces problèmes causent une inertie notamment dans le choix des primitives (caractéristiques) pertinentes qui décrivent la variabilité de la morphologie des caractères et dans la segmentation.

Le but de notre travail est de contribuer à l'amélioration du taux de reconnaissance des caractères arabes sur différents niveaux tout en prenant compte des différentes caractéristiques de cette langue.

Organisation du rapport :

Le premier chapitre est un rappel de certaines notions générales de la reconnaissance optique de l'écriture, nous présentons aussi toutes les étapes nécessaires à la réalisation d'un système OCR.

Le deuxième chapitre étudie la langue arabe, ses caractéristiques et ses données graphiques. Dans la dernière partie de ce chapitre nous présentons quelques systèmes et travaux de recherche réalisés au sein de la reconnaissance de la langue arabe.

Dans le troisième chapitre, la segmentation et ses différents types sont abordés en détail.

Finalement, nous présentons la partie pratique de notre travail, ou une description détaillée de nos contributions est donnée, ainsi que les résultats obtenus.

Chapitre 1

Généralités sur la reconnaissance automatique de l'écriture manuscrite

1.1 Introduction

La reconnaissance automatique de l'écriture est un traitement informatique, qui aide à numériser des documents écrits sur papier (manuscrits, imprimés, ou encore dactylographié) pour des utilisations ultérieures. Le système prend en entrée un document, et plusieurs opérations de traitement et de reconnaissance successives sont alors appliquées pour obtenir une description syntaxique de ses éléments significatifs. Nous proposons dans ce chapitre une description des algorithmes les plus souvent cités et utilisés dans les diverses étapes du processus de reconnaissance de l'écriture.

1.2 Types de reconnaissance de l'écriture manuscrite :

1.2.1 Reconnaissance de caractères :

C'est la tâche la plus basique d'un système de reconnaissance de l'écriture, L'effort d'analyse est concentré sur un seul élément à la fois du vocabulaire.

1.2.2 Reconnaissance de mots :

On s'intéresse à la reconnaissance du mot entier. Deux types de reconnaissance peuvent être considérés, la reconnaissance globale où le processus de la reconnaissance est ramené à un processus de reconnaissance de caractères avec des formes plus complexes que les caractères. Le deuxième type est une reconnaissance analytique où le mot est divisé en graphèmes. Elle cherche à identifier les caractères ou les sous-caractères (graphèmes) issus de la segmentation (séparation de mots, des caractères) pour reconstituer les mots. Elle permet une discrimination plus fine des mots car elle se base sur la reconnaissance des lettres qui la composent et il est

possible de récupérer l'orthographe du mot reconnu. Son inconvénient principal demeure la nécessité de l'étape de segmentation avec les problèmes de sous- ou de sur-segmentation que cela implique.

1.2.3 Reconnaissance de documents

Cette approche a une vision générale du mot; elle se base sur une description unique de l'image du mot, vue comme une entité indivisible. Disposant de beaucoup d'informations, elle absorbe plus facilement les variations au niveau de l'écriture. Cette méthode est pénalisante par la taille mémoire, le temps de calcul et la complexité du traitement qui croient linéairement avec la taille du lexique considéré, d'où une limitation du vocabulaire souvent appliquée pour réduire la liste de mots candidats

1.3 Généralités sur les images

1.3.1 Définition d'une image

L'image numérique est une représentation à deux dimensions d'une scène en trois dimensions, divisée en éléments de tailles fixes appelés cellules ou pixels, ayant chacun comme caractéristique un niveau de gris ou de couleur prélevé à l'emplacement correspondant dans l'image réelle, ou calculé à partir d'une description interne de la scène représenter (Gaudin, 2002). La numérisation d'une image consiste en la conversion de celle-ci en une image numérique représentée par une matrice bidimensionnelle de valeurs numériques f(x, y) où x, y sont les coordonnées cartésiennes d'un point d'une image.

1.3.2 Caractéristiques d'une image numérique

Une image numérique regroupe un ensemble de paramètres :

1.3.2.1 Pixel

Une image numérique est composée d'une grille de pixels qui apparaissent comme de petits carrés, porteurs d'une information de couleur élémentaire. Le pixel est le plus petit élément de

l’image, la quantité d'information que véhicule chaque pixel donne les nuances entre image en niveau de gris et l'image couleur. Dans une image en niveaux de gris, chaque pixel est codé sur un octet. Dans une image couleur (RVB) un pixel est codé sur trois octets Un octet pour chacune des couleurs : (R) Rouge, (V) Vert, (B) Bleu.

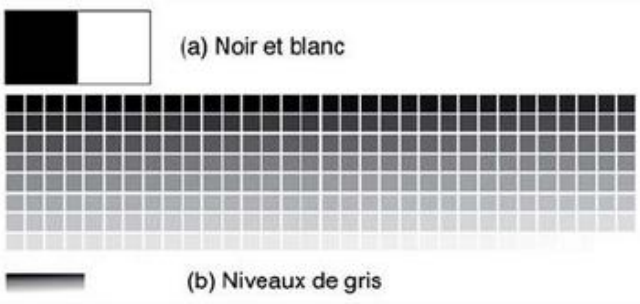


Figure 1.1 : Pixels sur une image noir et blanc dans (a) ou sur une image en niveau de gris dans (b)
Source :(Ameisen, 2012)

1.3.2.2 Dimensions

On appelle dimension, le nombre de points (pixel) constituant l'image, c’est-à-dire sa (dimension informatique). Cette dernière se présente sous forme de matrice dont les éléments sont des valeurs numériques représentatives des intensités lumineuses (pixels).
Le nombre de lignes de cette matrice multiplié par le nombre de colonnes nous donne le nombre total de pixels dans une image. (Fruitet, 2009)

1.3.2.3 Résolution de l’image

La résolution d'une image est le nombre de pixels par pouce qu'elle contient (1 pouce = 2.54 centimètres). Elle est exprimée en "PPP" (points par pouce) ou DPI (dots per inch). Plus il y a de pixels (ou points) par pouce et plus il y aura d'information dans l'image (plus précise). Par exemple, une résolution de 300dpi signifie que l'image comporte 300 pixels dans sa largeur et 300 pixels dans sa hauteur, elle est donc composée de 90 000 pixels (300x300 ppp). Grâce à cette formule, il est facile de connaître la dimension maximale d'un tirage.

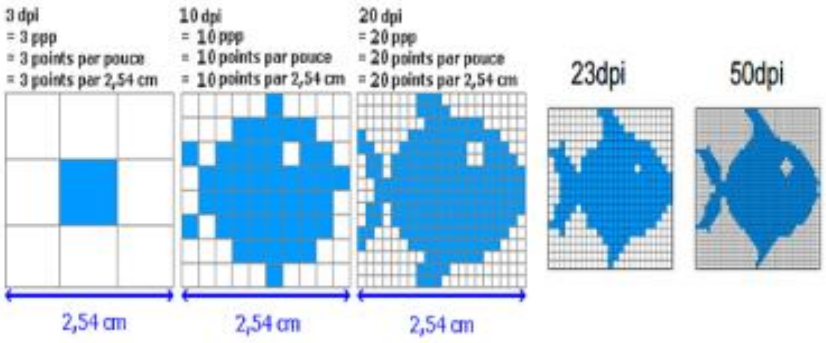


Figure 1.2: exemple d'une résolution d'image

1.3.2.4 La luminance

La luminance représente le degré de luminosité des différents points de l'image. Elle est définie aussi comme étant le quotient de l'intensité lumineuse d'une surface par l'aire apparente de cette surface, pour un observateur lointain, le mot luminance est substitué au mot brillance, qui correspond à l'éclat d'un objet. (Isdant, 2009)

1.3.2.5 Le contraste

Le contraste est une propriété de l'image qui désigne et quantifie la différence entre les parties claires et foncées d'une image.

1.3.3 Types d'images

Il existe généralement deux grandes familles d'images numériques, matricielles et vectorielles. Les images matricielles ou bitmap reposent sur une grille de plusieurs pixels formant une image avec une définition bien précise (Chakib et Sali, 1999). Les formats bitmap les plus répondus sont : BMP, GIF, JPEG, TIFF, PNG... il existe aussi deux types de format compressés (PNG, JPG) et non compressés (BMP, TIFF). Les images vectorielles quant à elles sont composées d'objets géométriques individuels (segments de droite, polygones, arcs de cercles,…) (Chakib, 1999), elles ont l'avantage de pouvoir être agrandie sans perdre leurs qualités initiales contrairement aux images matricielles.

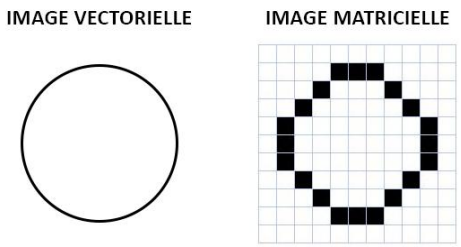


Figure 1.3: différence entre une image vectorielle et une image matricielle

1.3.4 Les différents modes colorimétriques d'une image

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

1.3.4.1 Image monochrome

Le monochrome représente le mode le plus simple à traiter, chaque pixel est soit allumé (blanc) ou éteint (noir), l'image obtenue n'est pas très nuancée.

1.3.4.2 L'image à niveaux de gris

On trouve dans une image à niveaux de gris différentes nuances de gris, l'intensité de chaque élément varie entre 0 (noir) à 255 (blanc)

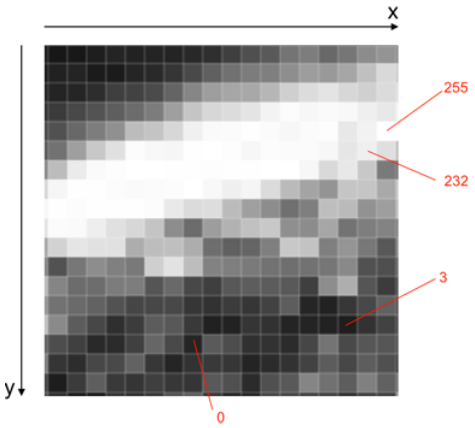


Figure 1.4: pixels d'une image en niveaux de gris avec quelques valeurs numériques

1.3.4.3 Image en couleurs

Les images en couleurs sont très souvent utilisées par les applications multimédias contrairement aux images noir et blanc ou en niveau de gris. La représentation des couleurs nécessite d'abord le choix d'un espace de couleurs à plusieurs dimensions pour donner suffisamment de composantes numériques pour décrire une couleur. On retrouve plusieurs représentations d'images en couleurs tel que :

- La représentation en couleurs réelle sur 24 bits.
- La représentation en couleurs indexées, avec utilisation d'une table appelée palette pour éviter la redondance de couleur.
- Le mode RGB (de l'anglais red, green, blue) représente l'espace le plus utilisé pour le maniement des images. Il est basé sur une synthèse additive des couleurs (256 teintes de rouge, 256 teintes de vert, 256 teintes de bleu), où chaque pixel est définit par 3 octets, ce qui permet quasiment la reproduction à l'écran toutes les couleurs du spectre visible (environs 16,8 millions de couleurs au total). (Isdant, 2009)

1.3.5 Région d'une image

La notion de région consiste à regrouper des zones possédant les mêmes caractéristiques. C'est-à-dire que si plusieurs pixels adjacents ont une couleur identique alors la zone qui qu'ils forment est une région

1.3.6 Contours d'une image

Les contours représentent la limite entre les objets de l'image, ou la limite entre deux pixels dont les niveaux de gris sont significativement différents '4)

1.3.7 Le bruit

Le bruit consiste en un signal qui s'ajoute à l'image lors de la phase d'acquisition, il est le résultat de certains défauts électroniques du capteur et de la qualité de numérisation.

Les sources de bruits d'une image sont nombreuses on peut citer :

- bruits liés aux conditions de prise de vue (bougé, éclairage de la scène).
- bruits liés aux capteurs (appareil numérique de bas de gamme).
- bruits liés à l'échantillonnage.
- bruits liés à la nature de la scène (poussières, rayures)

Ces bruits peuvent être catégorisés selon l'effet qu'ils ont sur l'image. Le bruit gaussien par exemple dégrade l'image en apparaissant sous forme de grains dans le cas d'une faible luminosité lors de l'acquisition de l'image. Le bruit sel et poivre lui apparait généralement sur des images en niveaux de gris (bruit impulsionnel) peut être représenté par des points noirs (poivre) et des points blancs (sel), il peut se joindre à une image suite à des erreurs de conversion ou bien de transmission de bit.

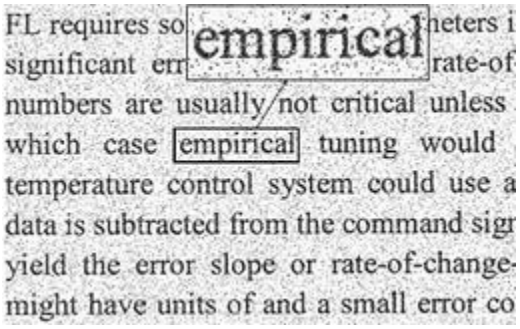


Figure 1.5 : bruit sel et poivre sur une image

1.4 Différents aspects d'un système de reconnaissance optique de caractères (OCR)

1.4.1 Reconnaissance de caractères imprimés ou manuscrits :

Les caractères imprimés simplifient généralement la phase de lecture vu qu'ils sont généralement alignés horizontalement et séparés verticalement, bien que certaines fontes présentent parfois des accolements qu'il faut repérer. Dans le manuscrit, les caractères sont souvent ligaturés et leur graphisme est inégalement proportionné .cela nécessite généralement l'emploi de techniques de délimitation spécifique, et souvent des connaissances contextuelles pour guider la lecture.

1.4.2 Reconnaissance mono-fonte, multi-fonte, omni-fonte

Dans le cas d'un texte imprimé, un système est dit mono-fonte s'il ne traite qu'une fonte à la fois, il est dit multi-fonte s'il est capable de reconnaître un mélange de fontes préalablement apprises, enfin un système est dit omni-fonte s'il est capable de reconnaître toute fonte sans l'avoir apprise.

1.4.3 Reconnaissance en ligne (on-line) ou hors ligne (off-line)

La méthode on-line se fait pendant l'écriture, elle permet de corriger ou de modifier l'écriture de manière directe et instantanée, tandis que l'utilisation de la méthode off-line nécessite l'acqusition du document entier avant de pouvoir commencer l'étape de reconnaissance, elle permet d'analyser un grand nombre de caractères, résultant d'un prétraitement coûteux.

1.5 Organisation générale d'un système de reconnaissance optique de l'écriture :

1.5.1 Phase d'acquisition :

La phase d'acquisition consiste à acquérir l'image numérique du document à l'aide de capteurs physiques (scanner, caméra, ..) et de la convertir par la suite en grandeurs numériques adaptés au système de traitement utilisé en évitant au maximum les dégradations qui peuvent réduire la qualité de l'image.

1.5.2 Phase de prétraitements :

Une fois l'acquisition terminée, l'image obtenue peut contenir du bruit qui dépend de plusieurs facteurs tel que :

- Le dispositif d'acquisition : capteur.
- Les conditions de prise de vue : éclairage, positionnement incorrecte du document
- La scène elle-même : poussière, rayure.
- La qualité du document d'origine : fond, composition des formes, nature de la matière.

Le but des prétraitements est la réduction du bruit, de la distorsion, et de la variation des styles pour faciliter les traitements ultérieurs tels que la segmentation et l'extraction de primitives. Ils comprennent principalement : la binarisation, la suppression du bruit, le lissage du contour, la squelettisation, l'estimation de la ligne de base, et les normalisations (de la taille, de l'inclinaison des lignes, et de l'inclinaison des caractères). L'utilisation des différentes méthodes et techniques existantes dépend de la nature des données traitées (scripte et qualité), et du type des primitives à extraire (invariante ou non à la distorsion et à la variation des styles).
Comme l'indique S. Madhvanath (S. Madhvanath et al, 1999), les prétraitements ont une influence major sur les performances des systèmes de reconnaissance de l'écriture manuscrite. Parmi les techniques les plus utilisées on retrouve :

1.5.2.1 La binarisation :

La binarisation est souvent la première étape dans les systèmes de traitement et d'analyse d'images (Trier et Taxt, 1995)(Leedham et al, 2002), et plus particulièrement d'images de documents . Cette opération permet de diminuer la quantité d'informations présentes dans l'image, et de ne garder que les informations pertinentes pour faciliter les traitements ultérieurs sur des images à niveaux de gris.

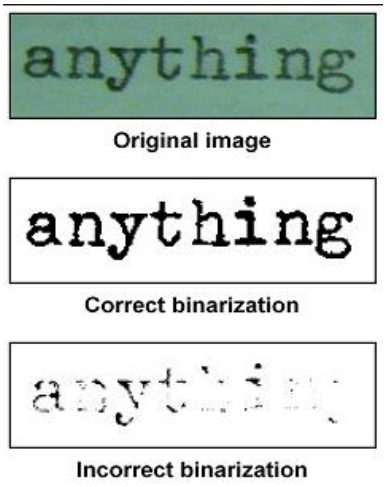


Figure 1.6 : Résultat d'une bonne et d'une mauvaise binarisation

Les performances des étapes suivantes dans les systèmes d'analyse de documents dépendent fortement des résultats de l'algorithme de binarisation utilisé, c'est pour cette raison que plusieurs techniques ont été proposées ces deux dernières décennies. Selon plusieurs travaux de recherche (Arica et Yarman-Vural, 2001) (Khurshid et al, 2009), les techniques de binarisation d'images en niveaux de gris peuvent être classées en deux catégories: seuillage globale, où un seul seuil est utilisé dans toute l'image pour la deviser en deux classes (texte et fond), et seuillage local où les valeurs des seuils sont déterminées localement, pixel par pixel ou bien région par région. D'autres (Sauvola et Pietinkäinen, 2000), ajoutent un troisième groupe de méthodes hybrides, ces méthodes combinent des informations globales et locales pour attribuer les pixels à l'une des deux classes.

- **Les méthodes globales :**

  Considérée comme la technique de binarisation la plus simple (Gupta et al ,2007), elle consiste à comparer le niveau de gris de chaque pixel p_i de l'image avec un seuil global fixe S (par exemple 130). On note n_i la nouvelle valeur du pixel, le seuillage est donné par l'expression suivante :

  $$n_i = 255 \text{ si } p_i \geq S \text{ et } n_i = 0 \text{ si } p_i < S$$

- *Méthode d'Otsu*

  Le but de cet algorithme est la binarisation d'images à niveau de gris en utilisant un seuillage automatique à partir de la forme de l'histogramme qui est calculée au préalable. La méthode d'*Ostu* sépare alors les pixels en deux classes, une classe qui contient *les objets* avec un seuil maximal (typiquement 255) et une classe contenant l'*arrière-plan* de l'image avec un seuil minimal(0). L'algorithme itératif balaye toutes les valeurs de seuil possibles sur une image en calculant la moyenne et la variance de chacune des deux classes. Le seuil optimal T est celui qui donne une variance intra-classe minimale (N. Otsu, 1979).Cette variance est donnée par :

  $$\sigma_w^2 = \omega_1(T) \times \sigma_1^2(T) + \omega_2(T) \times \sigma_2^2(T)$$

  ω_1 : Probabilité d'être dans la classe 1

  ω_2 : Probabilité d'être dans la classe 2

  σ_1^2 : Variance de la classe 1 ; σ_2^2 : Variance de la classe 2

- *Méthode ISODATA :*

  Le seuillage par ISODATA (Velasco, 1980) consiste à trouver un seuil en séparant l'histogramme en deux classes itérativement avec la connaissance a priori des valeurs associées à chacune d'elles .On détermine l'intervalle [min, max] des valeurs non nulles de l'histogramme. Après, on fait une estimation des valeurs moyennes initiales en divisant l'intervalle en deux parties équidistantes et en prenant m_1 et m_2 comme la moyenne arithmétique de chaque classe si la densité de probabilité était uniforme. A chaque itération on calcule le seuil T en prenant l'entier le plus proche de la moyenne des deux moyennes : $T = (m_1+m_2)/2$ Puis, on met à jour des moyennes en calculant la moyenne statistique pour chaque classe. On recalcule les seuils et moyennes tant qu'il n'y aura aucun changement jusqu'à la convergence.

- *Méthode de Cheng et Chen*

  La méthode de Cheng et Chen (Cheng et al, 1995) est basée sur le principe d'entropie maximale et la partition floue. Pour le choix du seuil, on considère deux ensembles flous d'*objet* et *fond* dont les fonctions d'appartenance sont définis par :

$$\mu_{objet} = \begin{cases} 1, & x \le a \\ \dfrac{x-c}{a-c}, & a < x < c \\ 0, & x \ge a \end{cases} \qquad \mu_{fond} = \begin{cases} 0, & x \le a \\ \dfrac{x-a}{c-a}, & a < x < c \\ 1, & x \ge c \end{cases}$$

  Dans cette méthode, le seuil de binarisation est choisi comme le niveau de gris dont la fonction d'appartenance=0.5, et donc c'est le centre de l'intervalle [a_{opt}, c_{opt}], tel que a_{opt} et c_{opt} sont les valeurs de a et c maximisant l'entropie de la division.

  Cheng et Chen proposent un algorithme pour trouver les valeurs optimales de a et c correspondants à l'entropie maximale de partitionnement. L'entropie de partitionnement est donnée par :

$$H = -P(objet)\ log\ (P(objet)) - P(fond)\ log\ (P(fond))$$

- **méthodes locales :**
- *Méthode de Niblack*

L'algorithme de Niblack (Niblack, 1986) calcule un seuil local à chaque pixel en glissant une fenêtre rectangulaire sur toute l'image. Le seuil T est calculé en utilisant la moyenne m et l'écart-type σ de tous les pixels dans la fenêtre (voisinage du pixel en question). Ainsi le seuil T est donné par :

$$T = m + k * \sigma$$

Tel que k est un paramètre utilisé pour déterminer le nombre de pixels de contours considérés comme des pixels de l'objet, et prend des valeurs négative

- *Méthode de Bernsen*

  C'est une méthode locale adaptative dont le seuil est calculé pour chaque pixel de l'image (Bernsen, 1986).Ainsi pour chaque pixel (x, y), le seuil est donné par :

$$T(x,y) = \frac{z_{bas} + z_{haut}}{2}$$

Z_{bas} : Niveau de gris le plus bas
Z_{haut} : niveau de gris le plus haut

  Dans une fenêtre carré r × r centré sur le pixel (x, y). Cependant si la mesure de contraste C(x, y)= (Z_{haut} - Z_{bas}) est inférieure à un certain seuil *l*, alors le voisinage consiste en une seule classe: fond ou texte.

- *Méthode de Sauvola*

  Sauvola (Sauvola, 1997) est une modification de la méthode de Niblack qui vise à améliorer les performances de ce dernier dans le traitement des documents avec une texture du fond claire ou une luminosité hétérogène. Pour déterminer le seuil T correspondant à chaque pixel de l'image, il est donc nécessaire de calculer la matrice des moyennes locales de l'image et la matrice des écarts-types locaux de l'image et la binarisation est donnée par la formule (dans ses tests, Sauvola utilise R=128 et k=0.5) :

$$T(x,y) = \left(m(x,y) \times (1 - k \times \left(1 - \frac{s(x,y)}{R}\right)\right)$$

R : valeur maximale de l'écart-type dans un document en niveau de gris
K : paramètre qui prend une valeur positive dans l'intervalle [0.2, 0.5]
s(x,y) : matrice des écarts-types locaux pour chaque pixel de l'image

m(x,y) : matrice des moyennes locales pour chaque pixel de l'image

Selon (Khurram et al, 2009), la méthode de Sauvola se montre plus efficace quand le niveau de gris du texte est proche de 0, et celui du fond de 255. Cependant les résultats sont moins satisfaisants sur des images ou le niveau de gris des pixels du fond et du texte se rapprochent.

1.5.2.2 L'élimination du bruit

La localisation du bruit consiste à identifier les pixels du fond qui n'appartiennent pas à la forme. Cette identification est facile pour des formes simples de bruit telles que *«salt and pepper»*, où la localisation est basée sur une analyse de la taille des composantes connexes. Les composantes connexes de taille inférieure à un seuil déterminé heuristiquement seront supprimées (S.Madhvanath et al, 1999).

Les bruits de forme compliquée peuvent être intersectés avec l'écriture et nécessitent des techniques de suppression complexes, la méthode de Verma (Verma, 2003) par exemple supprime les soulignages qui figurent sous quelques mots grâce à une recherche des lignes horizontales de longueur supérieure à un seuil fixe, pour des soulignements erratiques et inclinés, une suppression manuelle est appliquée. Un autre traitement est proposé par Suen et al (Lam et al, 1995) pour éliminer les lignes de guide, leur méthode peut engendrer une suppression des pixels qui appartiennent au mot. Pour résoudre ce problème, les chercheurs utilisent des opérations morphologiques et topologiques afin de restaurer l'information perdue.

Un autre traitement proposé par M. S. Khorsheed (Khorsheed, 2002) consiste à utiliser la transformée de Fourier pour chercher les pics de haute fréquence qui correspondent au bruit (lignes, grilles). Cette méthode présente l'avantage qu'elle soit applicable sur des images à niveau de gris.

1.5.2.3 Squelettisation

S'effectuant sur une image binaire, le processus de squelettisation transforme l'image du mot en sa représentation en 'fil de fer' appelée squelette en ayant les propriétés suivantes : aussi fin que possible (idéalement, 1 pixel d'épaisseur), préservation de la connexité, et approximation de l'axe médiane de la forme (Cheriet et al, 2007). Le squelette permet de réduire et de compacter la taille de l'image, il facilite également l'extraction de quelques primitives structurelles telles que les points d'embranchement, de croisement, et de fin.

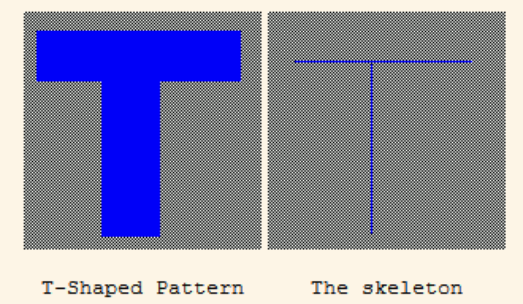


Figure 1.7 : Squelettisation de la lettre T

Il existe, principalement, deux classes d'algorithme de squelettisation : les algorithmes parallèles et les algorithmes séquentiels. Les algorithmes parallèles comme la méthode de ZHANG ET SUEN opèrent sur tous les pixels de l'image simultanément. En revanche, les algorithmes séquentiels comme l'algorithme de HILDITCH examinent les pixels et les transforment selon les résultats obtenus précédemment (H. Boukerma, 2010).

1.5.2.4 Lissage du contour

L'acquisition et la binarisation d'une image produisent généralement des absences de points (trou indésirable), ou des excroissances et des surcharges de points le long du contour du mot. L'étape de lissage permet de remédier à ces déformations, une technique simple consiste à examiner le voisinage d'un pixel et de lui attribuer la valeur 1 (noir) si le nombre de pixel noir dans cette zone est supérieur à un seuil fixe.

Cheriet et al proposent une méthode ou quatre masques simples sont appliqués sur l'image d'un mot, commençant du pixel le plus à droite de la dernière ligne et en parcourant l'image ligne par ligne vers le haut. Cette procédure peut être appliquée plusieurs fois jusqu'à ce qu'aucun changement ne soit achevé (Cheriet et al, 2007).

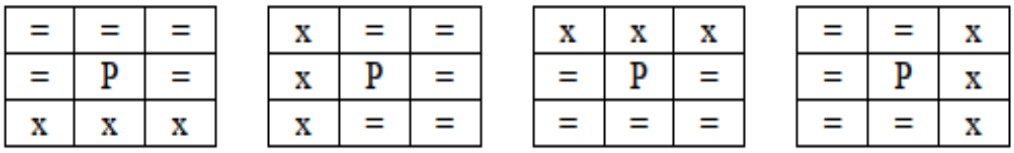


Figure 1.8 : Masques appliqués dans la méthode de Cheriet et al. Source: (Cheriet et al, 2007)

Une autre technique de lissage consiste à appliquer les opérations de la morphologie mathématique (Kanungo et Haralick, 1990), les deux opérations principalement utilisées sont: l'ouverture (une érosion suivie d'une dilatation) et la fermeture (une dilatation suivie d'une érosion). L'ouverture permet d'ouvrir les petits trous et les espaces entre les objets qui se rapprochent suffisamment l'un de l'autre. Tandis que la fermeture permet de remplir les petits trous dans la forme

1.5.2.5 Estimation de la ligne d'appui

La ligne d'appui de l'écriture ou ligne de base porte des informations importantes pour les différentes étapes de la chaîne de reconnaissance. Pour la segmentation, elle guide le processus de détection de points de liaison entre caractères ; outre, elle permet de préciser les positions des diacritiques et de localiser les ascendants et les descendants et donc d'aider le processus de normalisation et d'extraction de primitives. Elle permet également le redressement de l'écriture (Boukerma, 2010).

Il existe de nombreuses méthodes d'extraction de la ligne de base, la plus utilisée est celle basée sur l'histogramme de projection horizontale. Cette méthode part de l'hypothèse que la majorité des pixels se disposent sur la ligne de base, la ligne extraite est donc une ligne droite qui correspond au pic maximal de l'histogramme (voir figure 5). Cette méthode donne de bon résultat pour l'imprimé et le manuscrit de bonne qualité (Lorigo et Govindaraju, 2006), quand les mots sont aussi longs et l'écriture est bien droite.

Figure 1.9: Estimation de la ligne de base par la projection horizontale

1.5.2.6 Seuillage

Le seuillage d'image est l'une des méthodes les plus simples pour obtenir une image comportant uniquement deux valeurs , noir ou blanc à partir d'une image en niveau de gris. Le seuillage remplace un à un les pixels de l'image à l'aide d'une valeur seuil fixée au préalable manuellement ou automatiquement à partir de l'histogramme de projection. Ainsi, si un pixel détient une valeur supérieure au seuil, il prendra la valeur 255 (blanc), et si la valeur est inférieure, il prendra la valeur 0 (noir) (Plamodon et Srihari, 2000).

1.5.2.7 Filtrage

Le principe du filtrage est de modifier la valeur des pixels d'une image, généralement dans le but d'améliorer son aspect. En pratique, il s'agit de créer une nouvelle image en se servant des valeurs des pixels de l'image d'origine dans le but d'éliminer les fluctuations des niveaux de gris présentes à cause du bruit. Pour pallier à ces dégradations, on distingue généralement deux types de filtres (OueldDiaf, 2008) :

- Filtres linéaires :

  Un filtre est dit linéaire si la valeur du nouveau pixel est une combinaison linéaire des valeurs des pixels du voisinage tel que : $NouvelleValeur_{x,y} = \sum_{i,j} A_{i,j} * P_{x+i,y+j}$ avec i,j variant entre -h et +h , la demi taille du voisinage (pour 3x3 h=1, pour 5x5 h=2, …) et $A_{i,j}$ = valeur, entière ou réelle, spécifique au filtre linéaire.

- Filtres non linéaires

Si le filtre ne peut pas être exprimé par une combinaison linéaire, il est appelé " non-linéaire ". Les filtres non-linéaires sont plus complexes à mettre en œuvre que les filtres linéaires. Il existe plusieurs filtres non linéaires tel que :

- *Filtres moyenneur*

Il s'agit d'effectuer une moyenne des niveaux de gris au tour du pixel central à traiter. Ce filtre linéaire qui peut être mis sous la forme d'un masque qu'on déplace sur toute l'image. La taille du masque est un paramètre variable, plus la taille du masque est grande plus le filtrage est important. L'inconvénient de ce filtre est l'introduction de flou dans l'image.

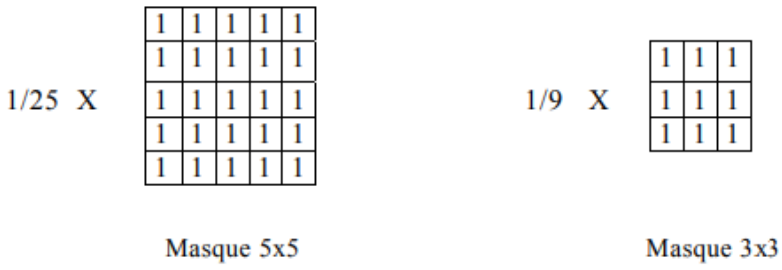


Figure 1.10 : Quelques types de masques utilisés par le filtre moyenneur
Source : (OueldDiaf, 2008)

- *Filtre Gaussien*

L'expression de la gaussienne en deux dimensions est donnée par :

$$H(x,y) = \frac{1}{2.\pi.\sigma^2}.\exp\left[-\frac{(x^2+y^2}{2\sigma^2}\right]$$

σ *est l'écart type*

La largeur du filtre est donnée par son écart-type ce qui permet de régler facilement le degré de filtrage, plus σ est grand, plus on réduit le bruit, mais l'image filtrée sera floue en sortie. Par rapport au filtre moyenneur, il accorde plus d'importance aux pixels voisins du pixel central, mais il dégrade lui aussi les contours.

- *Filtre médian*

Ce filtre non linéaire affecte au pixel central le niveau de gris séparant la population en deux effectifs égaux.

L'avantage de ce filtre est qu'il donne de très bons résultats sur le bruit impulsionnel (poivre et sel).Ses inconvénients sont le fait, qu'il supprime les détails fins et qu'il conduit à des temps de calcul élevés (OueldDiaf, 2008).

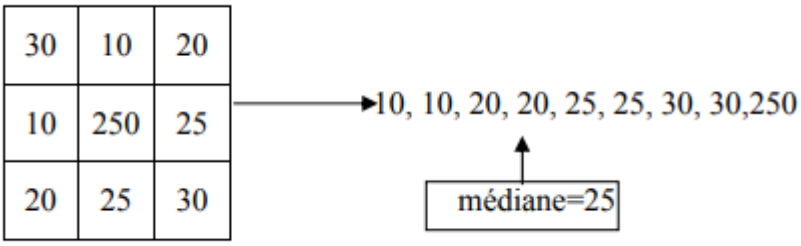


Figure 1.11 : Médiane des niveaux de gris

1.3.2.7.4 Filtrage morphologique

Ce type de filtrage est utilisé pour éliminer des pixels isolés dans des images binaires qui sont considérés comme du bruit. Il met en correspondance chaque pixel et son voisin par une fonction logique (ET, OU, XOR). Parmi les opérateurs morphologiques (Toumazet, 1987) :

- La dilatation

Elle permet d'éliminer les pixels blancs isoles. On effectue le ET logique des huit voisins du pixel considéré. La dilatation élimine les tâches (les trous) blanches dans les zones noires mais ajoute des pixels noirs au contour des objets présents dans l'image.

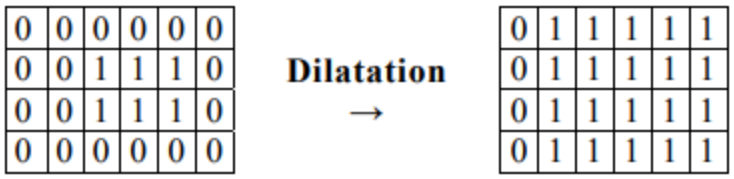


Figure 1.12: Principe de dilatation. Source (OueldDiaf, 2008)

- L'érosion

Elle permet d'éliminer les pixels noirs isolés au milieu des parties blanches de l'image, on effectue le OU logique des huit voisins du pixel considéré. En appliquant une érosion, ces tâches noires peuvent être éliminées mais la taille des objets présents dans l'image diminue car l'érosion enlève des pixels du contour, entraînant parfois une déformation de certains objets.

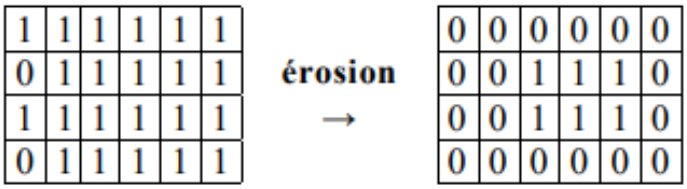


Figure 1.13 : Principe d'érosion. Source : (OueldDiaf, 2008)

1.5.2.8 Normalisation

L'écriture manuscrite a des styles et des tailles différentes ce qui rend complexe le reste des opérations de reconnaissance. Par conséquent, le processus de normalisation est l'une des tâches les plus importantes dans ce processus. Il permet de réduire la variation entre les images du texte et d'ajuster la taille du caractère ou du mot. La normalisation des tailles est couramment utilisée pour réduire la variation de taille et ajuster les tailles des caractères ou des mots à une taille identique (Liu et al, 2010).

- *Correction de l'inclinaison des lignes*

Un défaut d'orientation du document pendant son acquisition, ou une police d'écriture imprécise peut conduire à une inclinaison de la ligne de base. Des traitements pour la rendre horizontale peuvent donc être appliqués, ils incluent généralement deux étapes. La première permet l'estimation de l'angle d'inclinaison globale de la ligne de base. À cet effet, la transformée de Hough et les histogrammes de projection sont les deux méthodes les plus utilisées (Cheriet, 2007) (Hull, 1998). La deuxième étape sert à corriger l'inclinaison par l'application d'une rotation de l'image d'angle θ , cela peut être réalisé par l'application de l'équation suivante (Cheriet, 2007) :

$$\begin{pmatrix} \acute{x} \\ \acute{y} \end{pmatrix} = \begin{pmatrix} \cos(\theta) & \sin(\theta) \\ -\sin(\theta) & \cos(\theta) \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}$$

- *Correction de l'inclinaison des lettres*

Certains scripteurs écrivent leurs lettres de façon inclinée, ce qui forme un angle entre l'axe correspondant à la direction moyenne des caractères et l'axe vertical. Cette inclinaison de l'écriture est également appelée "slant". Les lettres peuvent être inclinées vers la droite ou vers la gauche. L'objectif de cette opération est de transformer le mot de façon à ce que son axe de direction principale devienne vertical, ce qui permet une réduction considérable de la variabilité de l'écriture et une amélioration de la qualité de la segmentation des mots en caractères (Bozinovic et Srihari, 1989) (Grandidier, 2003). Plusieurs approches ont été proposées :

R. M. Bozinovic et Srihari (Bozinovic et Srihari, 1989) utilisent des portions d'écriture proches de la direction verticale pour évaluer l'inclinaison globale du mot. La correction de l'inclinaison est obtenue par l'application de la transformation suivante :

$$\acute{x} = x - y \times \acute{\tan}(\beta - \text{def})$$
$$y \doteq y$$

β : Angle d'inclinaison globale
Def : Paramètre qui spécifie l'inclinaison normale du mot

D'autres travaux utilisent les contours (Britto et al, 2000) (Ding et al, 2000). En parcourant le contour de l'image et en comptant le nombre de fois qu'on se déplace dans les trois directions privilégiées. n_1, n_2 et n_3. L'angle total de l'inclinaison α est donné par l'équation suivante :

$$\alpha = \tan^{-1}\left(\frac{n_1 - n_3}{n_1 + n_2 + n_3}\right)$$

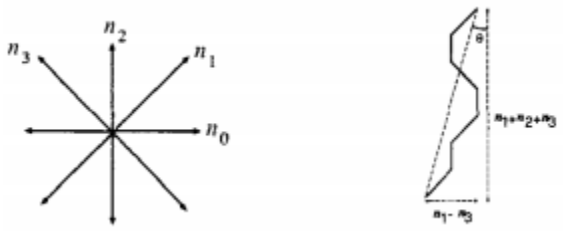


Figure 1.14 : Inclinaison moyenne de l'écriture évaluée sur le contour de l'image .A gauche : n_0, n_1, n_2 et n_3 sont compteurs associés à 4 directions privilégiées. A droite : l'angle d'inclinaison se déduit en parcourant le contour. Source : (Miyake, 2000)

La correction de l'inclinaison se fait par une translation des lignes (Shear Transform) (Cheriet, 2007) :

$$\acute{x} = x - y \times \tan(\alpha)$$
$$\acute{y} = y$$

L'inconvénient de ces méthodes est qu'elles s'appliquent toutes sur la globalité du mot, que ce dernier aie des lettres inclinées ou pas.

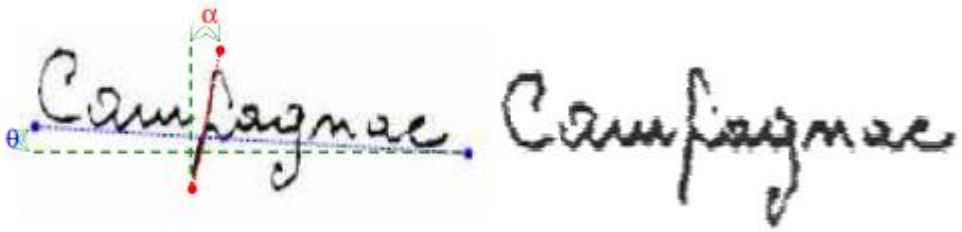


Figure 1.15 : Correction de l'inclinaison des lettres et des lignes .θ est l'angle d'inclinaison de la ligne, α : l'angle d'inclinaison de la lettre. Source : (H. Boukerma, 2010).

- *Normalisation des caractères*

Cette opération vise à normaliser la taille des caractères, en les ramenant tous à une même taille prédéfinie. Dans (Cheriet1, 2007), deux approches de normalisation des caractères sont évaluées, l'approche linéaire et non linéaire.

Dans (Grandidier, 2004), F. Grandidier et *al* remplacent l'étape de la normalisation en taille par une division de l'image de mot en trois zones : zone supérieure, zone inférieure, et zone médiane. Selon les auteurs, cette division permet l'extraction d'un nombre identique de primitives quel que soit la taille de l'image traitée. Les auteurs évaluent d'autres méthodes de division de l'image en fonction du taux de reconnaissance et concluent sur l'importance de cette méthode pour l'amélioration du pouvoir discriminant des primitives.

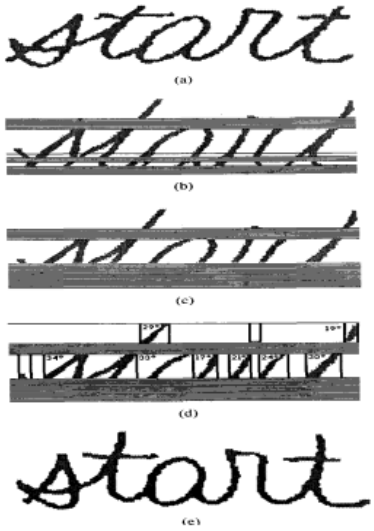


Figure 1.16 : Correction de l'inclinaison des lettres par la méthode proposée dans (Bozinovic, 1989). (a) image originale ; (b) écartement des bandes horizontales suffisamment longues ; (c) les bandes horizontales restantes de petite largeur sont également enlevées ; (d) les bandes de 'image conservées pour l'évaluation de l'angle de l'inclinaison des lettres ; (e) image du mot corrigé. Source :(Bozinovic, 1989)

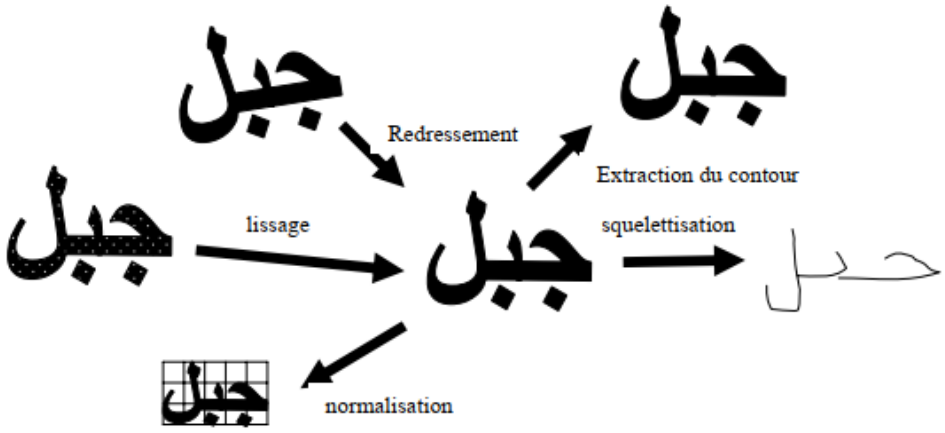


Figure 1.17: Effets de certaines opérations de normalisation. Source (Haitaamar, 2007)

1.5.3 Phase de segmentation

La segmentation permet l'extraction des différentes parties logiques d'une image donnée. Après l'acquisition d'une image, une séparation des blocs de texte et des blocs graphiques est faite, puis à partir des blocs de texte, les lignes des différents paragraphes sont extraites. A partir de ces lignes sont extraits les mots puis les caractères qui forment chaque mot (ou partie du caractère) (Al-Badr et Mahmoud, 1995). Cette phase sera revue en détail dans le chapitre 3.

1.5.4 Phase d'extraction des caractéristiques

L'étape d'extraction de primitives consiste à extraire les informations les plus discriminantes pour la tache de reconnaissance, ces informations doivent permettre de filtrer tous les attributs et préserver les propriétés qui rendent un caractère ou un mot différent d'un autre. Cette étape influe fortement sur les performances du système de reconnaissance. Le choix des primitives à extraire n'est pas aléatoire, selon Trier et al (Trier et al, 1996) il repose sur plusieurs critères tel que la nature de l'écriture à traiter (imprimée/manuscrite, latin, arabe,…), ainsi que sa qualité (variation des fontes et des styles, distorsion…), il dépend aussi de la méthode de classification utilisée. Les caractéristiques des textes manuscrits peuvent être classées dans les catégories suivantes :

- *Caractéristiques structurelles*

  Les caractéristiques structurelles sont les caractéristiques les plus utilisées par les chercheurs (Govindan et Shivaprasad, 1990). Ils illustrent les caractéristiques géométriques et topologiques d'une image textuelle en décrivant leurs propriétés locales et globales (Khorsheed, 2002). Ces caractéristiques dépendent de la catégorie du modèle à classer. Pour un texte en arabe par exemple, les caractéristiques incluent les points et leurs positions, les traits, la largeur et la hauteur du trait, les directions, l'intersection de segments de lignes et des boucles (Khorsheed et Clocksin, 1999), (Amin et Mari, 1989).

- *Caractéristiques statiques*

  Elles analysent la répartition spatiale des pixels en compotant les caractéristiques locales de chaque pixel et en dérivant un ensemble des statistiques majeures du caractère incluent le zonage, où le caractère se divise en zones qui se chevauchent ou non et la distribution de densité des pixels de caractère dans différentes régions est analysée. Mohiuddin et Mao dans (Mohiuddin et Mao, 1994) ont mesuré la direction du contour du caractère en divisant l'image du caractère en zones. Ensuite, des histogrammes des codes de chaîne sont utilisés pour calculer la direction du contour de ces régions.

- *Transformation globale*

  Les techniques de transformation globale ont la capacité de convertir la représentation des pixels en une forme plus compacte. Les techniques les plus courantes pour effectuer cette transformation sont (Toufik et Salih, 2017) :

- Transformée de Fourier
- Transformée de cosinus discret
- Ondelettes
- Transformée de Hough

1.5.5 Phase de classification

La classification est une tâche qui vise à identifier un objet en comparant ses caractéristiques à un ensemble de données répertoriées sous forme de classes grâce à deux opérations majeures qui se succèdent : l'apprentissage puis la reconnaissance et décision. Le modèle d'apprentissage se compose d'un ensemble d'instances qui lui permettent d'acquérir des connaissances et de les organiser en modèles de références. On distingue deux types d'apprentissage : l'apprentissage supervisé par un opérateur qui indique au système les différentes classes et l'apprentissage non supervisé ou le système construit automatiquement son modèle d'apprentissage sans l'intervention d'un opérateur. Une fois que le système a acquis ses connaissances, on passe à la phase de décision ou le système identifiera les différentes formes de test à partir de l'apprentissage réalisé. Les spécificités d'un objet en entrée seront

extraites et comparées avec les caractéristiques des modèles d'apprentissage pour aboutir à son identification.

Parmi les méthodes les plus courantes sur lesquelles se basent les classifieurs, on trouve :

1.5.5.1 Les K plus proches voisins

L'algorithme KNN (K nearest neighbors en anglais) figure parmi les plus simples algorithmes d'apprentissage artificiel. Dans un contexte de classification d'une nouvelle observation x, l'idée fondatrice simple est de faire voter les plus proches voisins de cette observation. La classe de x est déterminée en fonction de la classe majoritaire parmi les k plus proches voisins de l'observation x. La méthode KNN est donc une méthode à base de voisinage, non-paramétrique ; Ceci signifiant que l'algorithme permet de faire une classification sans faire d'hypothèse sur la fonction y=f(x1,x2,...xp)qui relie la variable dépendante aux variables indépendantes (Hastie & al, 2001).

- Quelques règles sur le choix de k :

  Le paramètre k doit être déterminé par l'utilisateur : k∈N. En classification binaire, il est utile de choisir k impair pour éviter les votes égalitaires. Le meilleur choix de k dépend du jeu de donnée. En général, les grandes valeurs de k réduisent l'effet du bruit sur la classification et donc le risque de sur-apprentissage, mais rendent les frontières entre classes moins distinctes (Hechenbichler, 2004). Il convient donc de faire un choix de compromis entre la variabilité associée à une faible valeur de k contre un 'oversmoothing' ou surlissage (i.e gommage des détails) pour une forte valeur de k. Un bon k peut être sélectionné par diverses techniques heuristiques, par exemple, de validation-croisée. Nous choisirons la valeur de k qui minime l'erreur de classification.

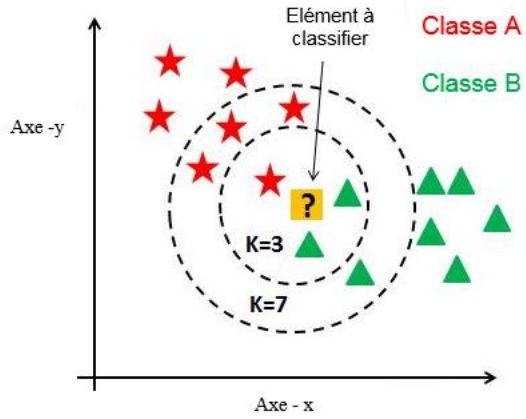


Figure 1.18 : Notion du voisinage de KNN. Exemple avec K = 3 et K = 7

La robustesse et la simplicité de la mise en œuvre sont les principaux avantages du KPPV. Cependant, Son efficacité dépend directement de la pertinence de la base d'apprentissage et notamment de sa densité dans les différentes régions de l'espace de données. Outre, le KPPV est réputé comme classifieur lourd. La recherche des plus proches voisins est coûteuse, cela d'autant plus que la métrique utilisée est complexe, et que la base et la valeur de k sont grandes.

1.5.5.2 Machines à vecteurs de support (SVM)

Les machines à vecteurs de support ou séparateurs à vaste marge (en anglais support vector machine, SVM) sont des classifieurs binaires par apprentissage supervisé destinés à résoudre des problèmes de discrimination ou de régression (prédiction). Elle représente la famille la plus connue des méthodes à noyaux inspirées de la théorie statistique de l'apprentissage de Vladimir Vapnik. Pour décider à quelle classe appartient un échantillon, les SVM utilisent une transformation non linéaire pour redécrire les données d'apprentissage dans un espace de plus grande dimension. L'objectif est alors de déterminer dans le nouvel espace, que l'on nomme espace de redescription, un hyperplan qui permet de séparer les données d'apprentissage de manière optimale, c'est la notion de *marge maximale.*

Les SVM ne permettent de séparer que deux classes, il est nécessaire alors d'en combiner plusieurs pour résoudre les problèmes multi-classes.

1.5.5.3 Modèles cachés de Markov (HMM)

Les HMM sont des modèles stochastiques qui permettent de prendre en compte la variabilité des formes et du bruit qui perturbent la reconnaissance de l'écriture. Ils ont la particularité de pouvoir prendre en compte des séquences de longueurs variables. Ce point est particulièrement important en reconnaissance de l'écriture manuscrite où la longueur des mots peut varier considérablement en fonction des styles d'écriture.

Le modèle est décrit par deux ensembles de probabilités :

- probabilités de passer d'un état à l'autre : probabilités de transition
- probabilités d'observer un symbole pour un état donné : probabilités d'émission

A ceci s'ajoute le choix de l'état initial.

Un HMM est donc défini par :

Un vecteur de probabilités initiales $\Pi = (\pi_i)$

•un vecteur de probabilités de transition (probabilité de passer de l'état i à l'état j) $A = a_{ij}$

•une matrice de probabilités d'émission (probabilité que le symbole b soit observé dans l'état i) $E = (e_i(b))$

La probabilité d'une séquence d'observation x et d'une séquence d'état (chemin) π est donnée par :

$$P(x, \pi) = a_{0\pi_1} \prod_{i=1}^{L} e_{\pi_i}(x_i) a_{\pi_i \pi_{i+1}}$$

1.5.5.4 La logique floue

La logique floue a été introduite par Lotfi A. Zadeh en 1965, elle permet la représentation des connaissances incertaines et imprécises afin de modéliser au mieux le raisonnement humain. Cette représentation est faite à travers la notion des ensembles flous qui permet de définir une appartenance graduelle d'un élément à une classe ou un ensemble, c'est-à-dire appartenir plus

ou moins à cet ensemble, ce dernier est alors qualifié de « flou ». L'appartenance d'un objet x à un sous ensemble flou A est définie par un degré d'appartenance $\mu_A(x)$ entre 0 et 1.

La reconnaissance de l'écriture est un domaine imprécis qui demande une modélisation incertaine. Dès lors, l'utilisation de la logique floue s'est naturellement imposée. On Dans ce qui suit nous nous intéresserons plus particulièrement à l'intégration de la logique floue au niveau de la reconnaissance.

- **Perceptron flou multicouches**

Dans [PAL 92], Sankar K. Pal et Sushmita Mitra introduisent une version floue d'un Perceptron Multi-Couches qui intègre la logique floue à différents niveaux. Ce travail consiste à fuzzifier les sorties du réseau lors de son apprentissage, avant de procéder à leur défuzzification en phase de test. Un mécanisme adéquat est développé aussi pour améliorer l'apprentissage dans le sens où ils changent les valeurs affectées au moment d'inertie (the damping coefficient ou momentum) et au pas d'apprentissage (learning rate) afin d'éviter les minimums locaux et d'accélérer la convergence.

- **K plus proches voisins flou**

Ce type de classifieur tire profit de la logique floue dans le calcul de la distance utilisée pour la sélection des k plus proches voisins. Dans [SIN 99], les auteurs définissent cette mesure floue comme suit :

$$\mu(x) = \left[1 + \left\{\frac{d(x,R)}{F_d}\right\}^{F_e}\right]^{-1.0}$$

Avec : d(x, R) la distance entre la forme inconnue x et l'échantillon R. F_d et F_e deux constantes positives utilisées pour contrôler le floue dans le calcul précédant.

1.5.5.5 Réseaux de neurones artificiels

Les réseaux de neurones (RN) formels sont des systèmes de traitement de l'information dont la structure s'inspire de celle du système nerveux. C'est un réseau fortement connecté de

processeurs élémentaire (*neurones*) fonctionnant en parallèle et disposant en *couches*. Tous les neurones d'une même couche ont la même *fonction d'activation*. L'apprentissage d'un RNA se fait le plus souvent de façon itérative, par rétro propagation du gradient d'erreur, cet algorithme d'apprentissage très efficace donne un essor important à ce classifieur.

Le premier réseau de neurones artificiel apparaît en 1958, grâce aux travaux de Rosenblatt qui conçoit le perceptron. Ce dernier est inspiré du système visuel (en terme d'architecture neurobiologique) et possède une couche de neurones d'entrée une couche de sortie ("décisionnelle"). Ce réseau parvient à apprendre, à identifier des formes simples et à calculer certaines fonctions logiques. Les chercheurs se sont tournés vers le domaine de l'intelligence artificielle par la suite et le projet de Rosenblatt connait un abandon financier. Il faudra attendre le début des années 80 pour que l'intérêt pour ce domaine soit de nouveau présent. En effet, Hopfield démontre en 1982 tout l'intérêt d'utiliser les réseaux récurrents « feed-back »pour la modélisation des processus. Les réseaux récurrents constituent alors la deuxième grande classe de réseaux de neurones, avec les réseaux type perceptron « feed-forward ».

- **Neurone formel**

En général, un neurone formel est un élément de traitement possédant *n* entrées (x_1, x_2, …, x_i, …, x_n) qui sont les entrées externes ou les sorties des autres neurones) et une ou plusieurs sorties. Son traitement consiste à effectuer à sa sortie y_i le résultat d'une fonction de seuillage *f* (dite aussi la fonction d'activation) de la somme pondérée.

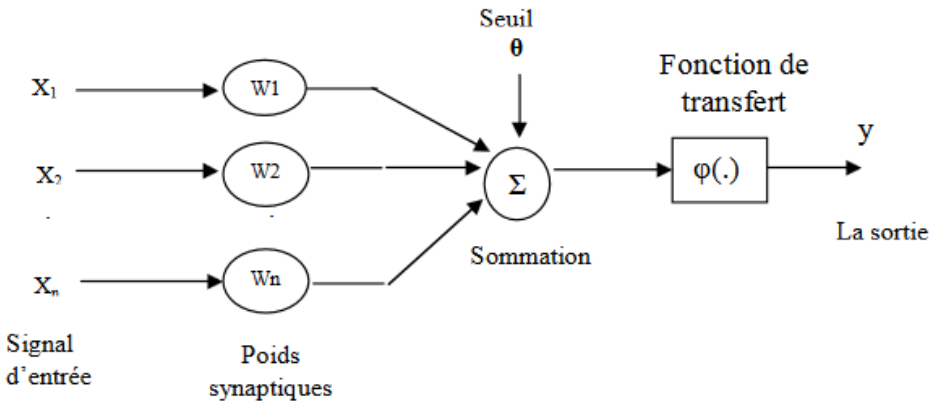


Figure 1.19 : Modèle d'un neurone formel

Avec :

- Les x_i sont les entrées du réseau
- S est le potentiel d'activation
- Les w_i représentent les poids synaptiques
- Y_i la sortie du réseau tel que : $y = f(s)$ et $s = \sum_{i=0}^{n} w_i . x_i + b$

- **Couches, connexions et fonctions d'activation**

Un réseau de neurones est constitué de cellules (ou neurones), connectée entre elles par des liaisons affectées de poids. Ces liaisons permettent à chaque cellule de disposer d'un canal pour envoyer et recevoir des signaux en provenance d'autres cellules du réseau. Chacune de ces connexions reçoit un poids (une pondération), qui détermine sont impact sur les cellules qu'elle connecte. Chaque cellule dispose ainsi d'une entrée, qui lui permet de recevoir de l'information d'autres cellules, mais aussi de ce que l'on appelle une fonction d'activation, qui est dans les cas les plus simple, une simple identité du résultat obtenu par l'entrée et enfin une sortie. Ainsi, pour un réseau de neurones avec N cellules dans la première couche, notées C(1),...,C(N) , et N poids affectées aux liaisons et notés w(1),...,w(N) l'entrée d'une cellule de la seconde couche sera généralement une somme pondérée des valeur de sortie des neurones précédents :

X = w(1)*C(1) + w(2)*C(2) + w(3)*C(3) + ... + w(N)*C(N)

Le choix d'une fonction d'activation se révèle être un élément constitutif important des réseaux de neurones. Ainsi, l'identité n'est pas toujours suffisante, bien au contraire, et le plus souvent des fonctions non linéaires et plus évoluées seront nécessaires. A titre illustratif voici quelques fonctions couramment utilisées comme fonctions d'activation :

- La fonction logistique: Y = F(X) = 1/(1 + exp(-d*X))
- La tangente hyperbolique : Y = 2 / (1 + exp(-2 * X)) – 1
- La fonction Gaussienne : Y = exp(-(X^2)/2)
- Une fonction à seuil : Y = 0 si X<0 et Y=1 si X>0

L'exemple le plus simple de réseau de neurones est souvent donné par le perceptron multicouches (qui est un cas particulier de réseau de neurones). Dans un perceptron, plusieurs couches contenant des neurones sont connectées entre elles de l'entrée vers la sortie.

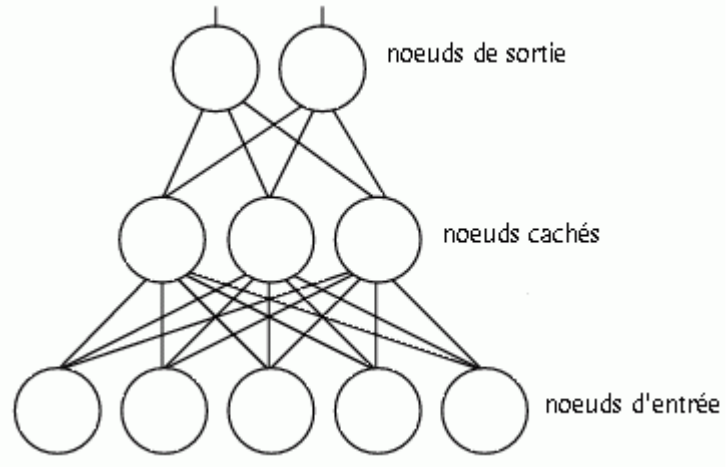


Figure 1.20 : schémas type d'un perceptron à trois couches

- Les nœuds d'entrée : La première couche est appelée couche d'entrée. Elle recevra les données source que l'on veut utiliser pour l'analyse. Dans le cas de l'aide au diagnostic

médical, cette couche recevra les symptômes. Sa taille est donc directement déterminée par le nombre de variables d'entrées.

- Les nœuds cachés : La seconde couche est une couche cachée, en ce sens qu'elle n'a qu'une utilité intrinsèque pour le réseau de neurones et n'a pas de contact direct avec l'extérieur. Les fonctions d'activations sont en général non linéaires sur cette couche mais il n'y a pas de règle à respecter. Le choix de sa taille n'est pas implicite et doit être ajusté. En général, on peut commencer par une taille moyenne des couches d'entrée et de sortie mais ce n'est pas toujours le meilleur choix. Il sera souvent préférable pour obtenir de bon résultats, d'essayer le plus de tailles possibles.

- Les nœuds de sortie : La troisième couche est appelée couche de sortie. Elle donne le résultat obtenu après compilation par le réseau des données entrée dans la première couche. Dans le cas de l'aide au diagnostic médical, cette couche donne le diagnostic. Sa taille est directement déterminée par le nombre de variables qu'on veut en sortie.

- **Types de réseaux de neurones**

Les types de RNA sont très nombreux, ils se distinguent généralement par la fonction d’activation des neurones. Nous présentons rapidement les trois types les plus populaires :

1. Les Perceptrons Multicouches : le perceptron multicouche est sans doute le plus simple et le plus connu des réseaux de neurones. La structure est relativement simple : une couche d'entrée, une couche de sortie et une ou plusieurs couches cachées. Chaque neurone n'est relié qu'aux neurones des couches précédentes, mais à tous les neurones de la couche précédente. La fonction d'activation utilisée est en générale une somme pondérée.

2. Les Réseaux de Hopfield : ces réseaux sont des réseaux récursifs, un peu plus complexes que les perceptrons multicouches. Chaque cellule est connectée à toutes les autres et les changements de valeurs de cellules s'enchainent en cascade

jusqu'à un état stable. Ces réseaux sont bien adaptés à la reconnaissance de formes.

3. Les Réseaux de Kohonen: ils décrivent dont on parle généralement sans les distinguer, décrivent en fait trois familles de réseaux de neurones :

– VQ :Vector Quantization. (apprentissage non supervisé) : Introduite par Grossberg (1976), la quantification vectorielle est une méthode généralement qualifiée d'estimateur de densité non supervisé. Elle permet de retrouver des groupes sur un ensemble de données, de façon relativement similaire à un k-means algorithm que l'on préfèrera d'ailleurs généralement à un VQ si la simplicité d'implémentation n'est pas un élément majeur de la résolution du problème.

– SOM :Self Organizing Map. (apprentissage non supervisé) : Les SOM sont issus des travaux de Fausett (1994) et Kohonen (1995). Ces réseaux sont très utilisés pour l'analyse de données. Ils permettent de cartographier en deux dimensions et de distinguer des groupes dans des ensembles de données. Les SOM sont encore largement utilisés mais les scientifiques leur préfèrent maintenant les LVQ.

– LVQ: Learning Vector Quantization. (apprentissage supervisé) : Les réseaux utilisant la méthode LVQ ont été proposés par Kohonen (1988). Des trois types de réseaux présentés ici, la LVQ est la seule méthode qui soit réellement adaptée à la classification de données par "recherche du plus proche voisin".

1.5.6 Phase de post-traitement

L’objectif du post-traitement est l’amélioration du taux de reconnaissance des mots (par opposition au taux de reconnaissance du caractère). Comme la classification peut aboutir à plusieurs hypothèses possibles, le post-traitement a pour objet de vérifier si la réponse est correcte en utilisant des niveaux d’informations plus élevés (niveau syntaxique, niveau lexical, niveau sémantique…)

.Dans (Carbonnel), on trouve une étude intéressante du post-traitement lexical. L'auteur distingue quatre niveaux de connaissances linguistiques :

1. Un *niveau pragmatique* qui correspond à l'analyse du langage associé à son utilisation et à l'action. Il a une valeur concrète et pratique.

2. Un *niveau sémantique* qui analyse des phrases ou des énoncés du point de vue du sens comme par exemple les relations action/acteur, objet/processus, mais aussi la synonymie, la polysémie, etc.

3. Un *niveau syntaxique* qui concerne l'analyse des phrases du point de vue grammatical, des règles régissant les relations entre les mots dans une phrase.

4. Un *niveau lexical* qui analyse les mots, cette analyse est liée à la notion de lexique qui désigne l'ensemble des mots d'une langue, ou l'ensemble des mots appartenant à un certain vocabulaire.

Pratiquement les connaissances linguistiques les plus utilisées dans les procédures de post-traitement sont syntaxiques et lexicales.

Dans un post-traitement syntaxique, un ensemble de règles grammaticales traduisant le contexte syntaxique du document, elles sont généralement appliqués afin de confirmer ou non la séquence de mots proposés. Ce type de post-traitement est largement utilisé quand les règles syntaxiques ne sont pas trop nombreuses (Lam, 1995).

Le rôle du post-traitement lexical est d'ordonner les mots du lexique par rapport à leur ressemblance aux propositions issues d'une approche de reconnaissance analytique (ou pseudo analytique). Les connaissances lexicales peuvent être modélisées par des modèles de langage de N-grammes ou par des dictionnaires (lexiques) (Menasri, 2008).

1.6 Problèmes liés à l'OCR

La tâche d'un OCR reste très compliquée à réaliser, divers problèmes compliquent le processus de reconnaissance, parmi ces problèmes nous pouvons citer :

- **La qualité du document** : un support de document télécopié ou photocopié plusieurs fois est plus difficile à traiter. L'écriture peut devenir plus mince ou au contraire plus épaisse, elle peut aussi être dégradée avec des parties du texte qui manque ou de tâches qui apparaissent, ainsi les caractères deviennent plus difficiles à reconnaitre.

- **La disposition du texte** : la présentation du texte peut subir deux types de contraintes : externes (écriture pré casée, zonée, guidée ou générale) ; et internes provenant des habitudes propres à chaque scripteur (écriture détachée, groupée, script (bâton), purement cursive ou mixte). L'écriture cursive nécessite plus d'efforts pour définir les limites entre les lettres.
- **L'impression**

  Un document composé est de meilleure qualité qu'un document dactylographié qui, à son tour, est plus clair qu'un texte imprimé. Une imprimante à jet d'encre peut introduire des tâches d'encre et un étalement des caractères, des lignes ou des fonds supplémentaires peuvent être générés par une imprimante laser…
- **Nombre de scripteurs**

  Dans le cas d'une écriture manuscrite, la difficulté varie selon le nombre de scripteurs. Dans le cas d'un multi-scripteur, le système doit s'adapter à l'écriture, tandis que pendant l'utilisation d'un omni-scripteur, le système doit être capable de généraliser son apprentissage à n'importe quel type d'écriture. Le mono-scripteur reste le moins compliqué des trois catégories existantes car le système prend un seul style d'écriture en compte.

- **La discrimination de la forme**

  Le graphisme d'un caractère change selon le style de la fonte utilisée. Le corps et la graisse sur le document ainsi que la similarité entre les caractères de la langue arabe et le nombre de formes d'un seul caractère qui peut être plus élevé que le nombre de styles d'écritures existants compliquent d'avantage la tache de reconnaissance.

- **L'acquisition**

  La numérisation en temps réel introduit souvent des bruits inutiles dans l'image. Dans le cas hors-ligne la qualité du texte numérisé est compromise entre les variations de la position (inclinaison, translation, rétrécissement...), la propreté de la vitre du dispositif de numérisation et sa résolution.

- **Les variations des dimensions des caractères sur un support**

  Un pitch (character per inch) de 10 implique des caractères plus grands aussi bien en largeur qu'en hauteur que ceux d'un pitch de 12.

- **La taille du vocabulaire utilisé**

  On peut retrouver des applications à vocabulaire limité (< 100 mots) et celles à vocabulaire très étendu (> 10 000 mots). Dans le cas d'un petit vocabulaire, la complexité est moindre, car la réduction du nombre limite l'encombrement mémoire et favorise l'utilisation de méthodes de reconnaissance directes et rapides

En plus de ces problèmes un système OCR devrait être capable de distinguer entre un texte et une figure, de reconnaître les caractères ligaturés et d'être indépendant des variations de l'espace aussi bien inter-mots que de l'interligne.

1.7 Domaine d'application des OCRs

Les OCRs sont de nos jours utilisés dans plusieurs taches de la vie quotidienne telles que :

- L'aide à la lecture pour les non-voyants : Les systèmes de reconnaissance associés à des synthétiseurs vocaux permettent la compréhension de documents et livres pour les aveugles.

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

1.8 Conclusion

Nous avons présenté dans ce chapitre les différents modules qui peuvent constitué un OCR, nous avons aussi vu en détail les techniques les plus utilisées dans la littérature pour réaliser les opérations nécessaires à la reconnaissance de l’écriture en général.

Chapitre 2

Reconnaissance de l'écriture arabe manuscrite

2.1 Présentation de la langue arabe

La langue arabe, langue officielle dans près de 25 pays, fait partie de la grande famille des langues sémitiques au même titre que l'akkadien, le phénicien, l'hébreu, etc. Elle se classe dans le sous-groupe qu'on appelle les sémitique méridional dont l'espace géographique fut l'Arabie, nom tiré de celui des tribus arabes qui la peuplaient (Baccouche, 1998). Elle est aussi utilisée comme vecteur de transmission religieux pour tous les musulmans au nombre de 1.8 milliards à travers le globe en 2015 soit 24% de la population mondiale. Cette langue qui est l'ascendante directe de l'écriture araméenne ancienne qui est un rejeton de l'alphabet phénicien entre linguistiquement dans l'histoire, à la fin du VI^{ème} siècle, d'abord avec les vers de ses premiers poètes connus puis avec le Coran qui va déterminer son destin. Et elle entre dans l'histoire comme une langue commune aux tribus (Roman, 1990).

2.2 Variantes de la langue arabe

Le mot arabe est un terme qui regroupe les nombreuses variétés existantes de cette langue. A l'époque préislamique (les débuts de la langue arabe) l'arabe possédait des dialectes différents l'un de l'autre parmi eux on peut citer les dialectes des tributs de Qahtane, Adnane et Himyar (Farghaly et Shaalan, 2009). Le nombre de dialectes existants n'a jamais été précis, de ce fait il existe plusieurs classifications pour distinguer ces variétés, par exemple : en 1959 le grand linguiste américain Ferguson définit deux variétés : la variété élevée reconnue par l'arabe classique et la variété basse utilisée dans la communication quotidienne des arabophones (les dialectes) (Ferguson, 1959). La classification du linguiste El-Said Badawi dans (Badawi, 1973) propose quant à elle cinq variantes : (1) L'arabe classique, (2) L'arabe standard moderne,

(3) L'arabe utilisé par les personnes instruites, (4) L'arabe utilisé par des personnes semi-instruites, et (5) L'arabe utilisé par des personnes analphabètes.

Suite aux différentes classifications proposées, la communauté de recherche se met d'accord sur une seule classification qui fait ressortir trois variétés de l'arabe : (1) L'arabe classique, (2) L'arabe standard, (3) L'arabe dialectal.

2.2.1 L'arabe classique

Langue sacrée de l'islam, la naissance de cette dernière est marquée par la révélation coranique (le Coran a été révélé au prophète Mahomet par Dieu à travers l'archange Gabriel, en arabe classique). Cette époque était appelée par certains linguistes et historiens, la première métamorphose de la langue arabe. La langue arabe est devenue une langue officielle du monde musulman en 685 , elle s'est par ailleurs développée au fil du temps à travers son utilisation dans le développement des sciences et techniques, et dans la traduction des manuscrits grecs, de philosophie et de sciences, entre le VIIIe et le Xe siècle , ce qui lui a permis de connaitre une deuxième métamorphose qui a fait d'elle une langue de civilisation qui a duré plus de quatorze siècles et qui est arrivée jusqu'en occident .Cette variété prestigieuse est maitrisée par peu d'arabes dans le monde , 120 millions de personnes seulement la connaissent comme langue seconde .

2.2.2 L'arabe standard (MSA)

Cette variété de la langue arabe est fondée syntaxiquement, morphologiquement et phonologiquement sur l'arabe classique avec un lexique plus récent (El Kassas, 2005), elle est souvent utilisée chez les locuteurs arabes instruits dans les situations formelles. L'arabe standard constitue la langue officielle et écrite de tous les pays arabophones sans être la langue maternelle des populations de ces pays, elle est utilisée dans la plupart des écrits administratifs, médiatiques, scientifiques, techniques, littéraires ainsi que dans la majorité des articles de presse et les journaux télévisés.

Le MSA possède par ailleurs des variations régionales dues à plusieurs facteurs parmi lesquels nous citons :

1) la création de nouveaux vocabulaires
2) l'influence de l'histoire coloniale propre aux régions sur la syntaxe et la stylistique du MSA

2.2.3 L'arabe dialectal

Généralement appelée « āmmiyya » "langue commune" ou « dārija » "langue courante", cette forme de la langue arabe est utilisée dans les communications et activités quotidiennes (Johnstone et Salih, 1969). Elle varie non seulement d'un territoire arabe à un autre, mais aussi d'une région à une autre au sein du même territoire (Saâdane, 2013). Ainsi, presque tous les pays arabes ont leurs propres dialectes qui sont plus ou moins différents les uns des autres au sein du même pays, et plus naturellement, de ceux des autres pays. Ces différences dépendent considérablement de l'histoire de chaque pays et de son emplacement géographique. La classification des différents dialectes arabes a longtemps intéressé plusieurs chercheurs et plusieurs classifications ont été proposées au cours des années selon certains critères comme le critère géographique et le critère social (Les dialectes des pays du Golf, les dialectes irakiens, les dialectes égyptiens, les dialectes maghrébins...).

2.3 Caractéristiques de la langue arabe

L'Arabe se distingue des autres écritures à différents points de vue, parmi ces différenciations on retrouve :

- **L'alphabet**

  L'alphabet arabe comporte 28 lettres fondamentales (29 en y ajoutant la hamza qui peut prendre le rôle d'un diacritique ou d'une lettre à part entière). Comme il est montré dans le *tableau 1*, on distingue 22 lettres de l'alphabet ayant quatre formes d'écriture selon leur position dans le mot , les six lettres restantes (ا, د, ذ, ر, ز, و) ont quant à elles juste deux formes d'apparitions et ne peuvent pas être liées à la lettre suivante, leur forme de début est simplement leur forme isolée et leur forme de milieu est exactement celle de fin.

Plus de la moitié des lettres arabes sont composées d'un corps principal et des composantes secondaire: par exemples la lettre Beh (ب) et possède un point au-dessus de sa corps principal, Teh (ت) possède deux points et Kef (ك) possède un zigzag ci-joint avec le corps principal.

Caractère.	Au début	Au milieu	Fin	Isolé	Car.	Au début	Au milieu	Fin	Isolé
alif	أ إ	ـأـ△	ـأ	إ أ	dhad	ضـ	ـضـ	ـض	ض
ba	بـ	ـبـ	ـب	ب	tad	طـ	ـطـ	ـط	ط
tâ	تـ	ـتـ	ـت	ت	thad	ظـ	ـظـ	ـظ	ظ
thâ	ثـ	ـثـ	ـث	ث	ayn	عـ	ـعـ	ـع	ع
jim	جـ	ـجـ	ـج	ج	ghayn	غـ	ـغـ	ـغ	غ
ha	حـ	ـحـ	ـح	ح	fa	فـ	ـفـ	ـف	ف
kha	خـ	ـخـ	ـخ	خ	kaf	قـ	ـقـ	ـق	ق
del	ـد	ـد△	ـد	د	kef	كـ	ـكـ	ـك	ك
dhel	ذ	ـذ△	ـذ	ذ	lam	لـ	ـلـ	ـل	ل
ra	ر	ـر△	ـر	ر	mim	مـ	ـمـ	ـم	م
zei	ز	ـز△	ـز	ز	noun	نـ	ـنـ	ـن	ن
sin	سـ	ـسـ	ـس	س	ha	هـ	ـهـ	ـه	ه
shin	شـ	ـشـ	ـش	ش	wew	و	ـو△	ـو	و
sad	صـ	ـصـ	ـص	ص	ya	يـ	ـيـ	ـي	ي

Tableau 2.1 : Lettres de l'alphabet arabe avec leurs variations de forme
(Le triangle montre les lettres qui ne peuvent pas être attachés à leur successeur)

Les points jouent un rôle important dans les caractères arabes par ce que la forme des certains caractères est similaire, mais la différence se pose avec la position et le nombre des points, cela peut se produire soit au-dessus ou en dessous des caractères

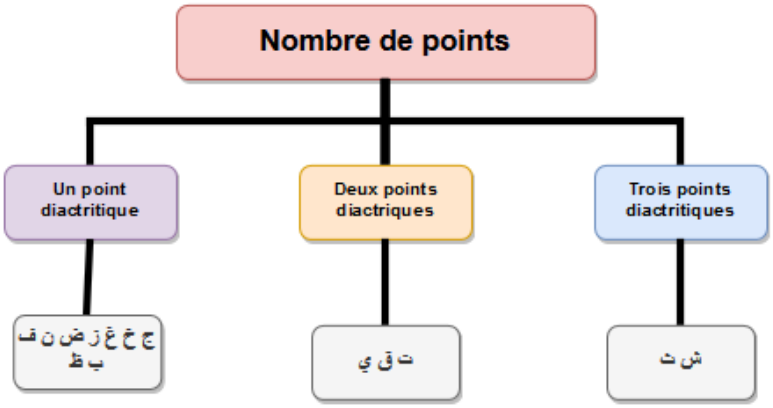


Figure 2.1: Classification des lettres de l'alphabet arabe ayant des points diacritiques

- **Sens de l'écriture**

Contrairement à une grande majorité de langues, la langue arabe est cursive et s'écrit en commençant de la gauche vers la droite dans le cas de l'imprimé et du manuscrit avec un sens du tracé respecté.

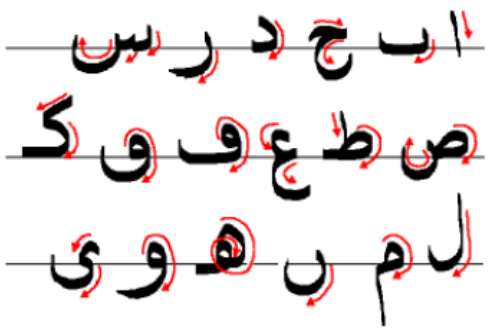


Figure 2.2: Sens de l'écriture des caractères arabes

- **Les voyelles**

En arabe, les voyelles ne sont pas des lettres, mais des signes diacritiques associés aux lettres sur lesquelles ils s'appliquent. Un seul mot peut avoir plusieurs voyellations

possibles et par conséquent une absence de voyelles de ce dernier peut porter à confusion. Les voyelles bréves :

- ***Fatha*** (ـَ) : elle surmonte la consonne et se prononce comme un « a » en Français

- ***Damma*** (ـُ) : elle surmonte la consonne et se prononce comme un « u » en Français

- ***Kasra*** (ـِ) : elle se note au-dessous de la consonne et se prononce comme un « i » en Français

Lorsque la consonne n'a aucune voyelle, on marquera une absence de voyelle représentée en arabe par une voyelle muette (Sukun ـْ)

Les voyelles longues :

Les voyelles longues sont des lettres prolongées, elles sont formées par une des voyelles brèves suivis d'une des lettres correspondantes suivantes :*Alef, waw, yeh* (ي, و,ا)

Voyelles longues	ـَا	ـُو	ـِي

Tableau 2.2: Voyelles longues arabes

Autres signes diacritiques :

Šadda : est un signe qui peut être placé au-dessus d'une consonne mais qui ne peut pas être à la position initiale du mot. La consonne surmontée de ce signe est considérée comme deux consonnes identiques géminées, la première avec une voyelle brève :Fatha, Damma ou Kasra dite *motaharika* ,et la deuxième sans voyelle avec sukun . Par exemple *Mada* ~ مَدَّ (donner) est analysé comme *Madad* مَدَدْ.

Tanwin : Il est considéré comme étant le double d'une voyelle brève. Ajouté seulement à la fin des mots indéterminés, ce signe n'apparait jamais sur un mot contenant l'article de détermination AL (ال)

Voyelle courte	Transcription	Nom
ـَ	A	Fatha
ـُ	U	Damma
ـِ	I	Kasra
ـْ	E	Sukun
ـّ	Doublement	Shadda
ـً	Aa	Fathatan
ـٌ	Uu	Dammatan
ـٍ	Ii	Kasratan

Tableau 2.3 : Voyelles courtes de la langue arabe

- **La cursivité**

L'arabe est une langue cursive du fait que la plupart de ses lettres s'attachent entre elles lors de leur écriture.

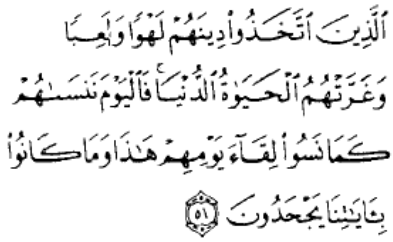


Figure 2.3 : Cursivité de la langue arabe

- **Le mot graphique arabe et ses constituants**

Un mot graphique est représenté par une séquence graphique de caractères, séparée par des délimiteurs comme le blanc ou les signes de ponctuation, ce dernier peut être simple s'il est formé par la concaténation d'une base avec d'éventuels affixes (préfixes et suffixes) nécessaires pour faire de lui un mot attesté ou complexe si ce dernier est formé par la concaténation d'un mot simple et d'un ensemble de clitiques (proclitiques et enclitiques).

Mot graphique arabe simple = Préfixes + Base + Suffixes
Mot graphique arabe complexe= Proclitiques + mot simple + Enclitiques

- *Les clitiques* : Les clitiques ou enclinomènes sont les unités lexicales qui ne comportent qu'une consonne et une voyelle brève. Ils ne peuvent pas être écrits isolément entre deux blancs et sont donc rattachés à la séquence qui suit, pour ne former avec elle qu'un seul mot.
- *La base* : La base représente la partie stable du mot après avoir retiré ses affixes sans former à elle seule un mot existant dans la langue arabe. Elle se compose de deux parties qui sont :
- *la racine* : constituée uniquement de consonnes, ordonnées de façon stricte.
- *le schème* : constitué de voyelles longues ou brèves et parfois aussi de consonnes

- **Typographie arabe**
- *Notion de fonte* : Une police (fonte) est un ensemble de caractères d'une même famille, d'une même graisse et pour un corps donné. La graisse ou chasse représente l'espace qu'occupe un caractère d'imprimerie, elle dépend du dessin, du style et de la grosseur du caractère, le corps quant à lui désigne la hauteur d'un caractère typographique comprenant le blanc de la séparation horizontale avec la ligne au-dessus, sa dimension s'exprime en points (Point est équivalent à 0.376 mm). Ces caractéristiques sont normalisées dans l'imprimerie, tant au niveau du symbole.

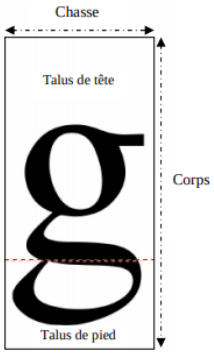


Figure 2.4: Chasse et corps d'un caractère latin

- *Styles d'écriture calligraphique arabe :*

  Les bases et les principes de l'écriture arabe ont été définis par Ibn Moqlah (886 – 940de l'hégire), qui a défini six styles d'écriture: *Kufi*, *Thuluth*, *Naskh*, *Ruq'a*, *Ta'līq* et *Diwani*.

- **Le Diwani** (ديواني): d'origine Turque, et ayant connu son summum durant le règne Ottoman il fut utilisé par les cours Ottmanes (*Diwan*) pour écrire des documents officiels, Certaines de ses variantes sont encore en usage aujourd'hui (par exemple, des documents écrits à la main par certains responsables religieux ('duls).
- **Le Koufique** (كوفي) : ce style d'écriture (anciennement appelé "Hiri" et issu de l'écriture syriaque) tire son nom de la ville iraquienne de Koufa. Il représente l'un des scripts arabes les plus anciens et les plus connus. Il se caractérise par ses formes géométriques décoratives et prononcées, bien adaptées aux dessins architecturaux. Le style s'est développé avec le début de l'Islam (vieux *Kufi*) pour satisfaire le besoin pour les musulmans de codifier le Coran.
- **Le Naskh** (نسخ): un des styles les plus anciens et les plus clairs de tous , respectant le caractère esthétique avec des lettres nettement distinguées qui facilitent la lecture et la prononciation. de l'écriture arabe, le style classique naskhi rassemble souplesse du style perse et harmonie de l'écriture koufique, il peut être écrit à de petites tailles ce qui convient à la production de textes plus longs utilisés dans les livres destinés à la population en général, en particulier le Coran.
- **Ruq'a** (رقعة) : un style manuscrit encore couramment utilisé dans les pays Arabes et reconnaissable par ses caractères audacieux écrits au-dessus de la ligne d'écriture. Conçu pour être utilisé dans l'éducation pour l'écriture quotidienne, il a été adopté dans les administrations de l'Empire Ottoman. Une de ses caractéristiques est que les calligraphes l'ont gardé et n'en ont pas dérivé des variations.
- **Ta'līq** (تعليق) : (suspension) est un beau script caractérisé par la précision et l'étirement de ses lettres, sa clarté et son manque de complexité. Conçu pour la langue persane, jusqu'à être remplacé par *Nasta'līq*.
- **Maghribi** (مغربي) : utilisé par le passé dans le monde occidental islamique (Andalousie), et encore aujourd'hui en Afrique du Nord. Utilisé pour écrire le Coran

ainsi que d'autres manuscrits scientifiques, juridiques et religieux. *Rabat*, une version *Mabsut* de ce script, est largement utilisé dans certaines imprimeries officielles au Maroc.

– **Thuluth** (ثلث) : (le tiers.) Reconnaissable par le fait que les lettres et les mots sont très entrelacés dans sa forme complexe. Peut-être le style le plus difficile à écrire (exigeant une bonne dose d'exercice), à la fois en matière de ses lettres et en matière de sa structure et sa composition.

– **Le Mohaqqaq** (محقق): était à l'origine une écriture dont les lettres étaient moins angulaires que le Koufi, avec des ligatures amplement séparées ; l'ensemble était « produit avec méticulosité » comme son nom le signale. Cette calligraphie arabe acquit une certaine rondeur qui la rendit plus facile à tracer et elle devint l'écriture privilégiée des scribes

Figure 2.5: différents styles d'écriture Arabe

- **La composition du lexique arabe**

La grammaire traditionnelle arabe ne connaît que trois sous-ensembles : verbes, noms et particules (Bourezg, 2017).

– *Les verbes* : Un verbe est une entité exprimant un sens dépendant du temps. La majorité des verbes arabes sont formés sur des radicaux de trois consonnes (Kelaiaia, 2010) tel est le cas du verbe كَتَبَ (écrire) et éventuellement quatre consonnes comme le verbe دَحْرَجَ (faire glisser). C'est un élément fondamental auquel se rattachent directement ou indirectement les divers mots qui constituent l'ensemble. Chaque verbe est donc l'origine d'une famille de mots. La conjugaison des verbes dépend de plusieurs facteurs:

- Le temps (accompli, inaccompli).
- Le nombre du sujet (singulier, duel, pluriel).
- Le genre du sujet (masculin, féminin).
- La personne (première, deuxième et troisième)
- Le mode (actif, passif).

– *Les noms* : L'élément désignant un être, un objet ou un état qui exprime un sens indépendant du temps. La fonction du nom est sa relation avec un mot ou une expression de la phrase, elle change avec le changement de cette relation sans perdre son sens linguistique. Les substantifs arabes sont de deux catégories, ceux qui sont dérivés de la racine verbale et ceux qui ne les sont pas comme les noms propres et les noms communs (Maraoui et al, 2017).

– *Les particules* : Ce sont principalement les mots outils comme les conjonctions de coordination et de subordination. Elles servent à situer les événements et les objets par rapport au temps et l'espace, et permettent un enchaînement cohérent du texte. La particule est tout ce qui n'est ni un verbe ni un nom et qui n'a de sens que dans une phrase construite par exemple : إلى, على, من. Les particules peuvent avoir des préfixes et suffixes ce qui rajoute une complexité quant à leur identification (Douzidia, 2005).

Elles sont classées selon leur sémantique et leur fonction dans la phrase en plusieurs types (introduction, explication, conséquence, …). Elles jouent un rôle important dans l'interprétation de la phrase (Bourezg, 2017).

2.4 Problèmes liés à la reconnaissance de l'écriture arabe

L'arabe, comme toutes les langues naturelles, est caractérisée par un ensemble difficultés et des problèmes lors de son traitement automatique, parmi ces problèmes on retrouve :

2.4.1 L'absence de voyellation

Dans la langue arabe, un mot sans voyellation engendre une combinaison de mots avec un sens et des classes grammaticales différentes, dont le nombre d'éléments diffère en fonction de leur existence dans le vocabulaire ou pas.

Chaque consonne faisant partie d'un mot peut prendre l'une des sept voyelles existantes de la langue arabe, ce qui peut mener à rencontrer des ambiguïtés compliquant ainsi la reconnaissance automatique d'un mot sans voyelles.

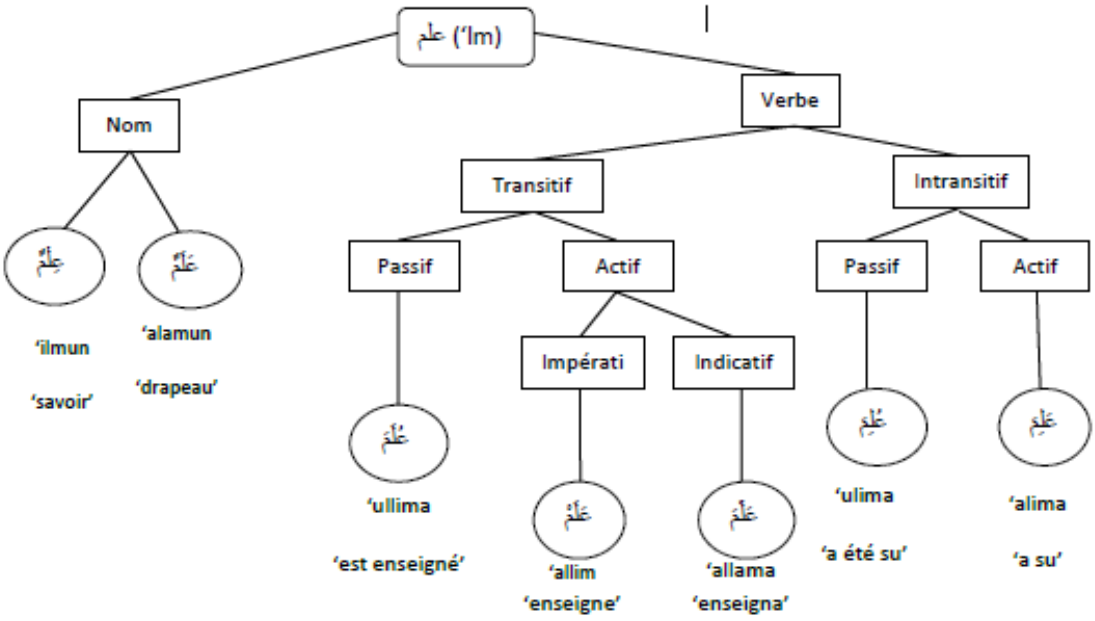


Figure 2.6: Ambiguïté causée par le manque de diacritiques

2.4.2 L’ordre des mots dans une seule phrase

L’ordre des mots dans une phrase donnée est relativement libre, ce qui donne une flexibilité à la phrase. Cette irrégularité qui la rend syntaxiquement ambiguë due a la prise en compte de toutes les règles de combinaison possibles comme l’illustre l’exemple suivant :

- Verbe + sujet + complément :
  تأهلت الجزائر إلى كأس العالم (L’Algérie s’est qualifiée pour la coupe du monde)
- Sujet + verbe + complément :
  الجزائر تأهلت إلى كأس العالم (C’est l’Algérie qui s’est qualifiée en coupe du monde)
- Complément + verbe + sujet
  إلى كأس العالم تأهلت الجزائر (C’est pour la coupe du monde que l’Algérie s’est qualifiée)

2.4.3 Les clitiques

Le terme « clitique » est une généralisation moderne des deux catégories traditionnelles : proclitique et enclitique. Ces unités léxicales ne comportent qu’une consonne et une voyelle brève comme par exemple : (fa ف), (wa و),(ka ك) ,(li ل) .Elles peuvent être considérées comme des éléments autonomes mais elles ont, en même temps, besoin de se rattacher au début ou à la fin de la séquence qui les suit pour ne former qu’un seul mot car elles ne peuvent pas être écrites isolément entre deux blancs.

Exemple : وَكتابهُ → | هُ | + | كتاب | + | وَ |

Enclitique = هُ

Proclitique = وَ

- *Les proclitiques :*

Les proclitiques réservés aux noms et adjectifs :

- L'article défini (' ال ' al- le)
- L a préposition (بِbi - avec), (لِli - pour),(ك ka – comme)

Les proclitiques réservés aux verbes :

- La particule du subjonctif : nasb (لِ li - pour)
- la particule du futur (سَ sa)
- La particule de l'apocopé (لِ li - pour)

Les proclitiques généraux utilisés indépendamment de la catégorie des mots auxquels ils s'attachent :

- Les conjonctions de coordination (ف fa), et (و wa)
- L'article d'interrogation (أ a - est ce que)
- Le marquer de corroboration (لَ la)

- *Les enclitiques :*

  Les enclitiques présentent les pronoms suffixes qui s'attachent toujours à la fin du mot graphique, leur liste est constituée des 17 éléments suivants :

  نِي ي نَا كَ كِ كُمَا كُمْ كُنَّ هُ هَا هُمَا هُمْ هُنَّ هِ هِمَا هِمْ هِنَ

  Un mot graphique ne contient qu'un seul enclitique à la fois. Ils s'attachent aux verbes comme étant un complément-objet et aux noms et prépositions comme un complément du nom ou complément d'objet indirect. Leurs utilisations est régie par certaines restrictions.

2.4.4 Le chevauchement des caractères

Certains caractères se chevauchent à l'intersection des composantes connexes (pseudo-mots ou mots) c'est à dire qu'il est impossible d'encadrer un caractère dans un rectangle sans croiser .son successeur

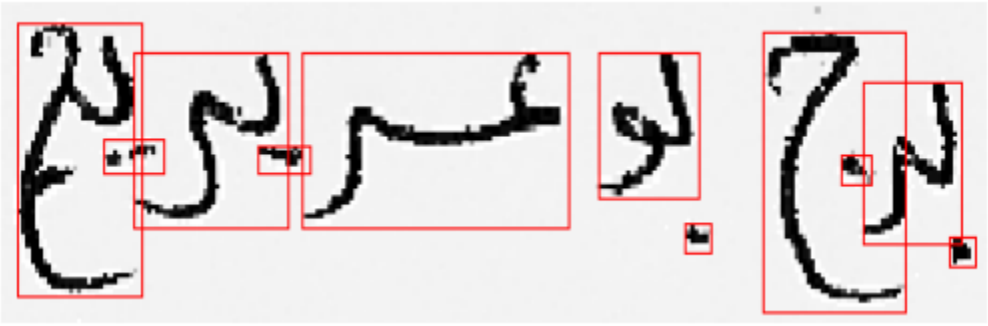


Figure 2.7: Chevauchement des caractères

2.4.5 Les coupures

Les six lettres: ا،د،ذ،ر،ز،و ne peuvent être rattachées à leurs successeurs ce qui peut introduire des coupures dans le mot. En effet, un mot arabe est une séquence d'entités connexes

entièrement séparées appelées pseudo-mot (ou PAWs : Pieces of Arabic Words) (Redouane, 2012). Un mot peut être composé d'un ou de plusieurs pseudo-mots, chaque pseudo-mot est une séquence de lettres liées, notons aussi qu'un caractère isolé peut constituer un pseudo-mot à lui seul.

Figure 2.8: Un mot arabe peut être composé de plusieurs composantes connexes (pseudo-mots ou PAWs).De droite à gauche, 1 seul PAW, 2 PAWs, 4 PAWs, et 3PAWs par mot

2.4.6 Les signes diacritiques

Les signes diacritiques jouent un rôle primordial dans la différenciation entre certaines lettres. En plus, leur manipulation n'est pas tout à fait facile : les prétraitements, notamment la squelettisation et la suppression du bruit, risquent d'altérer la forme des signes diacritiques ou même les supprimer, la variation des styles d'écriture des diacritiques rend complexe leur reconnaissance, en outre, la superposition verticale de ces diacritiques par rapport à la lettre originale n'est pas toujours préservée (Boukerma, 2010).

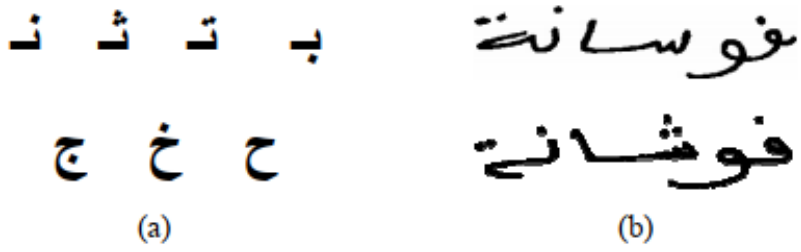


Figure2.9 : Exemple de lettres et mots arabes qui se différencient que par la présence, la position, ou le nombre de signes diacritiques. Source : (Boukerma, 2010)

2.4.7 Ligature verticale

Les liaisons entre les lettres arabes d'un pseudo-mot se situe au niveau de la ligne de base, toutefois certaines lettres peuvent être liées verticalement constituant ainsi des ligatures verticales généralement très complexes à segmenter.

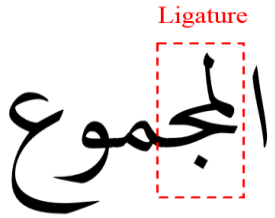


Figure 2.10: Exemple d'une ligature verticale de la langue arabe

2.4.8 Les élongations horizontales

La présence des élongations horizontales qui correspondent à insérer entre les caractères d'une même chaîne une ou plusieurs élongations, ces élongations se situent toujours à gauche du caractère courant et des ligatures verticales.

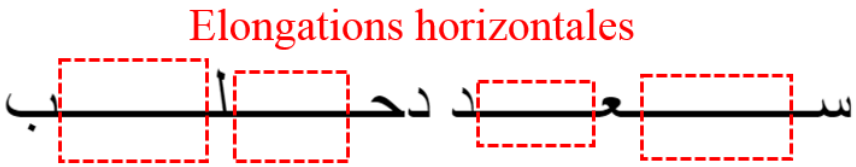


Figure 2.11: Exemple d'élongations horizontales

2.4.9 Forme des lettres

La forme d'une lettre écrite dépend de son contexte et le dessin du glyphe associé diffère selon son emplacement dans une chaîne de caractères. A chaque caractère peut correspondre jusqu'à quatre glyphes différents ce qui lève à environ 100 le nombre de formes à reconnaître. Les formes correspondantes à un même caractère, souvent appelées « formes internes », présentent parfois de sensibles différences ; dans certains cas, il est même difficile d'en déduire s'il s'agit d'une même lettre.

2.4.10 Segmentation des textes

Contrairement au latin, la langue arabe ne dispose pas de la notion de lettres majuscules qui délimitent le début d'une nouvelle phrase dans un texte donné (Maaloul, 2012). De plus les signes de ponctuation ne sont pas utilisés de façon régulière ce qui rend la séparation des phrases plus difficile.

2.5 Quelques OCR de la langue arabe

La reconnaissance l'écriture arabe (AOCR : Arabic OCR) remonte aux années 70, depuis, plusieurs solutions ont été proposées. Elles sont aussi variées que celles utilisées dans le latin. La synthèse des travaux étudiés, montre que les différents types de primitives (structurelles, géométriques, statistiques, transformations globales, corrélations…) et les différentes méthodes de classification (statistiques, structurelles, syntaxique, …) qui existent dans la littérature, ont été pratiquement toutes utilisées dans la description de l'écriture arabe.

Toutefois, nous constatons que le calcul des moments et l'utilisation des descripteurs de Fourier, pour l'extraction des primitives, sont appliqués dans un nombre relativement important de travaux (Alqaisy et Naser, 1985), (El-Sheikh et Guindi, 1988), (Al-Yousefi et Upda, 1992) et (Miled et al, 1998). Ces méthodes sont connues pour leur invariance à la translation, à la rotation et à l'homothétie, de plus, elles tolèrent les faibles variations de formes. De même la classification en arbre de décision est également populaire. Quatre arbres de décision sont élaborés, afin de déterminer l'identité du caractère selon sa position dans le sous mot

Les classifieurs connexionnistes constituent un nouveau paradigme en reconnaissance de formes, les travaux utilisant cette approche en AOCR, sont relativement récents [Amin 94], (Souici et al, 1997). Les modèles utilisés par la majorité des travaux, appartiennent à la famille des réseaux à couches. Le principe des réseaux à couches est de transmettre l'information recueillie sur une couche d'entrée vers une couche de sortie qui exprime la réponse du réseau. Par ailleurs, peu de travaux ont utilisés des méthodes de classification hybrides (Khella, 1992), (Amin et Al-Sadoun, 1994) et (Bouslama et Kishibe, 1999).Les études récentes en OCR

recommandent cette approche, toutefois le choix ainsi que nombre de classifieurs, qui devraient être complémentaires, dépend de l'application considérée.

Le tableau 4 (tiré de (Al-Badr et Mahmoud, 1995) et enrichi par des travaux récents), en fin de ce chapitre, regroupe certains systèmes de reconnaissance de l'écriture arabe en précisant pour chacun le mode utilisé en-ligne ou hors-ligne, l'approche de reconnaissance globale ou analytique, le type de segmentation, la représentation choisie ainsi que les scores réalisés

Référence	Système	Approche	Segmentation	Primitives	Classification	Performance
(Elgammal, 2001)	Hors-ligne, imprimé	Analytique	Externe	-	Grammaire régulière	RC 93.4 %
(El-Khaly,1990)	Hors-ligne, imprimé	Analytique	Externe	moments	Distance	RC 95-100 %
(El-Sheikh, 1988)	Hors-ligne, imprimé	Analytique	Externe	Descripteurs de Fourier	Classifieur topologique	RC 99 %
(El-Sheikh, 1990)	En-ligne, caractères isolés	-	-	Structurelles	Arbre « handcrafted »	RC 99.6 %
(Fehri, 1994)	Hors-ligne, MF	Analytique	Interne	Structurelles/ statistiques	Programmation dynamique	RC 98 %
(Fehri, 1998)	Hors-ligne, manuscrit	Analytique	Externe	Structurelles	Réseaux de neurones/ HMM	-
(Gillies, 1999)	Hors-ligne, imprimé	Analytique	Externe	Structurelles	Réseaux de neurones	RC 89-93.1 %
(Goraine, 1994)	Hors-ligne, imprimé	Analytique	Externe	Chaîne de codes	Struct./ Mesure géom./ contexte	RC 95.87 %
(Haj-Hassan, 1991)	Hors-ligne, imprimé	-	Externe	Structurelles	Syntaxique	RC 99 %
(Hassibi, 1994)	Hors-ligne, imprimé	Analytique	Interne	Structurelles	Réseaux de neurones	RC 99 %
(Jambi, 1993)	Hors-ligne, manuscrit	Analytique	Externe	Structurelles	Dictionnaire	SC 95 %
(Kurdy, 1993)	Hors-ligne, imprimé MF	Analytique	Externe	Structurelles	Morphologie mathématique	RC 98 %

Tableau 2.4 : Tableau récapitulatif précisant les caractéristiques et les performances de certains systèmes
RC : Taux de reconnaissance caractère, RM : Taux de reconnaissance mot, SC : taux de segmentation de caractères, MF : Multifonte, MS : Multiscripteur

Référence	Système	Approche	Segmentation	Primitives	Classification	Performance
(Ajla, 2001)	Hors-ligne, manuscrit	Analytique	Externe	Structurelles	Réseaux de neurones	SC 69.72 %
(Burrow, 2004)	Manuscrit	Globale	-	KNN, moments	-	RC 94%
(Brabols, 1997)	Hors-ligne, Imprimé	Analytique	Externe	-	-	SC 97.01 %
(Bahy, 2001)	Hors-ligne, Manuscrit	Analytique	Externe	Géométriques	Réseaux de neurones	RC 69.7 %
(Bachour, 2004)	Imprimé, caract isolés	Analytique	Externe	Morphologiques\ statistiques	Logique floue	-
(Kanhil, 2004)	Hors-ligne, imprimé	Analytique	Externe	-	-	-
(Kavianifar, 1999)	Hors-ligne, imprimé	Globale	-	Structurelles	KNN/MM	-
(Mamouni, 2002)	Hors-ligne, manuscrit	Analytique	Externe	-	Réseaux de Jeunones	-
(Mehdi, 2002)	Hors-ligne, imprimé	Analytique	Externe	-	-	SC 100%
(Ymaz, 2003)	Hors-ligne, imprimé	Analytique	Externe	Moments	Réseaux de Jeunones	RC 76 %
(Sarfraz, 2003)	Hors-ligne, imprimé	Analytique	Externe	Moments	Réseau de neurones	RC 73 %
(Sari, 2002)	Hors-ligne, Manuscrit	Analytique	Externe	-	-	SC 86 %

Tableau 2.4 (suite) : Tableau récapitulatif précisant les caractéristiques et les performances de certains systèmes.RC : Taux de reconnaissance caractère, RM : Taux de reconnaissance mot, SC : taux de segmentation de caractères, MF : Multifonte, MS : Multiscripteur

Seule une poignée de systèmes OCR commercialisés affirment pouvoir reconnaitre la langue arabe. On retrouve dans (Alghamdi et al, 2017) une évaluation limitées aux quatre systèmes OCR arabes les plus connus, à savoir :

- Automatic reader 11.2 software

  Automatic Reader est un produit commercial développé pour la première fois par la Sakhr Software Company en 1982 pour la reconnaissance textuelle de l'écriture arabe. Il supporte la langue arabe et plusieurs autres langues dont l'écriture comprend des caractères arabes tel que le perse et l'ourdou. Sakhr affirme que ce système a été classé comme le meilleur logiciel OCR arabe existant pour le texte de haute qualité images d'évaluateurs du gouvernement américain (Sakhr Software OCR, 2017). Il prend en charge multi-font d'images de type et multirésolution. Toutefois, la taille de police 8 n'est pas prise en charge par l'Automatic Logiciel OCR Reader .

- FineReader 12 software

  FineReader a été commercialisé par une société mondiale appelée ABBYY. Ses performances ont été longtemps améliorées par cette société, il supporte 190 langues dont le script arabe, avec des fontes, tailles et résolutions différentes.

- Clever page software

  Clever page a débuté comme un prnarojet de recherche de doctorat fait par El-Mahallawy en 2008. Depuis, il a été développé par RDI (Research and Devlopment International) comme un OCR arabe omni-police. Il ne marche que sur les pages ayant 300 dpi (dots per inch) mais supporte plusieurs fontes et tailles.

- Tesseract software

  Tesseract représente un moteur OCR conçu chez HP (Hewlett-Packard) entre 1984 et 1994. Google s'est ensuite occupé de sa maintenance et la publié comme un logiciel open source en 2005. Cependant le vecteur arabe n'a été ajouté qu'en 2013. C'est le seul OCR arabe disponible gratuitement. Il prend en charge plusieurs fontes, tailles et résolutions d'images.

Le tableau suivant montrera les résultats obtenus après évaluation :

Font Type	OCR System			
	Sakhr (%)	ABBYY (%)	RDI (%)	Tesseract (%)
Traditional Arabic	48.54	67.66	51.88	47.04
Tahoma	40.52	69.91	26.38	38.37
Simplified Arabic	52.97	67.69	44.94	46.75
M Unicode Sara	36.03	59.40	25.92	33.72
Diwani letter	18.13	18.47	18.13	23.32
DecoType Thuluth	36.12	37.71	24.26	32.48
DecoType Naskh	48.88	50.22	41.63	40.92
Arabic transparent	51.56	75.19	46.00	48.61
Andalus	28.07	37.53	21.68	25.34
AdvertisingBold	57.35	70.26	27.20	39.39

Tab*leau 2.5 : Evaluation des performances des OCR arabes sur des fontes différentes*
Source : (Mansoor et al, 2017)

2.6 Conclusion

Nous avons présentés dans ce chapitre, les principales propriétés morphologiques et typographiques de l'écriture arabe. La reconnaissance optique de l'arabe reste une tâche encore non résolue. Nous avons aussi passé en revue dans la dernière partie de ce chapitre certains travaux réalisés en OCR arabes. Les problèmes majeurs dans ce domaine se ramènent à la cursivité de l 'écriture et à la sensibilité de certaines caractéristiques topologiques de l'arabe à la dégradation, en l'occurrence les points diacritiques.

Chapitre 3
Segmentation de l'écriture et extraction des caractères

3.1 Introduction

La segmentation représente l'une des parties les plus cruciales de l'architecture d'un système OCR. Le but de ce chapitre est de présenter les techniques et approches utilisées dans la littérature pour la segmentation du texte manuscrit cursif, ainsi que les avantages et les inconvénients de chacune d'elles, pour justifier notre choix lors de notre contribution.

3.2 Les différents niveaux de segmentation d'un document

Lors d'une reconnaissance hors-ligne de l'écriture manuscrite, la segmentation consiste à extraire des différentes zones de l'image à segmenter, des lettres ou bien des sous-mots utiles pour des traitements futurs. Cette étape a un effet sur le taux de reconnaissance, une bonne segmentation qui détermine les bons segments donne un bon taux de reconnaissance, à l'inverse, une mauvaise segmentation quant à elle va entraîner une chute du taux de reconnaissance. Généralement, il existe quatre niveaux de segmentation :

- Segmentation de la page
- Segmentation du texte en lignes
- Segmentation de la ligne en mots
- Segmentation du mot en caractères

3.2.1 Segmentation de la page

La segmentation d'une page consiste à retrouver la structure physique du document en délimitant les différentes parties homogènes (les graphiques, blocs de texte …) et localiser dans chaque page les zones d'information conformément à leur apparence physique.

3.2.2 Segmentation du texte en lignes

Pour déterminer les différentes phrases d'un bloc de texte, la projection horizontale représente la méthode la plus utilisées à cet effet.

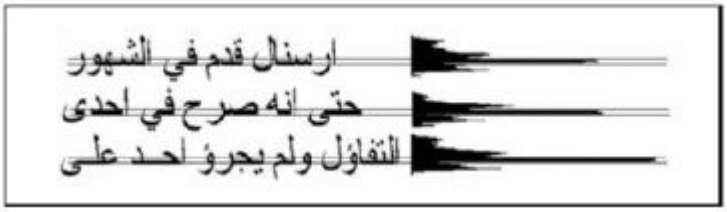


Figure3.1 : Projection horizontale des lignes

3.2.3 Segmentation de la ligne en mots

Après son extraction, une ligne est alors segmentée en mots et sous mots en se basant sur les différents espaces qu'elle contient. Pour ce fait, deux approches de segmentation de ligne en mots existent : Approche descendante en utilisant l'histogramme de projection verticale des lignes pour détecter les espaces entre les mots et pouvoir les séparer. Cette dernière est simple et similaire à la projection horizontale nous comptons le nombre de pixels noirs sur chaque colonne de l'image texte. Si le nombre de pixels noirs n'est pas égal à zéro, il indique une partie connectée et par conséquent le texte n'est pas segmenté. D'autre part, si le nombre de colonnes blanches successives rencontré est supérieur à un seuil, la partie courante est segmentée. Cette méthode n'est pas efficace lorsque les sous mots se chevauchent verticalement. Approche ascendante ou d'autres techniques sont utilisées comme : le suivi du contour, détermination du squelette ou la détermination des composantes connexes…

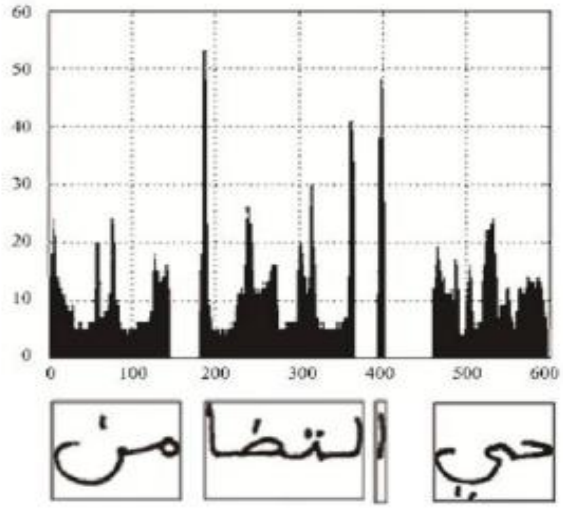


Figure 3.2 : Segmentation d'une ligne en sous-mots par projection verticale

3.2.4 Segmentation du mot en caractères

La segmentation des mots cursifs en graphèmes (ou caractères) est une étape indispensable pour les systèmes de reconnaissance analytique. Elle vise à découper le mot en ses éléments constructifs afin de les identifier isolément en repérant des points de segmentation potentiels sur la ligne de base. Les premières tentatives de segmenter les caractères arabes sont apparues en 1981 dans les travaux de Parhami et Taraghi (Parhami et taraghi, 1981) suivis par l'œuvre d'Amin et Masini en 1983 (Amin et Masini, 1983)

3.2.4.1 Classification des différentes approches de segmentation

D'une manière générale, il existe deux approches de segmentation (Chergui, 2013):

a. Segmentation explicite

Lors de la segmentation explicite, une sélection des points de segmentation les plus probables est effectuées (à partir du squelette, du contour, etc..). Une fois que ces points sont identifiés, il existe deux méthodes pour choisir la segmentation finale : les méthodes dites de segmentation puis de reconnaissance (segmentation-based) et les méthodes de segmentation-reconnaissance

(segmentation-free ou recognition-based). La différence entre ces deux méthodes réside dans la dépendance ou non entre l'étape de segmentation et celle de la reconnaissance.

La méthode de segmentation puis reconnaissance sélectionne les meilleurs points de segmentation sans avoir recours à la reconnaissance. Cette méthode nécessite la disponibilité d'un algorithme de segmentation fiable car la moindre erreur remet en cause la totalité des traitements ultérieurs.

La méthode de segmentation-reconnaissance quant à elle alterne les phases de segmentation et de reconnaissance pour valider les hypothèses de segmentation par la reconnaissance. Bien que cette méthode offre une segmentation beaucoup plus fiable que la précédente, elle souffre d'u inconvénient principal en temps de calcul lors de la comparaison de l'ensemble des hypothèses (Oliveira et al, 2000).

b. Segmentation implicite

Dans cette approche, la segmentation et la reconnaissance sont réalisées conjointement. En réalité, il n'y a pas de pré-segmentation ou dissection du mot, bien qu'un découpage a priori de l'image en intervalle de grandeur régulière est effectué classiquement, il peut le faire de deux manières, soit par fenêtrage, soit par recherche de primitives (Souici, 2006).

On retrouve plusieurs techniques de segmentation dans la littérature, dans (Menasri, 2008)l'auteur expose les techniques de segmentation suivantes :

- Segmentation basée sur les projections :

  Cette méthode qui fonctionne mieux avec les documents imprimés

3.2.4.2 Quelques méthodes de segmentation de caractères

a. Segmentation basée sur les projections

La segmentation basée sur les projections vise à simplifier les informations en changeant leur format 2D en un format 1D. Elle fonctionne mieux sur les documents imprimés, en particulier avec les polices ne formant pas de ligature telle que l'arabe simplifié.

Le principe fondamental de ces méthodes est de supposer que tout trait de connexion entre deux caractères détient moins d'épaisseur que les autres parties du mot ou il se trouve. Pour procéder, un calcul de la projection verticale de l'image est fait en premier. Après son obtention, cette projection peut être utilisée de différentes manières :

Zheng et al par exemple proposent dans (Zheng et al, 2014) un algorithme de segmentation des caractères arabes se basant sur l'histogramme vertical et quelques autres règles. De plus, les caractéristiques structurelles entre les régions de fond et les composants de caractère, les caractéristiques des caractères arabes, sont également utilisées pour vérifier si le pseudo-mot ne comprend qu'un seul caractère. Ensuite, l'histogramme vertical, et d'autres règles ont été utilisés pour trouver des points de segmentation réels. Enfin, les pseudo-mots ont été divisés en points de segmentation. Les résultats expérimentaux montrent que l'algorithme atteint environ 94% de segmentation correcte.

Les chercheurs dans (Nawaz et al, 2003) ont utilisé la projection verticale de la zone centrale au lieu du mot entier. Quatre zones de ligne de texte ont été identifiées ; À savoir la ligne de base, la zone centrale, et la zone inférieure. Si la valeur de la projection verticale de la zone centrale est inférieure à deux tiers de l'épaisseur de ligne de base, la zone est considérée comme une zone de connexion entre deux caractères. Alors que toute zone suit la zone de connexion avec une plus grande valeur étant considérée

comme le point de départ d'un nouveau caractère tant que le profil est supérieur à un tiers de la ligne de base. Dans (Lasri, 2014), Y. Lassri a appliqué une projection verticale simple sur le mot après avoir éliminé la ligne de base, cette méthode a rencontré plusieurs difficultés telles que la détection de la ligne de base comme étant la partie qui présente le nombre maximal des points noirs dans la direction horizontale, ainsi, la difficulté de savoir si le mot contient une ligne de base, ou s'il est constitué seulement des caractères isolés .On conclut que les méthodes basées sur les projections verticales sont indépendantes de la forme, la taille ou la police des caractères, elles se basent surtout sur la détermination de la ligne de base. Elles sont plus efficaces sur les textes imprimés ou les polices ne contiennent pas de chevauchements ou de ligatures.

b. Segmentation basée sur la squelettisation

Les informations essentielles d'une forme sont stockées dans son squelette. De nombreux algorithmes ont été proposés pour extraire le squelette, tels que la transformée de distance, l'amincissement homotypique avec ses deux approches : séquentielle et parallèle comme l'algorithme de Zhang et Suen (Zhang et Suen, 1984).

Dans la méthode d'El-Khaly et Sid-Ahmed (El-khaly et Sid-Ahmed, 1990), la ligne de base du mot aminci (squelette) est trouvée en première étape, c'est la ligne qui contient le nombre maximal des points noirs. Puis seulement les colonnes qui n'ont pas de pixels au-dessus ou en dessous de la ligne de base sont considérés pour trouver les points de segmentation. Le point de segmentation sera au milieu du segment de connexion.

La procédure de segmentation présentée dans (Goraine et al, 1992) décrite par H.Goraine, M.Usher, est similaire à celle d'AIEmami dans (AIEmami, 1988). L'idée de l'algorithme, après avoir obtenu le squelette du mot en utilisant l'algorithme de Hilditch, est de trouver le point de départ en utilisant un organigramme tout en se basant sur une classification des points rencontrés : point de connexion, point caractéristique ou jonction, trait. La deuxième étape est de trouver le point final et ensuite de tracer un trait depuis le début jusqu'à la fin. Par la suite, le tracé est isolé et éliminé de l'image.

Le premier point (à droite) détecté est pris comme point final de du tracé suivant, et la procédure de segmentation est répétée jusqu'à ce qu'il n'y ait plus de traits.

Al-Sadoun et Amin (Al-Sadoun et Amin, 1995) ont tracé le squelette de droite à gauche en utilisant une fenêtre 3*3 pour identifier les points potentiels de segmentation. Ensuite, un arbre binaire est construit et le squelette est représenté à l'aide du code de Freeman (Freeman, 1968). Chaque nœud de l'arbre binaire décrit la forme de la partie correspondante du pseudo-mot. L'arbre binaire est lissé pour minimiser le nombre de nœuds en éliminant les nœuds vides, en minimisant la chaîne de code Freeman et en éliminant (en minimisant) tout bruit dans l'image amincie. Enfin, l'arbre binaire est segmenté en sous arbre de sorte que chaque sous arbre décrit un caractère en utilisant des primitives, y compris des lignes, des boucles et des boucles doubles. Certaines règles ont été définies pour assurer les limites correctes de caractères. L'algorithme peut être appliqué à n'importe quelle police et la taille du texte arabe, en plus, il peut être appliqué au texte manuscrit. Un avantage de cette méthode est que l'identification de la ligne de base devient inutile puisque le pseudo-mot est décrit par un arbre binaire, ce qui permet d'économiser du temps de traitement.

Les différents algorithmes d'amincissement de ces méthodes peuvent modifier la forme des caractères surtout si l'image n'est pas de bonne qualité, rendant ainsi le caractère difficile à reconnaitre.

c. Segmentation basée sur le contour

L'utilisation des contours durant la segmentation réduit de manière significative la quantité de données à traiter en éliminant les informations qu'on peut juger moins pertinentes tout en préservant les propriétés structurelles de l'image. Dans (Bushofa et Spann, 1997), le contour supérieur est examiné pour extraire les points de segmentations en pseudo-mot potentiels. Cela se fait en traçant cette partie de gauche à droite en commençant par le premier point au-dessus de la ligne de base. Lorsqu'un point maximum dans la direction verticale est atteint, il est considéré comme un pic si

sa valeur est supérieure à une valeur du seuil (t1 = ligne de base + t/6) avec t est la distance entre le pic du mot et la ligne de base. Après cette étape, la valeur des coordonnées de contour commence à descendre jusqu'à ce qu'elle arrive à un point minimum. Si cette valeur est inférieure à la valeur du seuil (t1), elle est considérée comme un point de segmentation à condition qu'elle soit suivie d'un pic. Cette procédure se poursuit jusqu'à ce que toutes les coordonnées de la partie supérieure du contour soient examinées. Le point de segmentation est considéré entre deux pics. Si aucun pic n'a été trouvé après avoir rencontré un point minimum, ce point minimum est négligé.

C. Olivier, H. Miled dans (Olivier et al, 1996) ont proposé une méthode qui utilise le contour supérieur pour segmenter le texte arabe manuscrit. Tout d'abord, ils ont détecté le contour de chaque pseudo-mot avec le code de Freeman(Freeman, 1968) . Le contour est étudié dans le sens inverse des aiguilles de la montre, ensuite le contour supérieur est déterminé par les directions F3. F1 et F5. Figure 17 Un filtre est appliqué pour ne garder que les points du contour supérieur situés dans la partie supérieure du pseudo-mot. Puis un automate est utilisé pour détecter les minimums locaux du contour supérieur, ces minimums sont pris pour la segmentation du pseudo-mot. Dans ce travail, C. Oiivier, H. Miled ont évalué le système sur une base de données contenant 6000 mots arabe, qui sont les noms du peuple tunisien écris par 20 fontes différentes. Ils ont pu avoir un taux d'erreur de 2.59%.

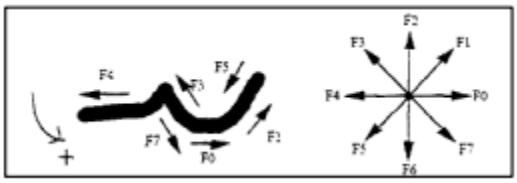


Figure 3.3: détection du contour supérieur en utilisant le code de Freeman

Les résultats des méthodes se basant sur le contour pour la segmentation des caractères sont très remarquables par rapport aux autres techniques malgré les problèmes qu'on peut rencontrer

lors de leurs utilisations telles que la discontinuité du contour lors de la présence de bruit sur l'image, l'échec de segmentation lors de la présence de ligatures ou la difficulté à retrouver la ligne de base en cas de besoin, ce qui peut influer sur tout le système. La variabilité des méthodes de détection de contours tel que le filtre de Sobel, filtre de Prewitt, filtre de Canny engendre l'obtention de plusieurs formes de contours selon la technique utilisée.

d. Segmentation basée sur le Template matching

Le template matching est une technique qui aide à rechercher les petites parties d'une image qui correspondent à une image modèle lors du traitement d'une image numérique. Elle est aussi utilisée lors de la reconnaissance de caractères ou de mots en utilisant comme référence une base de données ou des images de caractères ou de mots sont stockées. Cette méthode peut être appliquée à la segmentation de caractères en délimitant manuellement la caractère à reconnaitre sur l'image avant de le comparer.

Dans (Bushofa&Spann1, 1997), une technique a été proposée pour chercher l'apparition d'un angle formépar la jonction de deux caractères à la ligne de base Figure 18, en utilisant une fenêtre 7 × 7, pour examiner le voisinage des caractères. Cet angle est pris comme un point de segmentation potentiel. Pour valider les points précédemment calculés, le caractère extrait est comparé avec un model déjà stocké manuellement.

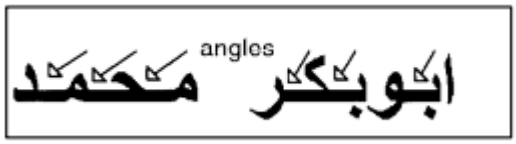


Figure 3.4: les angles de jonction entre les caractères

Malgré les bons résultats que peut apporter cette technique, sa réussite à trouver le bon angle dépend fortement du bruit de l'image. Le template matching n'est pas approprié à la segmentation du texte manuscrit ou imprimé dans le cas où les fontes et tailles de l'écriture ne sont pas uniformes.

3.3 Conclusion

Dans ce chapitre nous avons essayé d'exposer les différentes méthodes utilisées lors de la phase de segmentation. Ces méthodes ont connu beaucoup de progrès ces derniers temps. Des techniques variées influencées par l'évolution dans les domaines tels que la reconnaissance de la parole et la reconnaissance en ligne des caractères ont émergés. La difficulté dans la réalisation d'une segmentation performante dépend généralement de la nature du document à lire et de sa qualité. Le taux de mauvaise segmentation croit progressivement à partir de l'écriture imprimée à l'écriture manuscrite jusqu'à l'écriture manuscrite cursive où la difficulté devient plus importante. La performance d'un système de reconnaissance de l'écriture dépend fortement des résultats de la phase de segmentation.

Chapitre 4

Contribution à l'amélioration du taux de reconnaissance des caractères de la langue Arabe

Dans ce chapitre, nous allons présenter un système de reconnaissance hors-ligne de l'écriture arabe multi-fontes nommé « SRCAA » , qui permet la reconnaissance d'un texte à partir d'une image. Le document acquis subit en premier lieu quelques opérations de prétraitement pour améliorer sa qualité, puis une segmentation pour obtenir les entités à reconnaître. Dans un premier temps le système détecte les différentes lignes du texte pour les segmenter en mots. Le classifieur comparera l'image de chacun des mots présents sur le document avec les images de la base de données pour les reconnaître. Si un mot n'obtient aucune correspondance, une segmentation en caractères isolés et une reconnaissance de ces derniers seront effectuées. Ainsi en concaténant les résultats retournés, nous obtenons le texte reconnu. Une étape finalement de post traitement augmente le taux de la reconnaissance du système.

4.1 Corpus

4.1.1 Base de données APTI

La base de données APTI (Arabic Printed Text Image) a été mise à la disponibilité de la communauté scientifique gratuitement dans un but non-commercial à partir de juillet 2019 par Fouad Slimane, Rolf Ingold, Slim Kanoun, Adel M.Alimi et Jean Hennebert. Elle contient 45'313'600 images de mots arabes décomposables et non-décomposables, disponibles dans dix fontes (Andalus, Arabic Transparent, AdvertisingBold, Diwani Letter, DecoType Thuluth, Simplified Arabic, Tahoma, Traditional Arabic, DecoType Naskh, M Unicode Sara) avec une taille qui varie entre 6 et 24 points et quatre styles différents (plain, bold, italic, italic and bold). Elle peut être obtenue en contactant l'un de ses créateurs.

	Mots	Sous-mots	Caractères
Nombre	113’284	274’833	648’280
Nombre de fontes	10	10	10
Nombre de tailles de fontes	10	10	10
Nombre de styles de fontes	4	4	4
Total	***45’313’600***	***109’933’200***	***259’312’000***

Tableau 4.1 : Quantité de mots, sous-mots et de caractères disponibles dans la base de données APTI

Figure 4.1: Fontes utilisées lors de la génération de la base de données APTI : (A) Andalus, (B) Arabic Transparent, (C) AdvertisingBold, (D) Diwani Letter, (E) DecoType Thuluth, (F) Simplified Arabic, (G) Tahoma, (H) Traditional Aatbic, (I) DecoType Naskh, (J) M Unicode Sara

Chaque mot est représenté en détail par un fichier XML comme le montre la photo suivante :

```
<?xml version="1.0" encoding="UTF-8" ?>
- <wordImage id="78">
- <content transcription="ﺎﻟﻟ" nPaws="4">
    <paw id="1" nbChars="1">Alif_I</paw>
    <paw id="2" nbChars="2">Laam_B TildAboveAlif_E</paw>
    <paw id="3" nbChars="2">Laam_B Alif_E</paw>
    <paw id="4" nbChars="1">Faa_I</paw>
  </content>
  <font name="Arabic Transparent" fontStyle="Plain" size="24" />
  <specs encoding="png" width="96" height="36" effect="none" />
  <generation type="downsampling5" renderer="java" filtering="antialiasing" />
</wordImage>
```

Figure 4.2 : Exemple d'un fichier XML décrivant le caractére «âalaf »

4.1.2 Base de données de caractères isolés

En nous inspirant de l'organisation de la base de données APTI, nous avons tenté de créer une petite base de données multi-fontes qui comporte des images de caractères isolés à différents emplacements dans un mot ainsi que des chiffres. Chaque image est accompagnée d'un fichier XML pour décrire son contenu.

```
<?xml version="1.0" encoding="UTF-8"?>
<wordImage id="04">
  <content transcription="ه">
  </content>
  <font name="Traditional Arabic" fontStyle="Plain" size="8" />
  <specs encoding="png" width="18" height="10" />
</wordImage>
```

(a)

(b)

Figure 4.3: Représentation de la lettre « ha » dans notre base de données : (a) : Fichier XML déscriptif , (b) : La lettre « ha »

Pour chaque lettre de l'alphabet arabe, un dossier a été créé, il contiendra en moyenne quatre images (ou plus par exemple dans le cas de la « hamza »). Chaque image sera accompagnée d'un fichier XML qui contiendra :

- Un ID qui permet de reconnaître la lettre pour l'imprimer lorsqu'une similitude est trouvée
- La transcription de la lettre trouvée
- Le nom et la taille de la fonte utilisée
- Le format de l'image stockée ainsi que ses dimensions

Le traditional arabic est la seule fonte utilisée sur notre base de données et la taille des caracétres ne varie pas, elle est fixée à 8.

4.2 Architecture générale du système

Notre système contiendra trois modules importants qui sont le prétraitement, la segmentation et la reconnaissance avec extraction des caractéristiques de l'écriture. La figure montre le schéma général qui constitue notre système. Les détails de chaque composant seront discutés dans la suite de ce chapitre.

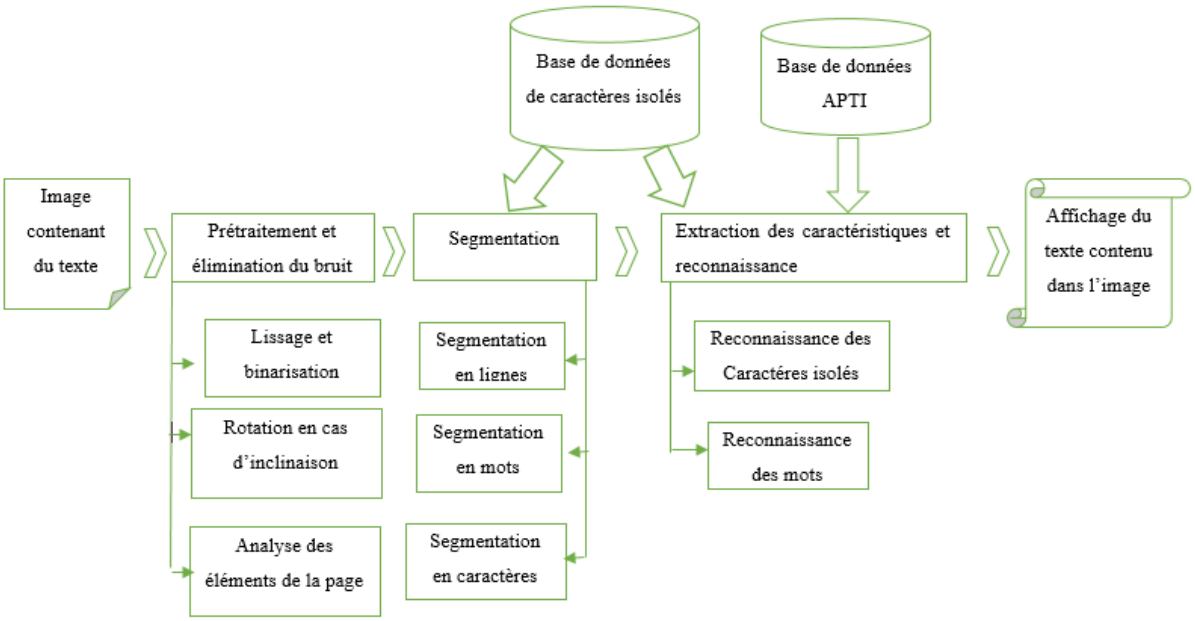


Figure 4.4 : Schéma général du système SRCAA

4.3 Prétraitement de l’image

L’entrée du système est une image capturée d’un document PDF, elle représente la plupart du temps une page scannée d’un livre avec une résolution supérieure à 300 dpi. La plupart des machines de reconnaissance se contentent d’une binarisation simple en utilisant l’une des méthodes existantes et citées dans le chapitre. Nous avons donc essayé plusieurs opérations sur les images de notre corpus dans le but d’effectuer un prétraitement efficace sans perdre beaucoup d’informations pertinentes présentes sur le document, tout en prenant en compte la complexité de l’écriture des caractères arabes. La méthode suivante à donner les meilleurs résultats :

4.3.1 Application d’un filtre gaussien

Nous avons vu qu'une des principales sources du phénomène de sur-segmentation ou de mauvaise classification est le bruit présent dans les images naturelles. Afin de pallier à ce bruit on opère un moyennage spatial à l'aide d'un filtre gaussien dans le but de lisser l’image.

4.3.2 Binarisation en niveaux de gris

Nous utiliserons la méthode d’Otsu pour transformer le document en une image en niveaux de gris en raison des bons résultats qu’elle a donnée après plusieurs tests. Ceci revient à séparer les pixels de l’image en deux classes après avoir calculé son histogramme au préalable, la première ayant un niveau maximal (typiquement 255) et la seconde un niveau minimal (0)

a. Calcul de l'histogramme

   Le calcul de l'histogramme est très simple. On initialise un tableau T avec des 0. Généralement, ce tableau est constitué de 255 cases correspondant aux 255 niveaux de gris d'une image. Ensuite, si p (i,j) représente la valeur du pixel au point (i,j), on balaye toute l'image et on compte le nombre de fois ou un niveau de gris apparait.

b. Séparation des pixels en deux classes

D'après l'article (Otsu, 1979), la séparation se fait à partir de la moyenne et de l'écart-type. Pour que le procédé soit indépendant du nombre de points dans l'image N, on normalise l'histogramme : pi = ni/N ou ni représente le nombre de pixels de niveau i. On peut calculer alors les 2 moments utilisés : mu(k) = sum(i=1,k) i*pi w(k) = sum(i=1,k) pi l'expression sum(i=1,k) représente la somme de i=1 à k. On note muT=mu(256), où 256 est le nombre totale de niveaux de gris. Si on appelle w0 la probabilité de la classe C0 et w1 la probabilité de la classe C1, alors : w0 = w(k*) où k* représente le niveau de seuil et w0 = 1 - w(k*). Si on note de même mu1 et mu0 avec : mu0 = mu (k*)/w (k*), mu1 = (muT -mu(k*))/(1-w(k)) . L'image totale conserve certaines propriétés, d'où on peut tirer les relations : w0mu0+w1mu1=muT w0+w1=1 En introduisant un paramètre pour évaluer la qualité du niveau de seuillage, on obtient : s2=w0w1(mu1-mu0)2. La valeur précédente est fonction de k. On calcule donc cette valeur pour les 256 niveaux de gris de l'image. (On peut enlever les valeurs 0 et 255 qui correspondent à affecter tous les pixels à la même classe). A partir de w(k) et mu(k), on calcule donc : s2(k)=w(k)(1-w(k))(mu Tw(k)-mu(k))2. La valeur du seuil k* est obtenue en trouvant la valeur maximum de tous les s2. Il ne reste plus qu'à comparer la valeur de tous les pixels de l'image au seuil ainsi trouvé pour définir leur classe.

4.3.3 Correction de l'inclinaison de la page

Les angles de l'inclinaison des lignes dans un document arabes sont relativement uniformes si le document est incliné dans sa globalité. Cette hypothèse est vérifiée pour les documents de texte imprimé ; dans ce cas, une correction de l'inclinaison globale du document donne des résultats satisfaisants. Ici, nous utiliserons la transformée de Hough, et les histogrammes de projection, elles représentent les deux méthodes les plus utilisées pour cette opération.

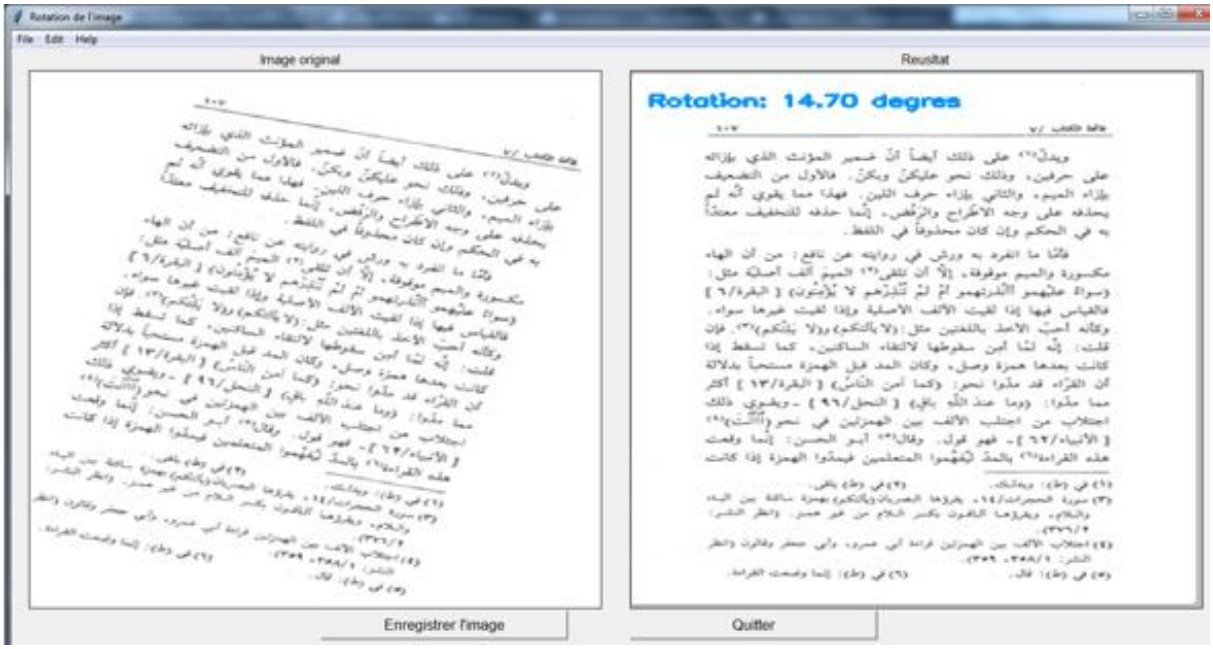


Figure 4.5: Résultat du redressement d'une page par notre système

4.3.4 Application de quelques opérations morphologiques

Les opérations morphologiques visent à modifier la valeur d'un pixel présent sur une image en fonction des valeurs de ses voisins, un masque binaire est alors utilisé pour prendre en compte le voisinage du pixel. L'érosion et la dilatation forment les deux opérations morphologiques de base que nous allons utiliser.

a. Erosion

Cette opération amincit les bordures de l'objet en premier plan. Le noyau se glisse sur l'image. Un pixel dans l'image originale (soit 1 ou 0) ne sera considéré « 1 » que si tous les pixels sous le noyau, dont il est le centre, sont à « 1 », sinon il est érodé (mis à zéro).

b. Dilatation

Elle forme l'opération inverse de l'érosion, un pixel obtient la valeur de « 1 » si au moins un pixel sous le noyau détient la valeur de « 1 », ce qui permet d'augmenter la taille de

l'objet en premier plan. Cette méthode est utilisée d'habitude après l'érosion pour ne pas agrandir le bruit de l'image.

Figure 4.6: Résultats de binarisation en utilisant quelques méthodes

4.4 Séparation des composantes de la page

Une page contenant du texte peut aussi contenir d'autres éléments (images, bordures, etc…). L'analyse des différents blocs constituant une image est une étape très importante qui permet au système de délimiter les paragraphes à reconnaitre. Nous utiliserons au niveau de cette étape notre propre implémentation de la méthode proposée par F. Shafait dans (Shafait et al, 2006) pour l'analyse des composantes des images de document en langue ourdou, qui est une langue utilisant des caractères arabes.

Image (a) Image (b)

Figure 4.7 : Analyse des éléments de deux images : (a) : image contenant un dessin (b) : image contenant du texte seulement

4.5 Segmentation de l'image en lignes de texte

Nous avons utilisé la projection horizontale pour segmenter les lignes du texte présent sur l'image. Nous utiliserons un tableau lors du parcours du graphe de projection obtenu, on fixera sa taille à la hauteur de l'image, lorsqu'on trouve un point noir la valeur de 1 sera ajoutée au tableau, dans le cas contraire, si un pixel blanc est trouvé on mettra un 0. Ainsi on pourra compter le nombre de zéros successifs pour fixer un seuil qui représente le saut de ligne.

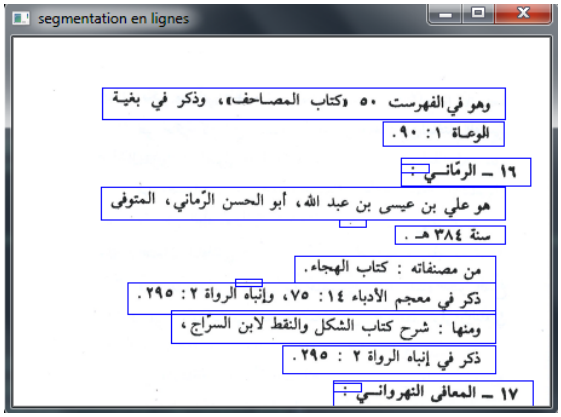


Figure 4.8 : Résultat de la segmentation en lignes d'un paragraphe

4.5.1 Reconnaissance de l'écriture

La reconnaissance se fait en premier lieu par mot, le système prend en entrée un vecteur qui contient les images des mots qui constituent chaque ligne. Pour chaque image on effectue les étapes suivantes

4.5.1.1 Prétraitement sur l'image du mot

L'image du caractère doit être redimensionnée pour correspondre à la taille des images de la base de données APTI (Hauteur de 15 pixels sur une largeur de 55 pixels sur 16 pixels) pour faciliter la comparaison, un lissage par filtre gaussien est ensuite effectué pour réduire le bruit.

4.5.1.2 Reconnaissance de l'image du mot

Pour cette étape nous nous sommes basés sur l'étude comparative de W. Yin dans (Yin et Al, 2017), et nous avons décidé d'utiliser un réseau de neurones à convolution pour la classification.

Avant la classification, avons utilisé la méthode de canny pour extraire le contour des mots à reconnaitre. Comparer le contour du mot avec les contours des mots présents sur la base APTI réduit le temps d'exécution et facilite la tâche de classification.

Nous utiliserons par la suite un réseau de neurones à convolution (CNN) déjà implémenté et disponible sur la bibliothèque Keras de Python. Les étapes d'utilisation de ce réseau de neurones sont les suivantes :

- **Choix et préparation des échantillons**
  Le processus d'élaboration d'un réseau de neurones commence toujours par le choix et la préparation des échantillons de données. Comme dans les cas d'analyse de données, cette étape est cruciale et va aider le concepteur à déterminer le type de réseau le plus approprié pour résoudre son problème. La façon dont se présente l'échantillon conditionne : le type de réseau, le nombre de cellules d'entrée, le nombre de cellules de sortie et la façon dont il faudra mener l8'apprentissage, les tests et la validation.
- **Elaboration de la structure du réseau**
  La structure du réseau dépend étroitement du type des échantillons. Il faut d'abord choisir le type de réseau : un perceptron standard, un réseau de Hopfield, un réseau à décalage temporel (TDNN), un réseau de Kohonen, un ARTMAP etc... Dans le cas du perceptron par exemple, il faudra aussi choisir le nombre de neurones dans la couche cachée. Plusieurs méthodes existent et on peut par exemple prendre une moyenne du nombre de neurones d'entrée et de sortie, mais rien de vaut de tester toutes les possibilités et de choisir celle qui offre les meilleurs résultats.
- **Apprentissage**
  L'apprentissage consiste tout d'abord à calculer les pondérations optimales des différentes liaisons, en utilisant un échantillon. La méthode la plus utilisée est la *rétropropagation* : on entre des valeurs du les cellules d'entrée et en fonction de l'erreur obtenue en sortie (le *delta*), on corrige les poids accordés aux pondérations. C'est un cycle qui est répété jusqu'à ce que la courbe d'erreurs du réseau ne soit croissante (il faut bien prendre garde ne ne pas sur-entrainer un réseau de neurones qui deviendra alors moins performant). Il existe d'autres méthodes d'apprentissage telles que le *quickprop* par exemple.

- **Validation et Tests**

  Alors que les tests concernent la vérification des performances d'un réseau de neurones hors échantillon et sa capacité de généralisation, la validation est parfois utilisée lors de l'apprentissage. Une fois le réseau calculé, il faut toujours procéder à des tests afin de vérifier que notre réseau réagit correctement. Il y a plusieurs méthodes pour effectuer une validation : la cross validation, le bootstrapping... mais pour les tests, dans le cas général, une partie de l'échantillon est simplement écarté de l'échantillon d'apprentissage et conservé pour les tests hors échantillon. On peut par exemple utiliser 60% de l'échantillon pour l'apprentissage, 20% pour la validation et 20% pour les tests. Dans les cas de petits échantillons, on ne peut pas toujours utiliser une telle distinction, simplement parce qu'il n'est pas toujours possible d'avoir suffisemment de données dans chacun des groupes ainsi crée. On a alors parfois recours à des procédures comme la cross-validation pour établir la structure optimale du réseau.

Dans le cas où le mot obtient une correspondance sur la base de données nous passons au mot suivant. Dans le cas contraire, une reconnaissance de caractères isolés sera utilisée en passant par une méthode de segmentation hybride que nous avons proposé.

4.5.1.3 Segmentation de l'image du mot en caractères avec extraction des caractéristiques

Comme nous l'avons déjà mentionné, cette étape ne sera effectuée que si l'image du mot à reconnaitre n'obtient aucune correspondance sur la base de données APTI. Après quelques opérations de prétraitement nous procédons de la manière suivante :

1. Une projection horizontale est appliquée pour séparer les sous-mots du mot, cela se fait en se basant sur les espaces blancs entre les caractères comme le montre la figure suivante. Chacun sous-mot sera sauvegardé à part.

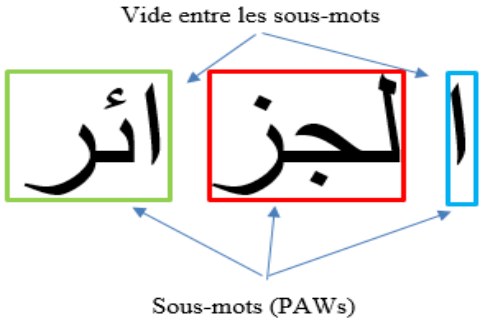


Figure 4.9 : Séparation des sous-mots d'un mot grâce à la projection horizontale

2. Le redimensionnement des images déjà réalisé aide à unifier la taille des caractères quel que soit le style de la police utilisée. Cela nous permet de faire une estimation manuelle de la largeur des caractères au sein de l'image. Nous proposons donc un seuil égal à 3 pixels comme largeur moyenne d'un caractère.

3. Nous comparerons la largeur des sous-mots extraits précédemment avec le seuil fixé à 3 pixels, si la valeur trouvée vient à être égale ou inférieure à celle du seuil cela indique forcément la présence d'un caractère isolé. Après squelettisation de ce dernier, la méthode de template-matching sera utilisée pour tenter de reconnaitre ce caractère en le comparant avec les photos de caractères isolés présentes sur la base de données des caractères. Le repérage et la reconnaissance des caractères isolés permettent de gagner du temps. Pour la squelettisation nous avons utilisé l'algorithme de Zheng-Suen (Chen et al, 2012)

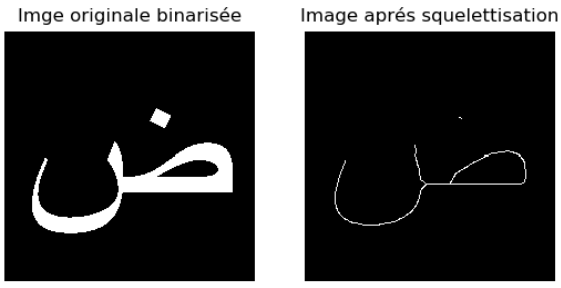


Figure 4.10: Squelettisation avec l'algorithme de Zheng-Suen

4. Pour les sous-mots dont la longueur dépasse le seuil fixé nous suivrons les étapes suivantes :

- Grâce à une projection horizontale du sous mot, l'ensemble des lignes identiques comportant le maximum de points noirs représentera la ligne de base. Le sous mot sera ainsi séparé en deux parties : partie supérieure qui constitue tout ce qui est en dessus de la ligne de base, et une partie inférieure qui représente tout ce qui en dessous de la ligne.

- Une extraction du contour du sous mot sera effectuée grâce au détecteur de Canny (Il est utilisé en traitement d'images pour la détection des contours. L'algorithme a été conçu par John Canny en 1986 pour être optimal)

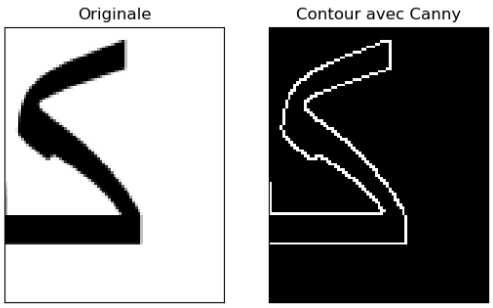


Figure 4.11: Extraction du contour avec Canny

- Généralement, les caractères se trouvant en dessous de la ligne de base représentent des caractères sous leur forme finale (à la fin du mot). Une projection horizontale sera donc faite pour délimiter les caractères qui ont une partie inférieure à la ligne de base. Ces derniers seront alors comparés aux images se trouvant dans le dossier « ending » qui contient des images de caractères dans leur forme finale en utilisant les réseaux de neurones.

- Une fois les caractères isolés et finaux trouvés, l'image contiendra que des caractères connectés entre eux.

- Les points et diacritiques seront colorés en blanc et sauvegarder pour ne garder que le corps principal de ces lettres.

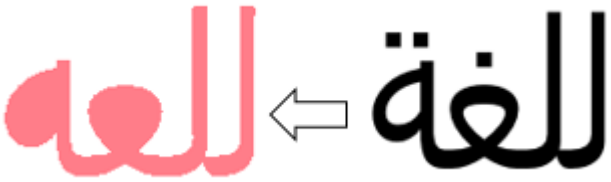


Figure 4.12: Elimination des points diacritiques

- En se servant du contour de canny et de la chaine de Freeman nous allons chercher les points locaux minimaux sur le contour de ces lettres dans le but de les projeter sur la ligne de base et de trouver des points de segmentation potentiels.

- Pour éviter une sur-segmentation, si deux points minimaux se trouvent à la même hauteur sur une distance inférieure au seuil fixé précédemment qui définit approximativement la taille du caractère, le dernier des points allant de la droite vers la gauche qui se trouvent sur la même hauteur sera retenu.

- Après la fin de ces étapes, nous obtiendrons une segmentation des caractères de chaque mot avec une reconnaissance des caractères isolés et finaux en même temps.

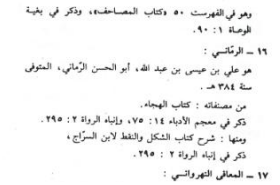
x	x	x	x	x	x
1	x	x	0	0	0
1	1	x	0	0	0
x	1	1	P	0	0
1	1	1	1	1	1

Figure 4.13 : Template utilisée lors du template matching

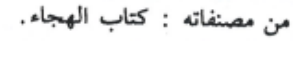
4.5.1.4 Reconnaissance des caractères isolés

Après obtention d'une segmentation en caractères isolés, les points diacritiques précédemment coloriés en blanc seront remis en place en les coloriant à nouveau en noir. Un réseau de neurones à convolution déjà disponible sera alors utilisés pour reconnaitre le caractère isolés chacun a son tour. Si le caractère ne figure pas parmi les lettres, on essayera de trouver une correspondance avec les chiffres de notre base de données, sinon un vide sera mis à la place.

Prétraitement de l'image :
Conversion en niveaux de gris
Binarisation
Lissage
Erosion et dilatation
Analyse des éléments de la page pour trouver les paragraphes
Redressement de l'écriture en cas d'inclinaison



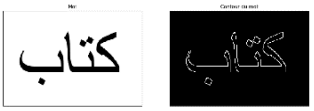
Segmentation des paragraphes en lignes en utilisant une projection verticale et stockage des lignes (avec un seuil)



Segmentation de chaque ligne en mots en utilisant une projection horizontale (avec un seuil) et obtention d'une image contenant un mot à reconnaitre

Prétraitement de l'image du mot et redimensionnement

-Obtention du contour du mot (Méthode de canny)
-Reconnaissance en utilisant un réseau de neurones à convolution (en se servant de la base de données APTI)
- Affichage du résultat

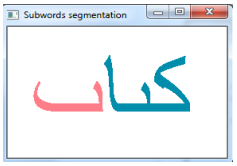


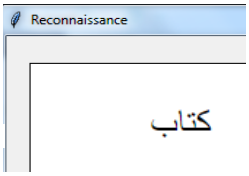
Si aucune correspondance n'a été trouvée avec les images d'APTI :

1) Segmentation du mot en sous-mots avec une projection horizontale
2) Fixer un seuil approximatif de la taille de chaque caractère
3) Extraction des caractères isolés, squelettisation et reconnaissance en se servant d'un réseau de neurones à convolution et de la base de données de caractères.
4) Estimation de la ligne de base
5) Détection des points et points diacritiques, sauvegarde de leur emplacement et coloriage des pixels qui les constituent en blanc
6) Recherche et délimitation des caractères dont une partie se situe sous la ligne de base (chaque caractère ayant une partie inférieure à la ligne d'appui représente un caractère à l'état final) à fin de les reconnaitre en se servant d'un réseau de neurones à convolution et de la base de données de caractères
7) Pour les caractères connectés qui restent :

Application du détecteur canny pour avoir le contour
Extraction du contour supérieur en utilisant l'algorithme de Freeman et le parcourir pour localiser les points de segmentation dans ses minimums locaux qui appartiennent à une zone précise.
Séparation des caractères

- Replacement et coloriage en noir des points et points diacritiques
- Utilisation d'un réseau de neurones à convolution pour reconnaitre les caractères isolés grâce à la base de données de caractères
- Affichage du résultat





Post-traitement pour améliorer les résultats de la reconnaissance

Figure 4.14 : Schéma général explicatif du fonctionnement de notre système

Dans la partie pratique de ce projet, nous avons tenté de réaliser un système hors-ligne et multi-fontes de reconnaissance optique de l'écriture arabe imprimée.
Dans ce chapitre, nous présentons l'application réalisée ainsi que les résultats obtenus de chaque étape en faisant quelques comparaisons avec d'autres systèmes.

4.6 Environnement de développement

4.6.1 Langage de développement: Python

Python est un environnement de programmation objet, multi paradigmes et multi plateformes. Il favorise la programmation impérative structurée, fonctionnelle et orientée objet. Il est doté d'un typage dynamique fort, d'une gestion automatique de la mémoire par ramasse-miettes et d'un système de gestion d'exceptions ; ce qui lui permet de combiner une puissance remarquable avec une syntaxe très claireLes bibliothèques

4.6.1.1 OpenCV 4.1.1

OpenCV (Open Source Computer Vision) est une bibliothèque proposant un ensemble de plus de 2500 algorithmes de vision par ordinateur, accessibles au travers d'API pour les langages C, C++, et Python. Elle est distribuée sous une licence BSD (libre) pour les plates-formes Windows, GNU/Linux, Android et MacOS.
Initialement écrite en langage C il y a 10 ans par des chercheurs de la société Intel, OpenCV est aujourd'hui développée, maintenue, documentée et utilisée par une communauté de plus de 40 000 membres actifs. C'est la bibliothèque de référence pour la vision par ordinateur, aussi bien dans le monde de la recherche que celui de l'industrie

4.6.1.2 Tensorflow

Développé par les chercheurs de Google, tensorflow est une bibliothèque open source, permettant de développer et d'exécuter des applications de machine Learning et de deep Learning. Cet outil permet notamment d'entraîner et d'exécuter des réseaux de neurones pour la classification de chiffres écrits à la main, la reconnaissance d'image, les plongements de

mots, les réseaux de neurones récurrents, les modèles sequence-to-sequence pour la traduction automatique, ou encore le traitement naturel du langage.

Elle a été créé en 2011, sous la forme d'un système propriétaire dédié aux réseaux de neurones de deep Learning, trensorflow s'appelait à l'origine distBelief, le code source de distBelief a été modifié et cet outils est devenue une bibliothèque basée application. Nous rappelons que les API TensorFlow peuvent être exploitées avec les langages python et C.

4.6.1.3 Matplotlib

Matplotlib est une bibliothèque graphique open source de traçage et de visualisation 2D, avec un support pour la 3D sur python. C'est à la fois une bibliothèque de haut niveau, fournissant des fonctions de visualisation de haut niveau (échelle logarithmique, histogramme, courbes de niveau, etc…), et de bas niveau, permettant de modifier tous les éléments graphiques suivantes (titres, axes, couleurs et styles des lignes, etc…). Elle est utilisable sur plusieurs systèmes d'exploitation (Unix, Mac Os, Windows, etc).

4.6.1.4 Numpy

Numpy est une bibliothèque numérique apportant le support efficace de larges tableaux multidimensionnels, et de routines mathématiques de haut niveau (fonctions spéciales, algèbre linéaire, statistiques, etc.).

4.6.1.5 Keras

Keras est une bibliothèque open source écrite en python et permettant d'interagir avec les algorithmes de réseaux de neurones profonds et de machine learning, notamment Tensorflow et Theano. Elle a été initialement écrite par François Chollet.

4.6.1.6 SciPy

SciPy est une distribution de modules destinée à être utilisée avec le langage interprété Python afin de créer un environnement de travail scientifique très similaire à celui offert par Scilab, GNU Octave, Matlab… Cette distribution offre plusieurs versions, le choix d'utiliser une version dépend des versions de : noyau algèbre linéaire, les statistiques, le traitement du signal ou encore le traitement d'images. Il offre également des possibilités avancées de visualisation grâce au module matplotlib.

4.6.1.7 Python Image Library (PIL)

PIL est une bibliothèque pour la manipulation des images. Vous retrouverez de nombreux modules qui vous permettront de traiter des images de n'importe quel format. Cette bibliothèque permet également aux interfaces graphiques construites avec Tkinter d'utiliser des images déclarées avec PIL.

4.6.1.8 Tkinter

Tkinter (de l'anglais Tool kit interface) est la bibliothèque graphique libre d'origine pour le langage Python, permettant la création d'interfaces graphiques. Elle vient d'une adaptation de la bibliothèque graphique Tk (ToolKit) écrite pour Tcl (Tool command language).

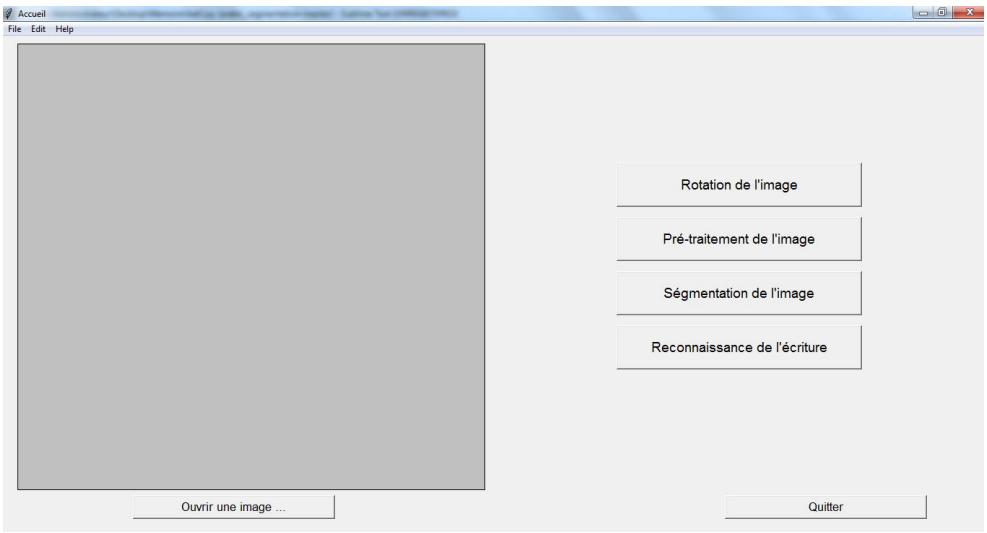
4.7 Interface graphique

Pour permettre une utilisation facile de notre système, nous avons développé une interface assez simple en langage Python. Elle offre les fonctionnalités de base dont un utilisateur risque d'avoir besoin, lors de la reconnaissance de l'écriture arabe. Quatre opérations de bases sont proposées :

- Rotation de l'image
- Prétraitement de l'image
- Segmentation de l'image
- Reconnaissance de l'écriture

Figure 4.15 : Interface d'accueil de notre OCR

Après avoir cliqué sur le bouton « ouvrir image », qui permet l'ouverture d'une image JPEG ou PNG, l'utilisateur pourra utiliser les fonctions disponibles sur notre



interface.

4.7.1 Rotation de l'image

En cliquant sur le bouton « rotation de l'image », l'utilisateur accédera à une nouvelle fenêtre ou il pourra redresser une image inclinée comme le montre la figure

4.7.2 Prétraitement de l'image

L'utilisateur pourra obtenir une image binarisée en utilisant la méthode déjà citée dans le chapitre 4 de notre travail. Le résultat peut être sauvegardé sous un format PNG.

4.7.3 Segmentation de l'image

Cette fonction permet à l'utilisateur de visionner le résultat des opérations suivantes :

- Séparation des éléments de l'image (texte, images, ..)
- Séparation des lignes des paragraphes
- Séparation des lignes en mots

Nous avons effectué des tests sur 25 images différentes, avec des fontes différentes (Arabic Transparent, AdvertisingBold, Diwani Letter, DecoType Thuluth, Simplified Arabic, Tahoma). Les fontes très anciennes n'ont pas été prises en compte lors de ces tests. Certaines images contiennent des phrases seulement et d'autres des paragraphes.

4.7.3.1 Séparation des différentes composantes de la page

Les paragraphes à reconnaître sont tous reconnus à 100% lorsque la police d'écriture est assez claire, dans le cas où elle est ancienne le taux atteint les 87% après avoir effectué des tests sur 25 images différentes.

4.7.3.2 Segmentation en lignes

Taux de segmentation en lignes : Après avoir testé une segmentation en lignes sur plusieurs images, le taux de réussite de cette opération est de 100% lorsque la fonte utilisée n'est pas très ancienne, dans le cas contraire un taux de 82% est atteint.

4.7.3.3 Segmentation en mots

Dans le cas d'une police simple le taux de segmentation en mots est de 100%. Dans le cas contraire un taux de 77% est atteint

4.7.3.4 Segmentation en sous-mots

La segmentation en sous-mots en utilisant les espaces atteint en moyenne un taux de 91% sur les images que nous avons testé jusque-là.

4.7.3.5 Segmentation en caractères

La segmentation en caractère avec la méthode que nous avons proposé évite beaucoup de problèmes de chevauchement et de sur-segmentation de caractères, sur des images simples elle peut atteindre 95% à 100%. Lorsque la police est ancienne ou très décorative ce taux peut baisser à 81%

4.8 Evaluation des performances de reconnaissance de l'écriture

Les techniques habituelles d'évaluation de résultats d'OCR supposent l'existence de textes de référence correctement transcris afin de calculer des métriques basées sur le nombre d'erreurs de reconnaissance. Malheureusement il n'y a pas de corpus de test unifié pour la langue arabe qui permettra de tester les performances de notre système.

Pour comparer notre système, nous avons tout d'abord utilisé l'image présente dans l'étude comparative de M. Alghamdi et W. Teahan dans (Alghamadi et Teahan, 2017). L'image suivante a été extraite de la base de données KAFD et a servie comme base de test pour comparer les compétences des OCR de la langue arabe les plus utilisés : Sakhr OCR, ABBYY OCR, RDI OCR et Tesseract OCR.

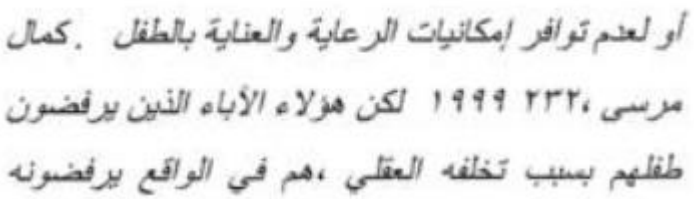

أو لعدم توافر إمكانيات الرعاية والعناية بالطفل . كمال
مرسى ،٢٣٢ ١٩٩٩ لكن هؤلاء الأباء الذين يرفضون
طفلهم بسبب تخلفه العقلي ،هم في الواقع يرفضونه

Figure 4.16: Image extraite de la base de données KAFD. Source (Alghamadi, Teahan, 2017)

Les résultats obtenus sur les autres systèmes sont les suivants :

ولعدكم تو/فر/م!نلات /لرطية و/لظية بلطفر كطلى/
مرسى ،٢٣٢ ١١١١ لكق هؤلاء /لأ؟ء/لذيق يرفضوق
طفلهم بسلبب تخلفه /لخفلي ،هم في /لو/قع يرفضؤله

(b)

أو لعدم توافر إمكانيات الرعاية والعناية بالطفل . كمال
مرسى،س ؟؟؟ر ككن هؤلا ءرردبإءسقزضرن
طفلهم بسبب تخلفه العقلى ،هم في الواقع يرفضونه

(c)

الى المستشفيات، و ٣ (سعب(وتر رمز فحتما السن ترلانتزاع عسف،عن
دنفر تستخدم تطرد تطلب لا يكاد فليات، لأغلب بكائيته ترافقنا نيته نبأ نعطا من٠
، تسلب يسعى ، لم يسء (٢ ٩ ٩ ٩ (لكن تعديلا،رجلا رع ، ونارلين، ليلب الصفرين

(d)

او ليددل٠دنا سلوكيلت الطفل العامة هير المرغوب فلها التي كترا ما يعجز
الأباء في التعامل معها بنجاح وفاعلية شاكر قنديل ،٩٣٦٠ ٦لم٩ ١ ، وقد
ترجع مثل هذه السلوكيات خير الطفل لمزيد من المشكلات النفسية

(e)

Figure 4.17: Résultat de la reconnaissance des 4 OCRs testés . (b)Sakhr OCR (c)ABBYY OCR (d) RDI OCR(e) Tesseract OCR (Source : (Alghamadi, Teahan, 2017))

Nous avons également tenté de reconnaitre cette image avec notre système :

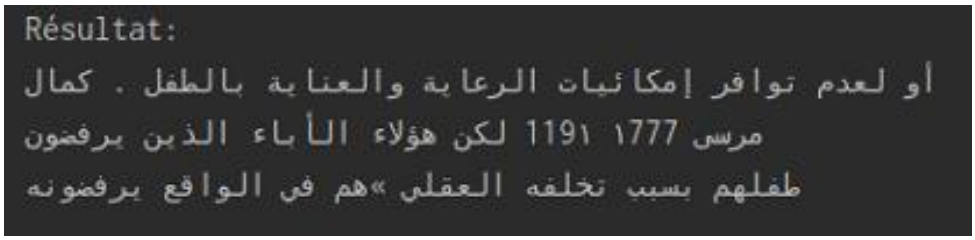


Figure 4.18 : Résultat de la reconnaissance de notre système

Le tableau représentera les résultats obtenus.

OCR	SAKHR	ABBYY	RDI	Tesseract	SRCAA
Taux de reconnaissance	83%	90%	44%	60%	94 %

Tableau 4.2 : Résultat de la reconnaissance de l'image de test obtenue par notre système

Notre système a pu obtenir un taux de 94% lors de sa comparaison avec d'autres systèmes.

Nous avons aussi effectué d'autres tests sur les 25 images qui constituent notre corpus de test, un taux de 100% peut être atteint lorsque l'image contient un texte simple avec une police non décorée. Lorsque SRCAA est utilisé sur une page de livre le taux de 84% a été atteint avec succès.

4.9 Conclusion

Dans ce travail, nous avons essayé d'implémenter au maximum nos propres idées dans les diverses étapes du système de reconnaissance. Nous nous sommes surtout concentrés sur la phase de segmentation, car la robustesse de tout système de reconnaissance dépend de son taux de réussite. Concernant la reconnaissance, les résultats sont acceptables et peuvent être optimisés en phase de post-traitement, les résultats restent tout de même très satisfaisants dans le cas où le document ne contient pas trop d'ambiguïtés ou une police d'écriture compliquée.

Notre système reste limité suite au peu de temps qu'on a eu pour former la base de données sur laquelle repose la reconnaissance, une base de données plus enrichie donnerai certainement de meilleurs résultats.

Conclusion et perspectives

Malgré les efforts et les travaux intensifs réalisés dans le domaine de la reconnaissance optique de l'écriture, aucun système OCR n'est jugé fiable à 100%. Les problèmes majeurs influençant la recherche en AOCR (Arabic Optical Character Recognition) sont le manque de normalisation des calligraphies des caractères arabes, l'absence d'outils tels que dictionnaires, bases de données et statistiques se rapportant à l'écriture arabe.

Dans le cadre de notre travail, un système pour la reconnaissance hors-ligne de la langue arabe est développé dans le but d'améliorer le taux de reconnaissance des caractères arabes manuscrits, en apportant certaines contributions à différents niveaux. Les approches développées ont été étudiées, dans le but, d'améliorer les performances du système de reconnaissance de caractères en termes de vitesse et d'exactitude

Dans un premier temps, notre étude bibliographique a présenté les algorithmes et les méthodes statiques qu'un système de reconnaissance utilise en général. Nous avons aussi présenté la langue arabe et toutes ses caractéristiques de langue cursive qui la rendent plus complexe.

Dans un second temps, nous avons présenté notre contribution à différents niveaux, nous avons tout d'abord commencé par utilisé un ensemble d'opération pour un bon prétraitement qui ne dégrade pas la qualité et les informations contenues par l'image. Nous avons par la suite proposé un algorithme de reconnaissance qui repose sur la reconnaissance des mots par leur contour. Mais vu l'immensité du vocabulaire arabe, la reconnaissance des mots ne peut pas toujours donner des résultats satisfaisant, c'est pour cela qu'on a proposé une méthode de segmentation de caractères isolés qui combine un certain nombre de techniques existantes pour surmonter le problème d'irrégularité, de chevauchements et des intersections dans l'écriture manuscrite arabe.

Contrairement à la plupart des systèmes qui utilisent des méthodes statiques lors de la classification, La reconnaissance de caractères à base des réseaux de neurones, offre une solution plus avancée par rapport aux méthodes que nous avons exposées. Ils offrent des améliorations en termes du temps de réponse, voire, de qualité de classification quand elles

sont acquises à partir des méthodes d'apprentissages efficaces. Néanmoins, ils prennent un temps d'apprentissage considérable mais une fois l'apprentissage terminé, les meilleurs poids trouvés seront enregistrés pour un usage rapide. Dans notre cas, un réseau de neurones à convolution a été utilisé et les résultats ont été satisfaisants.

Bien que les idées implémentées dans ce travail résultent d'une étude approfondie de l'état de l'art, le système proposé ne présente ni continuité ni adaptation de n'importe quel autre système; ce qui explique en grande partie notre incapacité de compléter l'étape de post-traitement en raison du temps limité. Toutefois, grâce à l'introduction de ces propositions, notre système donne des résultats très encourageants, dans ce qui concerne les performances et la complexité.

Cependant, plusieurs tâches restent encore à achever à différents niveaux, nous citerons comme perspectives :

- Agrandir notre base de données de caractères isolés. L'intégration d'autres images de caractères isolés avec différentes fontes et tailles permettra l'obtention de meilleurs résultats lors de la segmentation et de la reconnaissance.
- Munir notre système d'une phase de post-traitement du résultat obtenu pour réduire encore plus le taux d'erreurs de reconnaissance
- Amélioration de notre réseau de neurones pour réduire le temps d'exécution encore plus
- Amélioration de l'interface utilisateur, pour apporter de nouvelles fonctionnalités et une facilité d'utilisation de notre système.

Finalement, nous dirons qu'il nous reste encore du travail à faire pour achever le développement du système proposé afin d'améliorer encore plus le taux de reconnaissance de la langue arabe pour de meilleurs résultats.

Bibliographies

AL-BADAWI, al-Said. Mustwayat al-Arabiyya al-muasira fi Misr. Levels of contemporary Arabic in Egypt), Cairo: Dar al-Maarif, 1973.

AL-BADR, Badr et MAHMOUD, Sabri A. Survey and bibliography of Arabic optical text recognition. Signal processing, 1995, vol. 41, no 1, p. 49-77.

AL-EMAMI, Samir Yaseen Safa. Machine recognition of handwritten and typewritten Arabic characters. 1988. Thèse de doctorat. University of Reading.

ALGHAMDI, Mansoor et TEAHAN, William. Experimental evaluation of Arabic OCR systems. PSU Research Review, 2017, vol. 1, no 3, p. 229-241.

ALQAISY, E. K. et NASER, H. L. Recognition of Arabic numerals using probabilistic functions. In : Proc. of Computer Processing and Transmission of the Arabic Language Workshop. 1985.

AL-SADOUN, Humoud B. et AMIN, Adnan. A new structural technique for recognizing printed Arabic text. International journal of pattern recognition and artificial intelligence, 1995, vol. 9, no 01, p. 101-125.

AL-YOUSEFI, H. et UPDA, S. S. Recognition of Arabic characters. IEEE Transactions on Pattern Analysis & Machine Intelligence, 1992, no 8, p. 853-857.

AMEISEN, David, LE NAOUR, Gilles, et DANIEL,Christel. Technologie des lames virtuelles-De la numérisation à la mise en ligne. médecine/sciences, 2012, vol. 28, no 11, p. 977-982.

AMIN, Adnan et AL-SADOUN, Humoud B. Hand printed Arabic character recognition system. In : Proceedings of the 12th IAPR International Conference on Pattern Recognition, Vol. 3-Conference C: Signal Processing (Cat. No. 94CH3440-5). IEEE, 1994. p. 536-539.

AMIN, Adnan et MARI, Jean F. Machine recognition and correction of printed Arabic text. IEEE Transactions on systems, man, and cybernetics, 1989, vol. 19, no 5, p. 1300-1306

ARICA, Nafiz et YARMAN-VURAL, Fatos T. An overview of character recognition focused on off-line handwriting. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 2001, vol. 31, no 2, p. 216-233.

BACCOUCHE, Taieb. La langue arabe dans le monde arabe. L'information grammaticale, 1998, vol. 2, no 1, p. 49-54.

BERNSEN, John. Dynamic thresholding of gray-level images. In: Proc. Eighth Int'l conf. Pattern Recognition, Paris, 1986. 1986.

BOUKERMA, Hanene. Combinaison de classifieurs flous pour la reconnaissance de l'écriture arabe manuscrite. 2010. Thèse de doctorat. Master Thesis.

BOUREZG, Aissa. Implémentation d'une méthode d'analyse morpho-lexicale pour la langue arabe basée sur la position des lettres. 2017. Thèse de doctorat. FACULTE DES MATHEMATIQUES ET DE L'INFORMATIQUE-UNIVERSITE MOHAMED BOUDIAF-M'SILA.

BOUSLAMA, Faouzi et KISHIBE, Hiroki. Fuzzy Logic in the recognition of machine printed Arabic characters. In : ICONIP'99. ANZIIS'99 & ANNES'99 & ACNN'99. 6th International Conference on Neural Information Processing. Proceedings (Cat. No. 99EX378). IEEE, 1999. p. 1150-1154.

BOZINOVIC, Radmilo M et SRIHARI, Sargur N. Off-line cursive script word recognition. IEEE Transactions on Pattern Analysis & Machine Intelligence, 1989, no 1, p. 68-83.

BOZINOVIC, Radmilo M. et SRIHARI, Sargur N. Off-line cursive script word recognition. IEEE Transactions on Pattern Analysis & Machine Intelligence, 1989, no 1, p. 68-83.

BRITTO, A. S., SABOURIN, Robert, LETHELIER, Edouard, et al. Improvement in handwritten numeral string recognition by slant normalization and contextual information. In : Proc. 7th IWFHR. 2000. p. 323-332.

BUSHOFA, B. M. F. et SPANN, M. Segmentation of Arabic characters using their contour information. In : Proceedings of 13th International Conference on Digital Signal Processing. IEEE, 1997. p. 683-686.

BUSHOFA, B. M. F. et SPANN, Michael. Segmentation and recognition of Arabic characters by structural classification. Image and Vision Computing, 1997, vol. 15, no 3, p. 167-179

CARBONNEL, Sabine. Intégration et modélisation de connaissances linguistiques pour la reconnaissance d'écriture manuscrite en-ligne. 2005. Thèse de doctorat. Rennes, INSA.

CHAKIB, Kaddour et SALI, Aissa Brahim. La Compression des Images Fixes par les Approximations Fractales Basée sur la Triangulation de Delaunay et la quantification Vectorielle. Mémoire de fin d'étude, 1999.

CHAKIB, M. Généralités sur le traitement d'imagesarticle.1999

CHEN, Wei, SUI, Lichun, XU, Zhengchao, et al. Improved Zhang-Suen thinning algorithm in binary line drawing applications. In : 2012 International Conference on Systems and Informatics (ICSAI2012). IEEE, 2012. p. 1947-1950.

CHENG, H. D., CHEN, Jim-Rong, et LI, Jiguang. Threshold selection based on fuzzy c-partition entropy approach. Pattern recognition, 1998, vol. 31, no 7, p. 857-870.

CHERIET, Mohamed, KHARMA, Nawwaf, LIU, Cheng-Lin, et al. Character recognition systems: a guide for students and practitioners. John Wiley & Sons, 2007.

DING, Yimei, KIMURA, Fumitaka, MIYAKE, Yasuji, et al. Accuracy improvement of slant estimation for handwritten words. In: Proceedings 15th International Conference on Pattern Recognition. ICPR-2000. IEEE, 2000. p. 527-530.

DOUZIDIA, Fouad Soufiane. Résumé automatique de texte arabe. 2005.

EL KASSAS, Dina. Une étude contrastive de l'arabe et du français dans une perspective de génération multilingue. 2005. Thèse de doctorat. Paris 7.

EL-KHALY, Fatma et SID-AHMED, Maher A. Machine recognition of optically captured machine printed Arabic text. Pattern recognition, 1990, vol. 23, no 11, p. 1207-1214.

EL-SHEIKH, Talaat S. et GUINDI, Ramez M. Computer recognition of Arabic cursive scripts. Pattern Recognition, 1988, vol. 21, no 4, p. 293-302.

FARGHALY, Ali et SHAALAN, Khaled. Arabic natural language processing: Challenges and solutions. ACM Transactions on Asian Language Information Processing (TALIP), 2009, vol. 8, no 4, p. 14.

FERGUSON, Charles Albert. Structuralist studies in Arabic linguistics: Charles A. Ferguson's papers, 1954-1994. Brill, 1997.

FREEMAN, Herbert. On the encoding of arbitrary geometric configurations. IRE Transactions on Electronic Computers, 1961, no 2, p. 260-268.

FRUITET, Jean. Outils et méthodes pour le traitement des images par ordinateur. Université de Marne-La-Vallée-UMLV B, 2009, vol. 202.

GAUDIN, Jacques. Qu'est-ce qu'une image numérique? . 2002.

GORAINE, Habib, USHER, Mike, et AL-EMAMI, Samir. Off-line Arabic character recognition. Computer, 1992, no 7, p. 71-74.

GOVINDAN, V. K. et SHIVAPRASAD, A. P. Character recognition—a review. Pattern recognition, 1990, vol. 23, no 7, p. 671-683

GRANDIDIER, Frédéric, SABOURIN, Robert, et SUEN, Ching Y. Quelques techniques pour l'amélioration du pouvoir discriminant de primitives discrètes. In : Conférence Internationale Francophone sur l'Ecrit et le Document (CIFED 04). 2004.

GRANDIDIER, Frédéric. Un nouvel algorithme de sélection de caractéristiques: application à la lecture automatique de l'écriture manuscrite. 2003. Thèse de doctorat. École de technologie supérieure.

GUPTA, Maya R., JACOBSON, Nathaniel P., et GARCIA, Eric K. OCR binarization and image pre-processing for searching historical documents. Pattern Recognition, 2007, vol. 40, no 2, p. 389-397.

HAITAAMAR, Schahrazed. Segmentation de textes en caractères pour la reconnaissance optique de l'écriture arabe. 2007. Thèse de doctorat. Université de Batna 2.

HULL, Jonathan J. Document image skew detection: Survey and annotated bibliography. In : Document Analysis Systems II. 1998. p. 40-64.

ISDANT, R. Traitement numérique de l'image. 2009.

JOHNSTONE, Thomas Muir. Salih J. Altoma: The problem of diglossia in Arabic: a comparative study of Classical and Iraqi Arabic.(Harvard Middle Eastern Monographs, xxi.) ix, 167 pp. Cambridge, Mass.: Center for Middle Eastern Studies, Harvard University, 1969.(Distributed by Harvard University Press.

Distributed in GB by Oxford University Press. 34s.). Bulletin of the School of Oriental and African Studies, 1970, vol. 33, no 3, p. 619-620.

KANUNGO, Tapas et HARALICK, Robert M. Character recognition using mathematical morphology. In : Proc. of the Fourth USPS Conference on Advanced Technology. 1990. p. 973-986.

KELAIAIA, Abdessalem. Classification non supervisee de textes arabes appliquee a la recherche documentaire. 2010. Thèse de doctorat.

KHELLA, Fakhry. Analysis of hexagonally sampled images with application to Arabic cursive text recognition. 1992. Thèse de doctorat. M. Phil. Thesis, University of Bradford, Bradford, England, UK.

KHORSHEED, Mohammad S. et CLOCKSIN, William F. Structural Features of Cursive Arabic Script. In : BMVC. 1999. p. 1-10.

KHORSHEED, Mohammad S. Off-line Arabic character recognition–a review. Pattern analysis & applications, 2002, vol. 5, no 1, p. 31-45.

KHORSHEED, Mohammad S. Off-line Arabic character recognition–a review. Pattern analysis & applications, 2002, vol. 5, no 1, p. 31-45.

KHURSHID, Khurram, SIDDIQI, Imran, FAURE, Claudie, et al. Comparison of Niblack inspired Binarization methods for ancient documents. In : Document Recognition and Retrieval XVI. International Society for Optics and Photonics, 2009. p. 72470U.

LAM, L., SUEN, C. Y., GUILLEVIC, D., et al. Automatic processing of information on cheques. In : IEEE INTERNATIONAL CONFERENCE ON SYSTEMS MAN AND CYBERNETICS. INSTITUTE OF ELECTRICAL ENGINEERS INC (IEEE), 1995. p. 2353-2358.

LAM, L., SUEN, C. Y., GUILLEVIC, D., et al. Automatic processing of information on cheques. In : IEEE INTERNATIONAL CONFERENCE ON SYSTEMS MAN AND CYBERNETICS. INSTITUTE OF ELECTRICAL ENGINEERS INC (IEEE), 1995. p. 2353-2358.

LASRI, Y. Contribution à la reconnaissance optique (OCR) du texte arabe imprimé. 2014.

LEEDHAM, Graham, VARMA, Saket, PATANKAR, Anish, et al. Separating text and background in degraded document images-a comparison of global thresholding techniques for multi-stage thresholding. In: Proceedings Eighth International Workshop on Frontiers in Handwriting Recognition. IEEE, 2002. p. 244-249.

LIU, Zhi-Qiang, CAI, Jin-Hai, et BUSE, Richard. Handwriting recognition: soft computing and probabilistic approaches. Springer, 2012.

LORIGO, Liana M. et GOVINDARAJU, Venugopal. Offline Arabic handwriting recognition: a survey. IEEE transactions on pattern analysis and machine intelligence, 2006, vol. 28, no 5, p. 712-724.

MAALOUL, Mohamed Hedi. Approche hybride pour le résumé automatique de textes. Application à la langue arabe. 2012. Thèse de doctorat.

MADHVANATH, Sriganesh, KIM, Gyeonghwan, et GOVINDARAJU, Venu. Chaincode contour processing for handwritten word recognition. IEEE transactions on pattern analysis and machine intelligence, 1999, vol. 21, no 9, p. 928-932.

MARAOUI, Mohsen, ANTONIADIS, George, et ZRIGUI, Mounir. Un système de génération automatique de dictionnaires étiquetés de l'arabe. CITALA 2007, 2007, p. 18-19.

MENASRI, Farès. Contributions à la reconnaissance de l'écriture arabe manuscrite. UNIVERSITE PARIS DESCARTES, Thèse de doctorat, 2008.

MILED, H., CHERIET, Mohamed, OLIVIER, Claude, et al. Modélisation markovienne de l'écriture arabe manuscrite: une approche analytique. 1998.

MOHIUDDIN, K. M. et MAO, Jianchang. A comparative study of different classifiers for handprinted character recognition. In: Machine Intelligence and Pattern Recognition. North-Holland, 1994. p. 437-448

MOUSSA Richard. Segmentation Multi-Agents en Imagerie Biologique et Médicale: Application aux IRM 3D. 2011. Thèse de doctorat.

NAWAZ, Syed Nazim, SARFRAZ, Muhammad, ZIDOURI, A., et al. An approach to offline Arabic character recognition using neural networks. In: 10th IEEE International Conference on Electronics, Circuits and Systems, 2003. ICECS 2003. Proceedings of the 2003. IEEE, 2003. p. 1328-1331.

NIBLACK, Wayne. An introduction to digital image processing. Strandberg Publishing Company, 1985.

OLIVEIRA, L. Soares, LETHELIER, Edouard, BORTOLOZZI, Flâvio, et al. Segmentation de caractères manuscrits basée sur une approche structurelle. In : Colloque International Francophone sur l'Écrit et le Document, CIFED'2000. 2000. p. 231-240.

OLIVIER, G., MILED, Housem, ROMEO, K., et al. Segmentation and coding of Arabic handwritten words.In: Proceedings of 13th International Conference on Pattern Recognition. IEEE, 1996. p. 264-268.

OTSU, Nobuyuki. A threshold selection method from gray-level histograms. IEEE transactions on systems, man, and cybernetics, 1979, vol. 9, no 1, p. 62-66.

OULED-DIAF, MONCEF. Approche Analytique Pour La Reconnaissance Hors-Ligne De l'Ecriture Arabe Manuscrite. 2008. Thèse de doctorat.

PLAMONDON, Réjean et SRIHARI, Sargur N. Online and off-line handwriting recognition: a comprehensive survey. IEEE Transactions on pattern analysis and machine intelligence, 2000, vol. 22, no 1, p. 63-84.

REDOUANE, T. L. E. Reconnaissance des Formes-Intelligence Artificielle (RF-IA. 2012. Thèse de doctorat. usto.

ROMAN, A. Grammaire de l'arabe. Presses universitaires de France, 1990.

SAADANE, Houda, GUIDERE, Mathieu, et FLUHR, Christian. La reconnaissance automatique des dialectes arabes à l'écrit. In : colloque international «Quelle place pour la langue arabe aujourd'hui. 2013. p. 18-20.

SAADIA, Baza Halima. La reconnaissance d'écriture arabe manuscrite.2015.

SAUVOLA, Jaakko et PIETIKÄINEN, Matti. Adaptive document image binarization. Pattern recognition, 2000, vol. 33, no 2, p. 225-236

SHAFAIT, Faisal, KEYSERS, Daniel, BREUEL, Thomas M., *et al*. Layout analysis of Urdu document images. In : *2006 IEEE International Multitopic Conference*. IEEE, 2006. p. 293-298.

SOUICI .L,ZMIRLI .Z, SELLAMI M. : « Système connexionniste pour lareconnaissance de l'arabe manuscrit ». 1ères journées scientifiques et techniques (JST FRANCIL), pp. 383-388, Avignon, France, 1997

SOUICI-MESLATI, L. Reconnaissance de mots arabes manuscrits par intégration neuro-symbolique. 2006. Thèse de doctorat. PhD Thesis, Badji Mokhtar University, Annaba, Algeria.

TOUFIK, Lakhdari Ahmed et SALIH, Abbaci. La reconnaissance des caractères arabes manuscrits par les réseaux des neurones convolutionnels. Université Kasdi Merbah Ouargla, 2017.

TRIER, Oeivind Due et TAXT, Torfinn. Evaluation of binarization methods for document images. IEEE transactions on pattern analysis and machine intelligence, 1995, vol. 17, no 3, p. 312-315.

TRIER, Øivind Due, JAIN, Anil K., et TAXT, Torfinn. Feature extraction methods for character recognition-a survey. Pattern recognition, 1996, vol. 29, no 4, p. 641-662.

VELASCO, Flavio R. Thresholding using the ISODATA clustering algorithm. MARYLAND UNIV COLLEGE PARK COMPUTER SCIENCE CENTER, 1979.

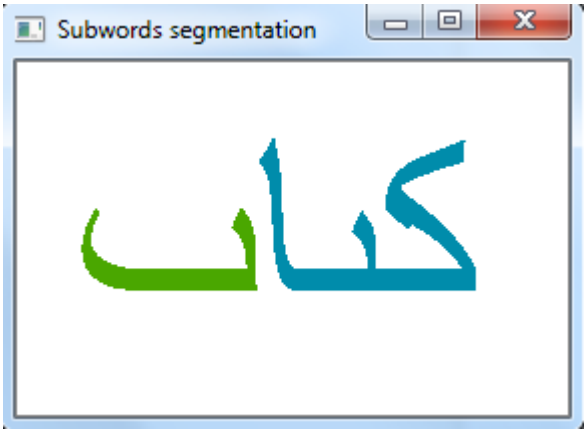
VERMA, Brijesh. A contour code feature based segmentation for handwriting recognition. In : Seventh International Conference on Document Analysis and Recognition, 2003. Proceedings. IEEE, 2003. p. 1203-1207.

YIN, Wenpeng, KANN, Katharina, YU, Mo, et al. Comparative study of CNN and RNN for natural language processing (2017). arXiv preprint arXiv:1702.01923, 2017.

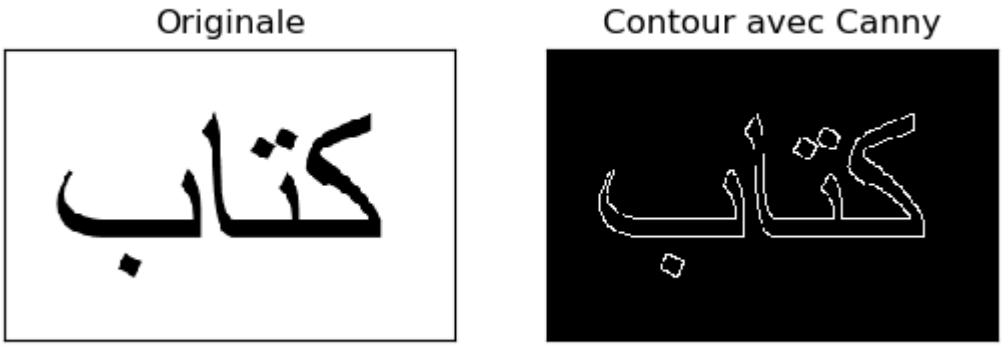
ZHANG, T. Y. et SUEN, Ching Y. A fast parallel algorithm for thinning digital patterns. Communications of the ACM, 1984, vol. 27, no 3, p. 236-239.

ZHENG, Liying, HASSIN, Abbas H., et TANG, Xianglong. A new algorithm for machine printed Arabic character segmentation. Pattern Recognition Letters, 2004, vol. 25, no 15, p. 1723-1729.

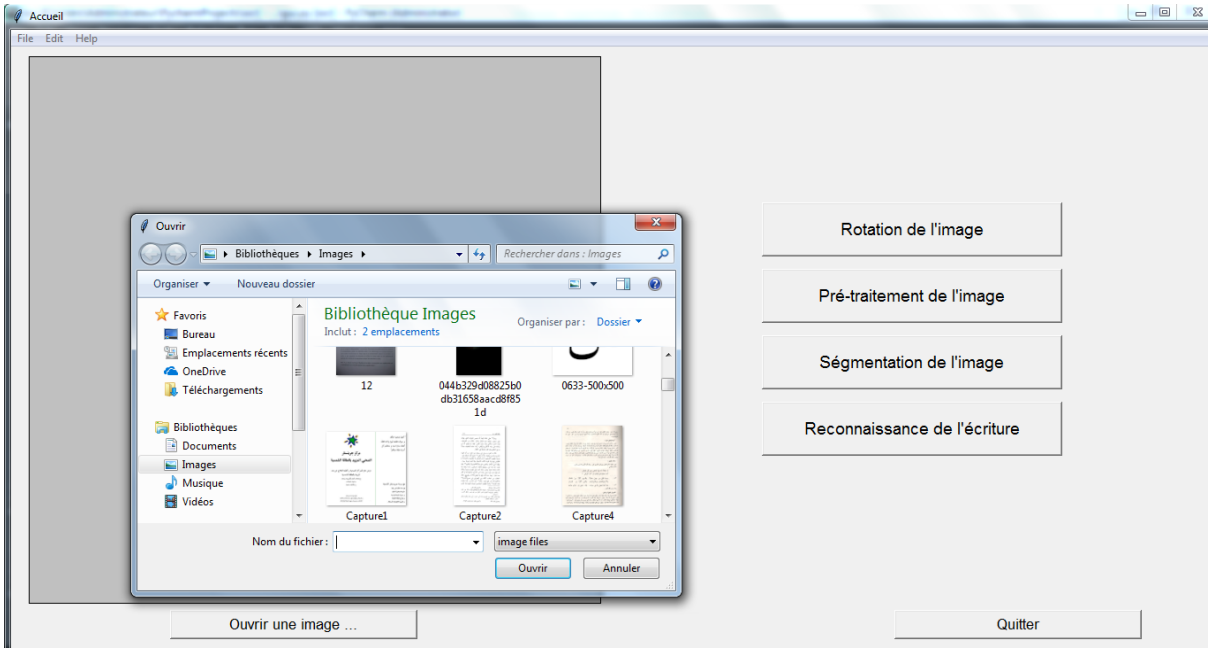
Annexe



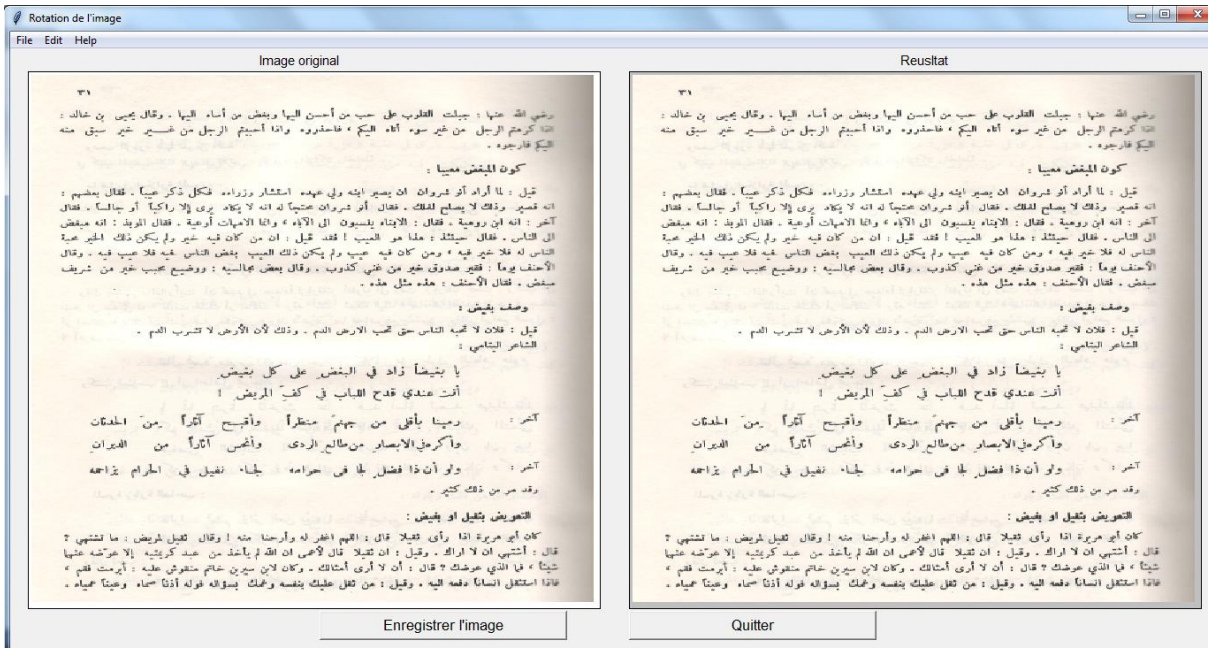
Segmentation en sous mot avec extraction des caractéristiques



Contour d'un mot entier avec la méthode de Canny implémentée en Python



Interface d'accueil



Fenêtre de rotation des pages

ain	Dossier de fichiers	25/07/2019 15:21
alif	Dossier de fichiers	29/08/2019 19:53
baa	Dossier de fichiers	25/07/2019 20:31
dal	Dossier de fichiers	25/07/2019 20:59
dhaal	Dossier de fichiers	05/08/2019 15:02
fa	Dossier de fichiers	29/07/2019 20:52
ghain	Dossier de fichiers	25/07/2019 16:08
ha	Dossier de fichiers	31/07/2019 21:34
hamza	Dossier de fichiers	29/08/2019 21:42
heh	Dossier de fichiers	25/07/2019 23:53
jim	Dossier de fichiers	31/07/2019 21:35
kaf	Dossier de fichiers	25/07/2019 20:11
kha	Dossier de fichiers	05/08/2019 14:48
lam	Dossier de fichiers	07/08/2019 18:39
Mim	Dossier de fichiers	07/08/2019 18:39
noon	Dossier de fichiers	25/07/2019 01:18
qaf	Dossier de fichiers	25/07/2019 20:10
ra	Dossier de fichiers	03/08/2019 21:24
sad	Dossier de fichiers	25/07/2019 23:55
sheen	Dossier de fichiers	25/07/2019 15:19
sin	Dossier de fichiers	25/07/2019 16:03
ta	Dossier de fichiers	31/07/2019 21:45
taa	Dossier de fichiers	07/08/2019 18:35
tha	Dossier de fichiers	31/07/2019 21:48
thaa	Dossier de fichiers	31/07/2019 22:03
thad	Dossier de fichiers	07/08/2019 18:33
waw	Dossier de fichiers	05/08/2019 14:56

Répertoire de quelques dossiers de caractères isolés de notre base de données

arabic_letter_yeh_with_hamza_above_medial_fo...	2 517	2 129	Image PNG	25/07/2019 13:43	0C514D59
final.png	1 844	1 844	Image PNG	25/07/2019 13:43	5CC5874B
final1.png	686	621	Image PNG	25/07/2019 13:44	57A66C2F
initial.png	5 996	4 705	Image PNG	29/08/2019 21:41	83F54699
initial1.png	5 998	4 713	Image PNG	29/08/2019 21:41	A36BEC70
isolated.png	5 887	4 239	Image PNG	29/08/2019 21:40	3502E7DB
medial.png	4 358	4 055	Image PNG	25/07/2019 13:47	97443FC2
medial1.png	2 461	2 378	Image PNG	25/07/2019 13:48	7B73360F
téléchargé.png	730	720	Image PNG	25/07/2019 13:42	28AFC2DA

Images des différentes positions et formes de la Hamza dans un mot