République Algérienne Démocratique et Populaire

Ministère de l'Enseignement Supérieure et de la Recherche Scientifique

Université Ahmed Draia - Adrar

Faculté des Sciences et de la Technologie

Département des Mathématiques et Informatique

Mémoire de fin d'étude, en vue de l'obtention du diplôme de Master en informatique

Option : Réseaux et Systèmes Intelligents

Thème

Réalisation d'un système de reconnaissance de l'écriture arabe manuscrite à base d'automates à états finis pondérés.

Préparés par

Safia KHERRASSI et Khayra BARKA.

Soutenu le 17 mai 2017, devant le jury composé de:

Mr. KOHILI Mohammed	Encadreur
Dr. OMARI Mohammed	Président
Mr. MAMOUNI El mamoun	Examinateur
Mr. CHOUGUEUR Djilali	Examinateur

Année Universitaire 2016/2017

DEDICACE

Si je suis arrivé là c’est grâce à Dieu.

Et,j' indique la bonne voie A: mon Père, A ma Mère,A tous la famille

Mes tantes et mes oncles A ceux qui ont fait preuve de soutiens,

et qui m'ont donné une motivation sans prix: mes amis surtout Latifa, Zohra,

,Yamina, Fatima.

SAFIA

.

DEDICACE

A mes parents

A tout ma famille

khayra

Remerciment

Avant tout, Nous remercions Dieu le très haut qui nous avons données le courage et la volonté de réaliser ce modeste travail.

Nous remercions le seigneur tout puissant de nous avoir accordé

volonté et patience dans l'accomplissement de ce travail à terme.

Nous tenons à exprimer nos vifs remerciements à

Mr. KOHILI Mohammed". "

Nous remercions « TOUS » les Messieurs et dames, les professeurs pour leurs précieux conseils.

Notre remerciements vont également aux membres du jury d'avoir accepté d'évaluer mon travail.Sans oublier de remercier mes amis.

Sans oublier bien sûr de remercier profondément tous ceux qui ont contribué de près ou de loin à la réalisation de ce travail.

Ssfia et khayra

Résumé:

On présente une approche de reconnaissance d'écriture manuscrite arabe à partir d'automate pondéré et fondée sur une modélisation par les matrices poids-position PSSM (position specific scoring matrices en anglais). Son originalité réside dans la combinaison d'une analyse fenêtrée de l'image et d'une modélisation matricielle qui permet une présentation optimal de ces fenêtrée. Un aspect important de ces travaux est la méthodologie de développement employée qui est centrée sur l'évaluation systématique des apports algorithmiques et des paramètres utilisés. Ces modélisations sont en partie empruntées aux techniques utilisées dans le domaine de la détection des protéines éloignées.

L'approche proposée est validée sur des bases de données standard et librement disponibles. L'application de cette méthode extrêmement générique à une tâche de reconnaissance de lettres manuscrites a permis d'obtenir des résultats comparables à ceux de l'état de l'art.

Mots-clés: automates à état fini pondéré, la reconnaissance des formes, caractéristiques structurelles, d'identifier des caractères arabes.

ملخص:

صور لمعالجة الأحرف العربية على التعرف تُحقق فَعّالة خوارزمية تطوير عملية في جديدة مساهمة البحث هذا يقدم وحل لمواجهة ,والخوارزميات ,الجديدة بعض الأفكار تضمن البحث وقد,(طريقة العمل) باستخدام النصية المستندات ,الأنماط على التعرف طرائق حول مرجعية دراسة تقديم سياق العمل في تم كما العربية الكلمات في تقطيع البالغ التعقيد على التركيز مع الصور معالجة بعلم الاهتمام إلى بالإضافة ,باللغة العربية الكتابة على للتعرف المستخدمة والخوارزميات طريقة العمل.

نهاية في التوصل عليه و التعرف د ا المر استخراج عناصر الحرف أو السمات البنيوية للحرف تم الاعتماد على طريقة والاستنتاجات النتائج من مجموعة إلى البحث.

المفتاحية الكلمات: العربية الأحرف على التعرف ,البنيوية السمات ,الأنماط على التعرف , طريقة العمل.

Sommaire

Chapitre 01: traitement d'image

Chapitre 02: La reconnaissance des manuscrit arabe

Chapitre 04: Application et résultats

Liste des tableaux

Chapitre 2

Chapitre 4

Liste des figures

Chapitre 1

Chapitre 2

Chapitre3

Chapitre4

.

Introduction générale:

La compréhension de l'écriture par un ordinateur est encore loin d'être pleinement satisfaisante. La raison est liée au fait que l'étude de la reconnaissance de l'écriture est un domaine très vaste tant par ses applications que par ses techniques. De plus, de grandes différences de complexité existent entre les problèmes de reconnaissance de caractères, de mots ou de documents imprimés ou manuscrits, plus ou moins dégradés.

Et malgré ça, des progrès considérables ont été réalisés dans le domaine de la reconnaissance de l'écriture manuscrite. Ce progrès est du d'une part aux nombreux travaux effectués dans ce domaine et d'autre part a la disponibilité de bases de données internationales standards relatives à l'écriture manuscrite qui permettait aux chercheurs de rapporter de façon crédible les performances de leurs approches dans ce domaine, avec la possibilité de les comparer avec d'autres approches vu qu'ils utilisent les mêmes bases.

La langue arabe n'a pas eu cette chance, contrairement au latin, elle reste encore au niveau de la recherche et de l'expérimentation, c'est-à-dire que le problème reste encore un pari ouvert pour les chercheurs. L'écriture arabe étant par nature cursive, elle pose de nombreux problèmes aux systèmes de reconnaissance automatique.

Le problème le plus difficile lors de la conception d'un système de reconnaissance de l'écriture manuscrite est la segmentation des mots manuscrits en vue de leur reconnaissance, qui n'est pas toujours triviale et demande beaucoup de temps et de calcul. D'autre part, les informations locales sont un peu négligées dans les systèmes se basant sur une analyse globale ce qui peut diminuer considérablement leurs performances. Pour remédier à ces problèmes, Notre travail consiste à la conception d'un système de reconnaissance d'écriture manuscrite arabe dans un vocabulaire limité, c'est une approche basé sur les automates pondérés fondée sur une modélisation par les matrices poids-position PSSM (*position specific scoring matrices*). Pour réaliser ce travail on a 4 chapitres a présenté:

Après que nous arborons une généralité sur l'image numérique dans le chapitre 1. Nous présenterons au chapitre 2 les techniques et algorithmes classiques de la reconnaissance de l'écriture manuscrite, ainsi qu'un aperçu des méthodes de segmentation utilisées. Au chapitre 3, nous détaillerons les concepts d'automate pondéré mises en œuvre au cours de notre recherche qu'à la modélisation par les matrices poids-position PSSM (*position specific scoring matrices).* Le chapitre 4 d´écrira nos principales contributions sur l'adaptation et l'implantation de ces techniques à une tâche de reconnaissance de caractères manuscrits.

Chapitre 01
traitement d'image

1.1. Introduction

Depuis longtemps, le public sait que: « Une image vaut mieux que mille mots », combinée avec la parole, l'image constitue un moyen essentiel dans la communication homme-machine « Une vidéo vaut mille phrases ». De ce fait, le traitement d'image est devenu une discipline nécessaire, mettant en place un ensemble de techniques permettant d'extraire des informations, de modifier une image numérique (dans le but de l'améliorer), et d'automatiser son traitement. En résumé cette discipline permet d' « Apprendre à voir aux machines ».

Le domaine de traitement d'image a vigoureusement évolué, et ses techniques sont actuellement utilisées pour résoudre une variété de problèmes pour lesquels on adapte des solutions selon la nature, les situations et les objectifs à atteindre. [1]

Dans ce chapitre en parle à les notions de base d'image ensuite en parle à les techniques de traitement exact filtrage et la segmentation

1.2. L'image

Une image est une représentation planaire d'une scène ou d'un objet situé en général dans un espace tridimensionnel, elle est issue du contact des rayons lumineux provenant des objets formants la scène avec un capteur (caméra, scanner, rayons X, …). Il ne s'agit en réalité que d'une représentation spatiale de la lumière.

L'image est considérée comme un ensemble de points auquel est affecté une grandeur physique (luminance, couleur). Ces grandeurs peuvent être continues (image analogique) ou bien discrètes (images digitales). [1]

1.2.1. Image numérique

L'image numérique est l'image dont la surface est divisée en élément de taille fixe appelée cellules ou pixels. [2]

1.2.2. Types des images

a) Image binaire : L'image binaire est l'image la plus simple, ou un pixel peut prendre uniquement les valeurs noir et blanc, tell que le blanc est représenté par 1et le noir par 0[1].

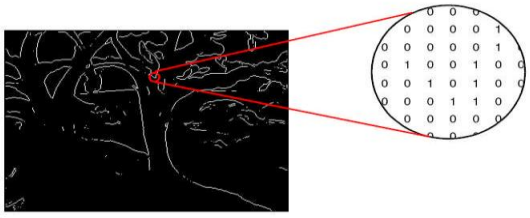


Figure 1.1: *image binaire*

b) L'image en niveau de gris : C'est la valeur de l'intensité lumineuse d'un pixel. Cette valeur peut aller du noir (0) jusqu'au blanc (255) en passant par les nuances qui sont contenues dans l'intervalle [0, 255] [2].

c) Image couleur : L'espace couleur est basé sur la synthèse additive des couleurs. C'est-à-dire le mélange de trois composants (R.V.B : rouge, vert, bleu) [2].

1.3. Le filtrage

Le filtrage est un traitement local utilisé principalement pour réaliser une analyse spatiale de l'image, son objectif est d'accentuer les variations d'intensité de l'image, ou de détecter les contours ou de réduire les bruits existants. Il existe un grand nombre de filtres possibles, on peut les classer en deux grandes catégories: les filtres linéaires et filtres non linéaires [1].

1.3.1. Types de filtrage

On distingue généralement les types de filtres suivants :

a) Filtre passe-bas

L'intensité du pixel considéré est remplacée par la moyenne des pixels de son voisinage. La taille de la zone (fenêtre) entourant le pixel est un paramètre important, plus cette dimension est grande, plus sa sensibilité au bruit diminue, et le lissage devient important (le flou s'accentue).

Le filtre passe-bas laisse passer les basses fréquences (les faibles changements d'intensité de l'image) et atténue les hautes fréquences (variations rapides).

b) Filtre passe-haut

Il est utilisé pour amplifier les détails de hautes fréquences. Il peut permettre par exemple de restaurer des images qui ont été défocalisées et d'accentuer les contours en faisant ressortir les pixels compris entre des zones homogènes.

Il met en évidence les changements rapides de l'intensité de l'image (les hautes fréquences) et laisse les zones uniformes inchangées (basses fréquences).

$$H=\begin{bmatrix} -1 & -1 & -1 \\ -1 & 9 & -1 \\ -1 & -1 & -1 \end{bmatrix}$$

Figure 2.2: *masque de filtre*

1.3.2. Techniques du filtrage

a) Filtre linéaire

Le filtre local est dit linéaire si la valeur du nouveau pixel est une combinaison linéaire des valeurs des pixels du voisinage y compris le pixel en question [3].

b) Filtre non linéaire

Ils sont conçus pour régler les problèmes des filtres linéaires, surtout pour ce qui concerne la mauvaise conservation des contours. Leur principe est le même que celui des filtres linéaires, il s'agit toujours de remplacer la valeur de chaque pixel par la valeur d'une fonction calculée dans son voisinage. La différence majeure, est que cette fonction n'est plus linéaire mais une fonction quelconque (elle peut inclure des opérateurs de comparaisons ou de classification). Les filtres non linéaires les plus connu sont [1] :

3.2.2.1. Le filtre médian: Déplacer une fenêtre de taille impaire sur le support image ,Remplacer le pixel central (sur lequel est positionnée la fenêtre) par la valeur médiane des pixels inclus dans la fenêtre [1].

1.4. La segmentation

La segmentation est un traitement de bas-niveau qui consiste à effectuer une partition de l'image en régions homogènes par rapport à un ou plusieurs critères. Les régions obtenues se distinguent les unes des autres par des différences significatives selon ces mêmes critères. Après ces étapes, nous pouvons introduire un traitement sectoriel de différentes manières.

La segmentation consiste à extraire des points, des lignes ou des régions. Le choix d'une technique de la segmentation est liée à plusieurs facteurs comme : la nature de l'image, les conditions d'acquisition (bruit), les primitives à extraire (contour, texture,…).

La segmentation fait référence aux notions de similarité comme les perçoit, le système visuel humain et ceci donne naissance à deux approches couramment qualifiées d'approche « région » et d'approche « frontière » [4].

1.4.1. Approche frontière (contour)

Un contour est défini comme étant la frontière entre deux régions, la détection de contour est équivalente à la détection de la discontinuité entre ces deux région.

Les approches contour opérant a trouver les zones de variations significatives d'intensité lumineuse (niveau de gris) ou de couleur dans l'image .nous pouvons citer les approches se basant sur les différences finies comme l'operateur de gradient, l'operateur de laplacien et les

différents filtres à savoir: le filtre de Sobel , Prwitt et Robert ou bien des approches analytique comme le filtre de Canny. Mais ce genre de techniques est peu exploitable car elles donnent souvent des contours non fermés, bruités ou non détectés: une utilisation des propriétés des régions comprise entre ces contours pourrait nettement améliorer la détection de ce derniers [5].

1.4.2. Approche région

Il est très difficile de mettre au point un algorithme de segmentation qui fonctionne correctement dans toutes les situations comme le fait aussi bien le système visuel humain. Nous allons dans ce qui suit exposer quelques techniques de détection des zones homogènes de l'image [4].

a) Segmentation par seuillage

La segmentation par seuillage utilise l'histogramme pour extraire les différentes régions de l'image.

Le seuillage est une technique simple, non contextuelle, globale, qui repose sur une mesure quantitative d'une grandeur. Il permet de classer les pixels en deux catégories, ceux dont la mesure est inférieure au seuil (S) et ceux dont la mesure excède ou égale le seuil.

$$g(x,y) = \begin{cases} 0 & \text{si } f(x,y) < S \\ 1 & \text{si } f(x,y) \geq S \end{cases} \qquad (1.1)$$

La transformation produite une image binaire. Les techniques de seuillage présentent de nombreuses variantes [4].

b) Segmentation par croissance de région (région growing)

- La méthode de croissance de régions est une méthode de fusion.
- L'image est décomposée en primitives 'régions' (une région ≡ un seul pixel).
- Celles-ci sont ensuite regroupées de manière itérative selon un ou plusieurs critères de similarité, jusqu'à ce qu'il n'y ait plus de fusion possible [4].

c) Segmentation par division/rassemblement (split and merge)

Le processus est décomposé en deux étapes :

- **Division :** analyse de chaque région R_i . Si celle-ci ne vérifie pas le critère

d'homogénéité, alors on divise cette région en blocs (le plus généralement en 4 quadrants) et l'on réitère le processus sur chaque sous-région prise individuellement [4].

- **Rassemblement :** Si l'union de deux régions voisines (R_i ,R_j) vérifie le critère

d'inhomogénéité: fusion des régions.

- Soit R_i (i = 1, ...n) un ensemble de régions effectuant la partition d'une image, et soit P un prédicat mesurant l'homogénéité de ces régions.
- Si P (R_i) = faux : R_ non homogène (subdivision des régions ou split).
- Si P (R_i U R_j) = vrai (i ≠ j et R_i U R_j = ensemble connexe): R_i et R_j sont homogènes et doivent être fusionnées (fusion des régions ou merge) [4].

d) Segmentation par classification

Les méthodes de classification permettent de regrouper des objets en groupes ou classes d'objets plus homogènes. Les objets regroupés ont des caractéristiques communes, ils sont similaires mais se distinguent clairement des objets des autres classes [4].

1.4.3. La morphologie mathématique

La morphologie mathématique constitue une technique d'analyse d'images à part entière et peut être utilisée pour résoudre un grand nombre de problèmes de traitement d'images

Le principe de base de la morphologie mathématique est de comparer l'image à analyser par rapport à un ensemble de géométries connues appelées éléments structurants que l'on déplace de façon à ce que leurs origines passent par toutes les positions de l'image, pour mettre en évidence certaines caractéristiques de l'image.

a) L'érosion

Dans l'opération de construction on pose le centre G de l'élément structurant sur les pixels (appartenant à la région à segmenter) et on vérifie si tout l'élément structurant est inclus totalement dans l'objet à segmenter. L'opération d'érosion consiste à prendre que les pixels pour lesquels l'élément structurant est inclus dans l'objet.

Dans l'opération d'érosion les points blancs isolés disparaissent. Soit E^N un espace euclidien à N dimension et soit I et S deux sous ensembles de E^N , l'érosion de I par S est définie à partir de la soustraction de Minkowski par [3]:

$$I\Theta S = \{c\epsilon\, E^N / i = c + s\,, i\epsilon I, \forall\, s\epsilon S\} \qquad (1.2)$$

b) La dilatation

On fait passer le centre de l'élément structurant sur tout les pixels de l'image et seuls les pixels pour les quels l'intersection entre l'élément structurant et la zone à dilater seront garder pour construire la zone dilatée.

Elle consiste à éliminer les points noirs isolés au milieu des parties blanches.

Soit E^N un espace euclidien à N dimension et soit I et S deux sous ensembles de EN , la dilatation de I par S est définie à partir de l'addition de Minkowski par [3]:

$$I \oplus S = \{c \in E^N / c = i + s, i \in I, s \epsilon S\} \qquad (1.3)$$

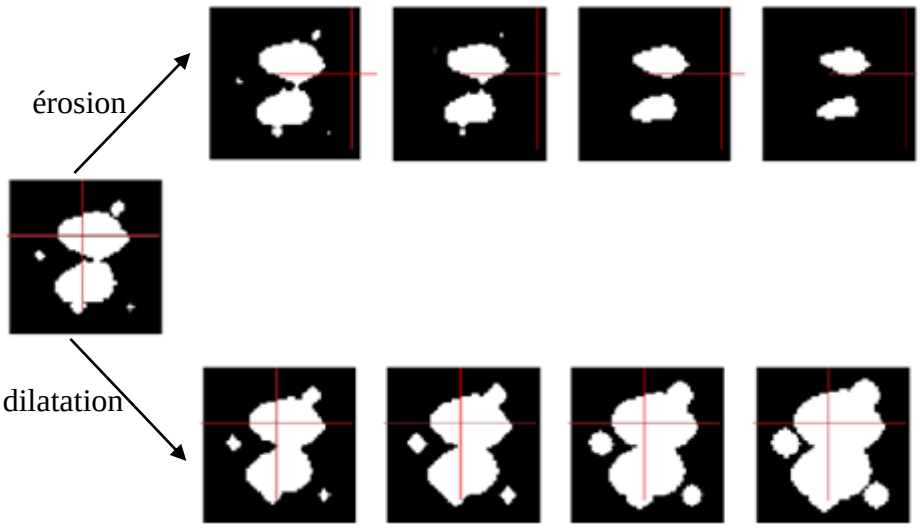


Figure2.3: *Exemples d'érosions et de dilatation*

c) L'ouverture

L'ouverture de I par S notée IoS est le résultat d'une érosion de I suivie d'une dilatation de l'ensemble érodé par le même élément structurant.

$$IoS = (I \Theta S) \oplus S \qquad (1.4)$$

L'ouverture adoucit les contours et élimine les pics aigus [3].

d) La fermeture

C'est l'opération duale de l'ouverture, notée I•S, c'est le résultat d'une dilatation suivie d'une érosion en utilisant le même élément structurant.

$$I \bullet S = (I \oplus S) \Theta S \qquad (1.5)$$

La fermeture fusionne les coupures étroites, élimine les petits trous, et comble les vides sur les contours [3].

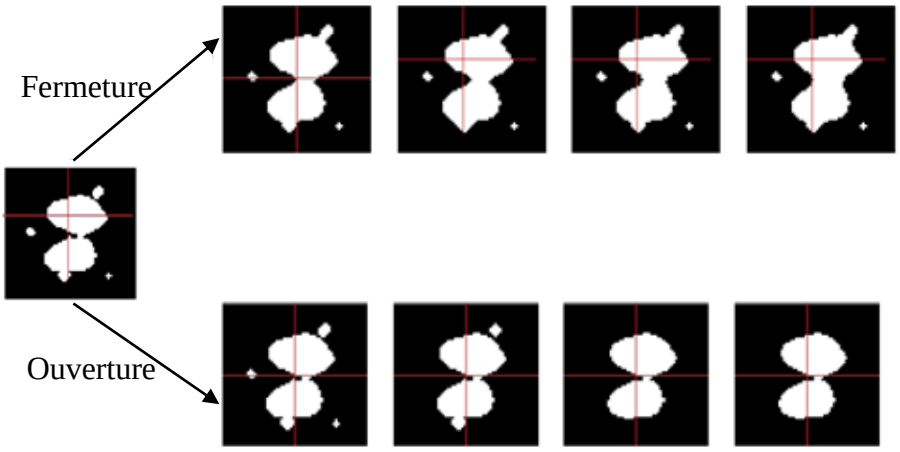


Figure2.4: *Exemples ouverture et de fermeture*

1.5. Indexation

L'indexation consiste à extraire et à modéliser les caractéristiques de l'image qui sont principalement la forme, la couleur et la texture. Chacune de ces caractéristiques pouvant être considérée pour une image entière ou pour une région de l'image [6].

1.5.1. Couleur

C'est le premier descripteur qui est employé pour la recherche d'images. Une technique très utilisée pour la couleur est l'intersection d'histogrammes. Les histogrammes sont faciles à calculer, et robustes à la rotation et à la translation.

Cependant l'utilisation d'histogrammes pour l'indexation et la recherche d'images pose quatre problèmes. Premièrement, ils sont de grandes tailles, donc par conséquent il est difficile de créer une indexation rapide et efficace en les utilisant tels quels. Deuxièmement, ils ne possèdent pas d'informations spatiales sur les positions des couleurs. Troisièmement, ils sont sensibles à de petits changements de luminosité, ce qui est problématique pour comparer des images similaires, mais acquises dans des conditions différentes. Et quatrièmement, ils sont inutilisables pour la comparaison partielle des images, puisque calculés globalement sur toute l'image. Pour remédier à ces problèmes, deux approches sont possibles.

La première approche ajoute des informations spatiales aux histogrammes. Tels que les moments d'inertie et la cohérence spatiale. Un pixel est cohérent s'il appartient à une région validée par la segmentation et incohérent autrement, la comparaison entre deux histogrammes devient la comparaison entre les valeurs d'histogrammes dans les classes correspondantes. La deuxième approche recherche d'autres espaces de couleurs qui se basent sur la perception de couleur de l'humain. L'espace RVB est un espace de couleur utilisé **[7].**

1.5.2. Forme

la forme est importante Caractéristique d'identifier et de distinguer les objets.

Les descripteurs de formes sont classés en base à la limite (ou à base de contour) et à base de région Méthodes. Cette classification prend en compte si les caractéristiques de la forme sont extraites Du contour uniquement ou de toute la région de la forme. Ces deux classes, à leur tour, peuvent Être divisé en des descripteurs structurels (locaux) et globaux. Cette subdivision repose sur la question de savoir si La forme est représentée dans son ensemble ou représentée par des segments / sections. Un autre possible La classification classe les méthodes de description de la forme dans les techniques de domaine spatial et de transformation, selon que des mesures directes de la forme sont utilisées ou une transformation est appliqué [8].

- **Moment Invariants:** Pour Moment Invariants, chaque objet est représenté par un 14-

Vecteur de caractéristiques dimensionnelles, y compris deux ensembles d'invariants de masse normalisés , un du contour de l'objet et un autre de sa silhouette d'objet solide. Encore une fois, la distance euclidienne est habituellement utilisée pour mesurer la similitude entre les différentes formes telles qu'elles sont représentées par leurs Moment Invariants.

- **Curvature Scale Space (CSS)** Le descripteur CSS représente une organisation à plusieurs niveaux des points de passage à zéro de la courbure d'une courbe planaire. Dans ce sens, la dimension de ses vecteurs caractéristiques varie selon les contours différents, donc un algorithme de correspondance spécial est nécessaire pour comparer deux descripteurs CSS.
- **Statistiques d'angle de faisceau (BAS):** Le descripteur BAS est basé sur les faisceaux issus d'un pixel de contour. Un faisceau est défini comme l'ensemble des lignes reliant un pixel de contour au reste des pixels le long du contour. A chaque pixel de contour, l'angle entre une paire de lignes est calculé et le descripteur de forme est défini en utilisant les statistiques de troisième ordre de tous les angles de faisceau dans un ensemble de quartiers. La similitude entre deux fonctions de moment BAS est mesurée par un algorithme optimal de sous-séquence correspondante (OCS), tel qu'indiqué dans.
- **Tensor Scale Descriptor (TSD):** TSD est un descripteur de forme basé sur le tenseur Concept d'échelle -un paramètre morphométrique donnant une représentation unifiée de l'épaisseur, de l'orientation et de l'anisotropie de la structure locale. C'est-à-dire, à n'importe quel point d'image, son échelle de tenseur est représentée par la plus grande ellipse (2D) centrée à ce point et dans la même région homogène. Le TSD est obtenu en extrayant les paramètres de l'échelle de tenseur pour l'image d'origine, puis en calculant

l'histogramme d'orientation de l'ellipse. Les TSD sont comparés en utilisant une fonction de distance basée sur la corrélation.

- **Contour Saliences (CS):** Le calcul CS utilise la transformation de la forêt d'image pour calculer les valeurs de salience des pixels de contour et pour localiser les points de survie le long du contour en exploitant la relation entre un contour et ses squelettes internes et externes. Le descripteur salience contour se compose des valeurs de salience des pixels saillants et de leur localisation le long du contour, et sur une fonction de distance heuristique correspondant.
- **Salences du segment (SS):** Le descripteur de salivité du segment est une variation de la Descripteur de niveau de contour qui intègre deux améliorations: les valeurs de salience des segments de contour, au lieu des valeurs de points isolés et un autre algorithme de correspondance qui remplace la concordance heuristique par une approche optimale. Les valeurs de salience le long du contour sont calculées et le contour est divisé en un nombre prédéfini s de segments de même taille. Les zones d'influence interne et externe de chaque segment sont calculées en résumant les zones d'influence de leurs pixels correspondants. SS montre qu'il présente une meilleure efficacité que plusieurs autres descripteurs de forme.

1.5.3. Texteure

Il n'y a pas de définition largement acceptée de la texture. cependant, Cette propriété d'image peut être caractérisée par l'existence de primitives de base, dont l'espace La distribution crée certains modèles visuels définis en termes de granularité, de directionnalité et Répétitivité. Il existe différentes approches pour extraire et représenter des textures. Ils peuvent Être classé dans des modèles basés sur l'espace, basés sur la fréquence et des signatures de texture. Prochain, Certaines de ces approches sont décrites.

- **Matrice de co-occurrence:** Est l'une des techniques les plus traditionnelles pour l'encodage de la texture information. Il décrit les relations spatiales entre les niveaux de gris dans une image. Une cellule Défini par la position (i, j) dans cette matrice enregistre la probabilité à laquelle deux pixels de gris Les niveaux i et j se produisent dans deux positions relatives. Un ensemble de probabilités de co-occurrence (par exemple, Énergie, entropie, contraste) a été proposé pour caractériser les régions texturées.

Les descripteurs de texture basés sur la fréquence comprennent, par exemple, les coefficients d'ondelettes Garbor . Ce descripteur vise à caractériser l'information sur la texture en termes de contraste, La grossièreté et la directionnalité. L'initiative a proposé trois descripteurs de

texture: Descripteur de navigation de texture, descripteur de texture homogène et descripteur d'histogramme de bord local [8].

1.6. Conclusion

Les techniques de traitement d’image sont des techniques très diverses et le choix de l’une parmi elle, est un choix qui dépend essentiellement de la nature de l’application et des résultats qui peuvent être obtenus par l’application de l’une ou de l’autre. Cependant, chaque ensemble de ces techniques est destiné à une application spécifique. Certains ensembles ont des applications communes, mais les résultats obtenus seront différents du point de vue avantages et inconvénients.

Notons que la majorité de ces techniques sont supposées comme des opérations de bas niveau, c'est-à-dire qu’elles ne cherchent pas à comprendre la sémantique de l’image, et se limite à la transformation et à la manipulation brute des pixels sans prendre en compte leur signification.

Chapitre 02
La reconnaissance des manuscrit arabe

2.1. Introduction

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 [9].

2.2. Définitions

2.2.1. L'écriture

L'écriture est l'un des plus anciens modes de communication dans notre civilisation, ce mode s'est beaucoup développé et a évolué significativement à travers les siècles. Un individu apprend à écrire en copiant des formes à partir d'un cahier d'écriture standard qui, lui aussi, diffère selon la localisation géographique ainsi que les circonstances temporelles, sociales et culturelles.

2.2.2. Un texte manuscrit

Un texte manuscrit peut présenter des intérêts variés. Les manuscrits anciens, par exemple, pourraient servir à étudier l'évolution de la forme et du style d'écriture d'une société au fil du temps, ce qui, à son tour reflète les changements historiques et culturels de cette société. Des connaissances sur les différentes lettres, les ligatures, les signes de ponctuation, les abréviations et la façon avec laquelle ils ont évolué, permetauxpaléographes2 et historiens d'identifier les périodes au cours desquelles un manuscrit a été écrit. La quantité de manuscrits anciens stockés dans des archives, des bibliothèques et des collections privées est considérable et il serait très utile de développer des systèmes informatiques qui pourraient aider les paléographes à dater, classer et authentifier ces manuscrits **[10].**

Figure2.1: *Un manuscrit arabe ancien extrait de la bibliothèque de Tombouctou au Mal****.[B]***

2.3. L'écriture arabe

L'écriture arabe a vu le jour aux alentours du VI ème siècle avant l'apparition de l'écriture cursive nabatéenne, et s'est progressivement répandue avec l'existence de l'Islam et la révélation coranique. Les principales caractéristiques de la langue arabe sont:

• L'alphabet arabe comprend vingt-huit lettres fondamentales . Chacune a entre deux et quatre formes selon sa position dans le mot. La figure donne toutes les formes possibles pour chaque lettre de l'alphabet arabe.

Début	Milieu	Fin et lié	Fin	Début	Milieu	Fin et lié	Fin
أ		أ	أ	ضـ	ـضـ	ـض	ض
بـ	ـبـ	ـب	ب	طـ	ـطـ	ـط	ط
تـ	ـتـ	ـت	ت	ظـ	ـظـ	ـظ	ظ
ثـ	ـثـ	ـث	ث	عـ	ـعـ	ـع	ع
جـ	ـجـ	ـج	ج	غـ	ـغـ	ـغ	غ
حـ	ـحـ	ـح	ح	فـ	ـفـ	ـف	ف
خـ	ـخـ	ـخ	خ	قـ	ـقـ	ـق	ق
د		ـد	د	كـ	ـكـ	ـك	ك

ذ		ـذ	ذ	ـل	لـ	لـ	ل
ر		ـر	ر	مـ	مـ	مـ	م
ز		ـز	ز	نـ	نـ	نـ	ن
س	سـ	ـس	س	هـ	هـ	هـ	ه
ش	شـ	ـش	ش			ـو	و
ص	صـ	ـص	ص	يـ	يـ	ـي	ي

Table1.1: *Les différentes formes possibles d'apparence des caractères de l'alphabet arabe.*

- Quelques caractères arabes incluent dans leur forme un, deux ou trois points diacritiques. Ces points peuvent se situer au-dessus ou au-dessous du caractère mais jamais en haut et en bas simultanément.
- L'existence du "hamza" (le zigzag), qui se comporte, soit comme une lettre à part entière, soit comme un diacritique.
- Certaines formes de lettres ne peuvent dans aucun cas être rattachées à la lettre suivante, ce qui fait qu'un mot unique peut être entrecoupé d'un ou plusieurs espaces, lesquels sont aussi utilisés pour séparer les mots.
- Les voyelles "a", "i" et "ou" ne sont pas utilisées systématiquement dans l'écriture arabe ; des signes qui correspondent à des voyelles sont employés pour éviter des erreurs de prononciation.
- On trouve également des chevauchements (**Figure**2.2).et des ligatures **(Figure**2.2) .dans l'écriture manuscrite ce qui complique la tâche de reconnaissance

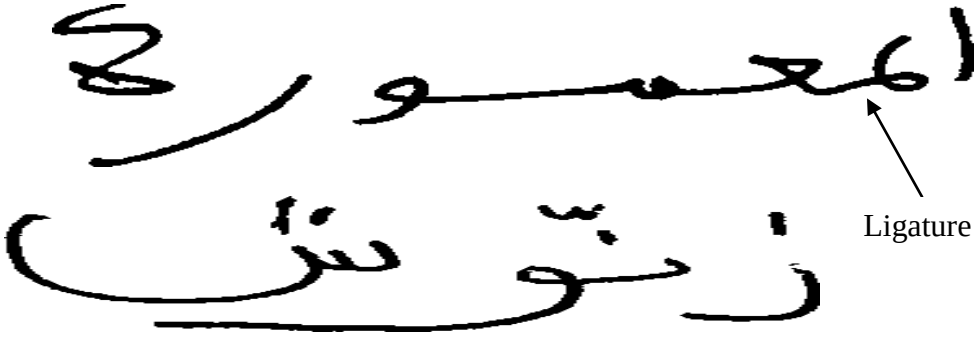


Figure 2.2: *Les ligatures et les chevauchements dans un mot arabe.*

2.4. Caractéristique morphologique de l'écriture arabe

- ❖L'écriture arabe est semi-cursive aussi bien dans sa forme imprimée que manuscrite. Les caractères d'une même chaîne (ou pseudo-mot) sont ligaturés horizontalement et parfois verticalement (dans certaines fontes deux, trois et même quatre caractères peuvent être ligaturés verticalement), occultant ainsi toute tentative de segmentation en caractères.
- ❖ De plus, la forme d'un caractère diffère selon sa position dans le pseudo-mot et même dans certains cas, selon le contexte phonétique. En outre, plus de la moitié des caractères arabes incluent dans leur forme des points diacritiques (1,2 ou 3). Ces points peuvent se situer au-dessus ou au-dessous du caractère, mais jamais en haut et en bas simultanément. Plusieurs caractères peuvent avoir le même corps mais un nombre et /ou une position de points diacritiques différents.
- ❖ D'autre part, le caractère arabe présente une forme cursive voyellée nécessitant, pour la majorité des lettres, des matrices de dimensions importantes. Ceci laisse jusqu'à présent les formes informatisées des caractères arabes non encore normalisées.
- ❖ Le mot arabe n'a pas de longueur fixe, il peut comprendre un ou plusieurs pseudo-mots incluant chacun un nombre souvent différent de caractères. L'étude de la morphologie des pseudo-mots montre que l'écriture arabe présente des variations dans des bandes horizontales plus ou moins complexes en fonction de la calligraphie des caractères contenus dans le pseudo-mot. La bande centrale est généralement la plus chargée au point de vue densité d'informations en pixels. Elle correspond aux lieux des ligatures horizontales, aux caractères centrés (sans extensions), aux boucles [12].

2.5. Architecture générale d'un système de reconnaissance

C’est la partie la plus étudiée des systèmes de reconnaissance d’écrits car la plus ancienne et la plus simple quand il s’agit de caractères numériques ou latins (vocabulaire limité). Le dictionnaire est constitué d’images de caractères étiquetées. Une image de caractère inconnu pourra être étiquetée comme l’élément du dictionnaire dont elle est la plus similaire. La notion de similarité implique une description des caractères dans un espace de représentation muni d’une métrique. Un système de reconnaissance se présente sous forme d'un ensemble de modules en partant de l'acquisition des caractères à reconnaître pour arriver à leur reconnaissance effective (figure). Dans ce qui suit, nous nous intéressons aux systèmes de reconnaissance hors-ligne [13].

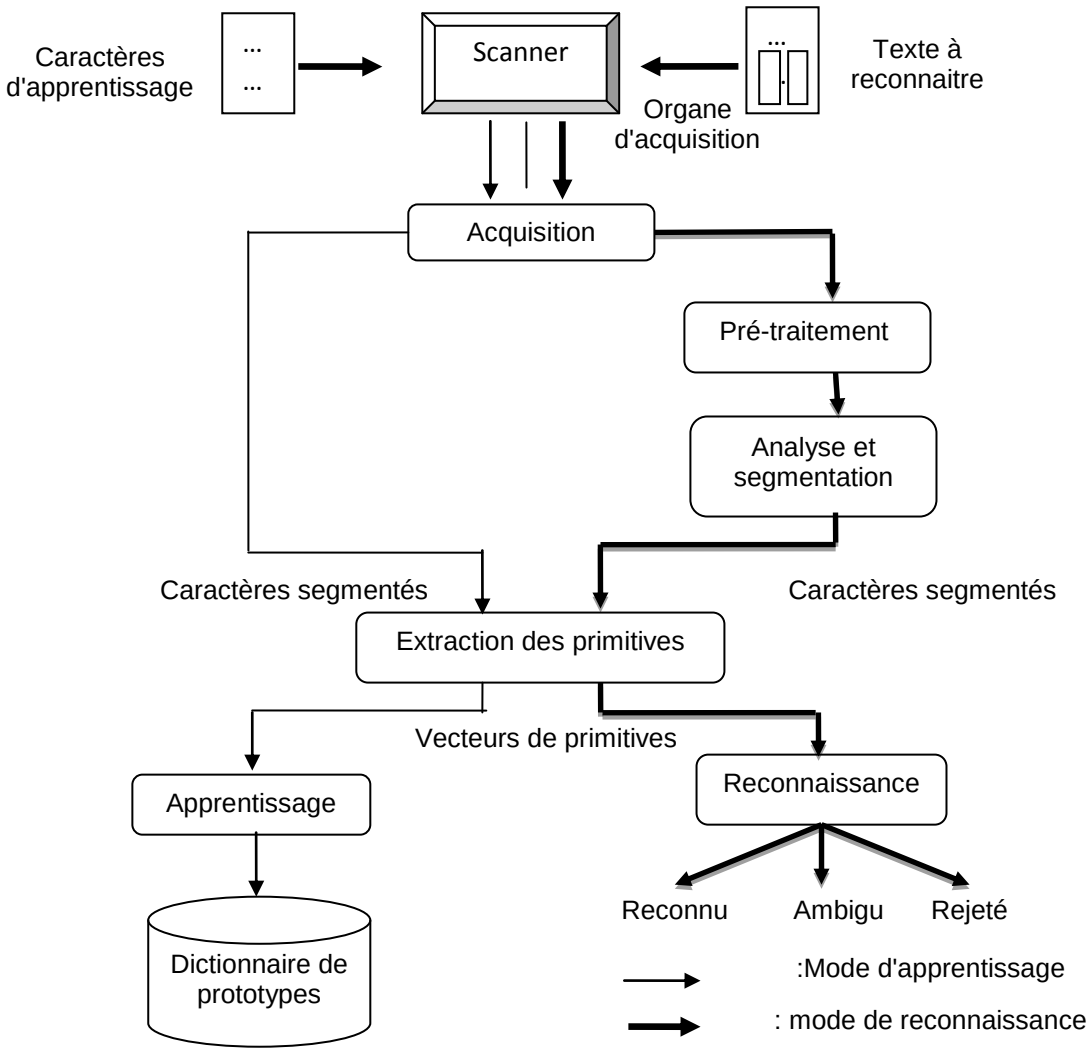


figure2.3: *système de reconnaissance*

2.5.1. Acquisition

La phase d'acquisition consiste à capter l'image d'un texte au moyen des capteurs physiques (scanner, caméra,…) et de la convertir en grandeurs numériques adaptés au système de traitement, avec un minimum de dégradation possible [14].

2.5.2. Prétraitement

A. Binarisation

La première étape à laquelle est soumise une image de texte manuscrit est la binarisation de cette image. Les images considérées n'étant pas trop bruitées [15].

On distingue en général deux approches: les approches à seuillage global, et les approches `a seuillage adaptatif.

1) Seuillage globale

Le seuillage global consisté à prendre un seuil ajustable, mais identique pour toute
l'image. Chaque pixel de l'image est comparé à ce seuil et prend la valeur blanc ou noir selon qu'il est supérieur ou inférieur. Cette classification ne dépend alors que du niveau de gris du pixel considéré **[**16**].**

2) Seuillage adaptatif

Dans les documents pour lesquels l'intensité du fond et l'intensité de la forme peuvent varier au sein du document, un seuillage global est inadapté. Il devient nécessaire de choisir le seuil de binarisation de manière locale. On calcule un seuil de binarisation pour chaque pixel de l'image, en fonction de son voisinage.

B. Squelette

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

Il n'existe pas de définition unique du squelette. Le squelette doit seulement remplir trois conditions :

- ✓ Il doit être aussi fin que possible (typiquement, 1 pixel d'épaisseur);
- ✓ Il doit respecter la connexité;
- ✓ Il doit être centré dans la forme qu'il représente;

Il existe de nombreuses méthodes de squelettisation. L'une des manières d'évaluer le squelette est de calculer l'axe médian de la forme, comme l'ensemble des centres des boules maximales de cette forme **[**16**]**

2.5.3. Segmentation de caractères

La segmentation de manière générale est la séparation des zones connexes appartenant au même environnement spatial. La segmentation de caractères est le moyen d'isoler des fragments dans un mot manuscrit pour être des unités d'information de base pour la reconnaissance.

❖ **Difficultés dans la segmentation de caractères**

La segmentation de caractères manuscrits est une tâche difficile due à la dégradation des images acquises, l'écriture cursive, la grande variété des styles d'écriture, le chevauchement ou la fragmentation de caractères, ou encore l'appartenance des caractères à un fond texturé [20]. L'un des principaux problèmes qui rendent cette tâche pénible est sa forte dépendance avec la reconnaissance due à la similarité visuelle de caractères.

Il est souvent extrêmement difficile de prendre la bonne décision en choisissant un caractère candidat tout en rejetant le reste d'un ensemble d'hypothèses pour un caractère donné.

L'existence de beaucoup de caractères visuellement semblables (` ح ', 'ج' et ` خ ', `د 'et ` ذ ') aussi bien que l'existence des caractères qui replient exactement les parties d'autres caractères (`ص ' et ` ط ', ` ض ' et ` ظ ') augmente ce problème [19]. Si, par exemple, la segmentation adoptée accepte plusieurs coupures, certaines lettres peuvent être divisées en sous-lettres. La lettre " ـشـ" par exemple, pourrait être interprétée comme une paire de lettres "ـنـ ". Ces deux interprétations ne peuvent pas être considérées en même temps, Ceci a comme conséquence un certain nombre de contraintes sur les interprétations possibles des sous lettres.

a. Les méthodes de la segmentation de caractères

Les anciennes méthodes de la segmentation de caractères sont principalement basées sur l'analyse du profil de projection et l'estimation de pitch (nombre de caractères par unité de distance horizontale), ces méthodes donnent de bons résultats sur les caractères imprimés ayant des espacements de même taille. L'analyse de composantes connexes est ensuite employée pour traiter les ligatures et la police italique. Les évolutions les plus récentes combinent la segmentation avec la reconnaissance et emploient les résultats de la reconnaissance aussi bien que le contexte linguistique comme feedback pour améliorer les décisions de segmentation [19].

Les méthodes de la segmentation peuvent être distinguées en quatre grandes catégories qui sont :

la segmentation par dissection, qui est l'isolement de diverses unités d'écriture avant la reconnaissance, *la segmentation par reconnaissance*, qui est l'isolement de lettres,

particulièrement, dans les mots cursifs en se servant de la reconnaissance, *la segmentation hybride* qui combine la dissection avec des méthodes de recherche et *la segmentation holistique* qui reconnaît le mot tout entier sans faire la segmentation en caractères.

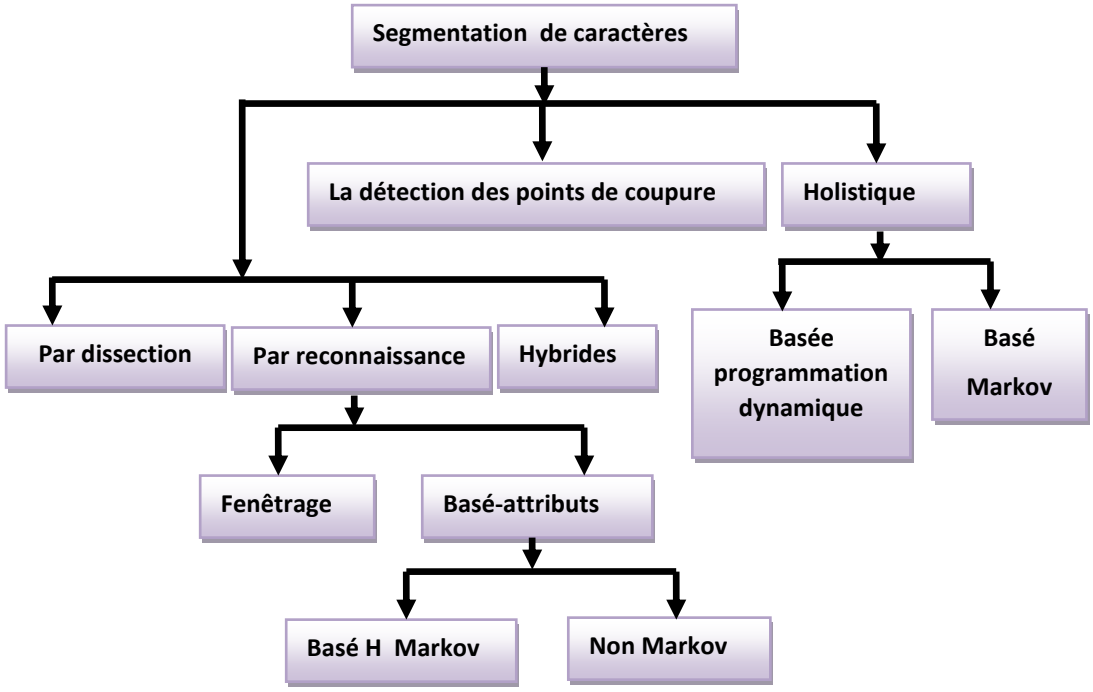


Figure 2.4: *Classification des méthodes de segmentation de caractères*

A. La segmentation par dissection

La segmentation par dissection est un processus indépendant de la reconnaissance qui s'effectue avant la reconnaissance, son but est de trouver l'endroit exact de séparation de caractères. Elle identifie les segments basée sur les propriétés de caractères. Ce processus de coupure de l'image en composants significatifs est appelé "dissection". La dissection est un processus qui analyse une image sans employer des informations structurelles.

Les techniques de la segmentation par dissection incluent :

- ✓ L'analyse de composants connexes
- ✓ L'espace blanc et l'estimation de pitch
- ✓ Landmarks (points de repère)
- ✓ Analyse de profils

Nous décrivons par la suite quelques méthodes de cette classe [18].

a. L'analyse de composants connexes

Une composante connexe est la région maximale de Pixels connexes (non séparés par un contour), la définition de l'ensemble de composantes connexes est la division de l'image ensegments.

Il existe de différents algorithmes basés sur différentes structures pour déterminer les composants connexes d'images binaires tels que :

- La diffusion récursive des composants connexes en balayant récursivement la matrice d'image.
- Les chaînes de codage en traçant le contour autour d'un composant dans la matrice d'image
- analyse de composants connexes basée RLC (Run-Length-Contour).

Le principe de l'algorithme le plus simple pour déterminer les composants connexes est comme suit :

- Trouver le prochain Pixel non étiqueté dans la matrice d'image
- Diffuser récursivement aux voisins
- Marquer les Pixels visités par l'étiquette courante
- quand le composant est épuisé, sélectionner la prochaine étiquette et répéter

b. L'espace blanc et l'estimation de pitch

La notion de l'espace blanc vertical entre les caractères successifs est un concept important pour séparer les caractères et principalement pour les applications de police imprimée où chaque caractère occupe un bloc de largeur fixe. Le pitch, ou le nombre de caractères par unité de distance horizontale, fournit une base pour estimer les points de segmentation. La combinaison de l'espace blanc et l'estimation de pitch peuvent produire des résultats de segmentation fiables sur des caractères bien espacés avec une largeur fixe. L'estimation de pitch a permis également la segmentation correcte en cas de caractères fragmentés ou chevauchés [18].

c. Détection des bornes limites (landmarks)

Les attributs caractéristiques de l'image tels que les ascendantes et les descendantes d'un caractère peuvent servir comme bornes limites pour aider la segmentation des caractères dans un mot manuscrit. La segmentation de caractères basée sur la détection des ascendantes et des descendantes a été appliquée au texte imprimé aussi bien qu'à l'écriture cursive [18].

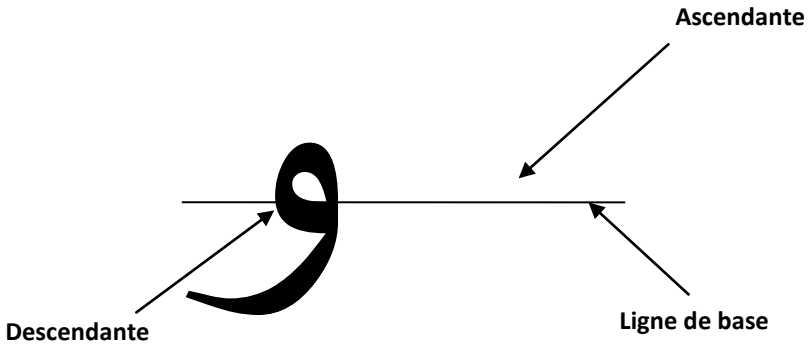


Figure 2.5: Ascendantes et descendantes d'un caractère

d. Les profils de projection

Le profil de projection horizontal est généralement utilisé pour segmenter les lignes d'un texte.

Les profils de projection sont caractérisés de basse performance quand ils sont appliqués à l'écriture cursive, les caractères chevauchés ou italiques.

Soit S(N, M) une image binaire de N lignes et M colonnes [18]

- ✓ **Le Profil horizontal :** est la somme de Pixels noirs perpendiculaires à l'axe des x; ceci est représenté par le vecteur ph de taille M définit par :

$$P_h[j] = \sum_{i=1}^{N} S[i,j] \qquad (2.1)$$

Les espaces entre mots ou même les caractères aussi bien que les tracés verticales peuvent être détectées par le profil de projection horizontal.

- ✓ **Le Profil vertical :** est la somme de Pixels noirs perpendiculaires à l'axe des y; ceci est représenté par le vecteur P_v de taille N définit par :

$$P_v[j] = \sum_{j=1}^{M} S[i,j] \qquad (2.2)$$

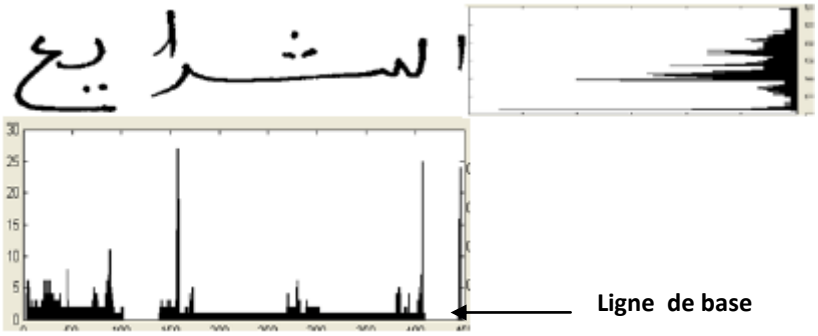


Figure2.6: *profils de projection horizontal et vertical*

Dans la figure, le profil de projection vertical montre deux crêtes à la ligne des x et à la ligne de base. Le profil de projection horizontal montre des zones sans hauteur qui correspondent aux blancs séparant les caractères d'un mot. L'analyse des profils de projection horizontaux et verticaux permet donc la séparation des mots et lignes du texte imprimé basée sur les crêtes et les zones sans hauteur.

L'analyse des profils de projection a été appliquée dans pour la segmentation des caractères manuscrits en se servant de l'analyse de composants connexes. Dans cette approche une ligne de texte est balayée horizontalement de gauche à droite. S'il existe moins de deux Pixels dans une colonne verticale, le balayage retourne "0", sinon il retournera le nombre de Pixels dans cette colonne. Un histogramme de balayage horizontal est alors construit. S'il existe dans l'histogramme une tranche de K "0" consécutifs (la valeur de K est déterminée expérimentalement) alors le point médian de cette tranche est considéré comme la frontière d'un caractère. L'approche considérée combine le profil de projection horizontal avec étiquetage de composants connexes pour séparer les caractères chevauchés. Si la distance entre deux frontières consécutives est grande (une fausse segmentation est détectée dans cette position) alors cet étiquetage de composants connexes est employé pour une segmentation ultérieure. Au cours de l'étiquetage le recouvrement vertical des composants est vérifié. Si deux composants ou plus sont entièrement recouverts verticalement, les composants représentent donc différentes parties du même caractère. Par contre, si le recouvrement horizontal entre deux boîtes de deux composants consécutifs est moins de 35% les deux composants, font partie de deux caractères différents.

B. La segmentation par reconnaissance

La segmentation par reconnaissance est un processus basé reconnaissance, une lettre ne peut être segmentée sans être reconnue et ne peut être reconnue sans être segmentée,

c.-à-d. la segmentation et la reconnaissance sont faites en même temps. La reconnaissance interprétant l'exactitude syntaxique ou sémantique des hypothèses de la segmentation est un processus qui sert généralement pour valider ou rejeter une hypothèse .Les méthodes de segmentation par reconnaissance peuvent être, eux mêmes distinguées en deux classes [18].

a. Les approches de Fenêtrage

Dans cette catégorie, des hypothèses de segmentation se sont générées en balayant l'image. La détection des points de coupure est confirmée ou validée par reconnaissance. Le principe de base est d'employer une fenêtre glissante de largeur variable pour fournir des séquences de segmentation expérimentales qui doivent être confirmées par reconnaissance de caractères. Des approches séquentielles et parallèles ont été employées.

L'approche séquentielle identifie des mots itérativement en glissant la fenêtre de gauche à droite.

L'approche parallèle procède d'une manière plus globale. Elle produit une grille d'attributs aux combinaisons possibles de lettres. La décision finale est trouvée en choisissant le chemin optimal à travers le graphe des attributs possibles [18].

b. Les approches basées attributs caractéristiques

Dans ce type d'approche, les attributs caractéristiques de l'image sont extraits, les correspondances possibles entre les attributs et les lettres sont ensuite générées, et puis une recherche de la meilleure combinaison parmi toutes les correspondances est effectuée Ce type d'approche réalise une segmentation implicite par classification des sous ensembles d'attributs spatiaux de l'image. Cette méthode peut être divisée en deux catégories:

*Les approches basées modèle de Markov caché (*HMM) et les approches *non-Markov*. L'objectif des modèles de Markov cachés est de modéliser les variations de l'écriture cursive comme structure probabiliste qui n'est pas directement observable. L'idée fondamentale des modèles de Markov cachés est de modéliser le texte manuscrit comme un processus de Markov. Ce processus permet d'affecter des probabilités de transition aux diverses combinaisons de lettres.

Les modèles de Markov cachés peuvent être employés pour la segmentation des caractères parce que les contraintes contextuelles telles que les fréquences de mots et de lettres aussi bien que des règles syntaxiques, peuvent être formulées par des probabilités de transition entre les différents états d'un caractère.

Un exemple de la segmentation par reconnaissance basée sur le modèle de Markov caché est la méthode de la reconnaissance des chaînes numériques proposée dans [28].

Cette approche suit deux étapes : La première emploie une méthode de segmentation implicite et d'une information contextuelle de la chaîne numérique pour fournir les chemins de segmentation possibles. Les hypothèses produites par cette étape sont ensuite reclassifiées dans une étape de vérification, cette dernière est basée sur un classificateur de chiffres manuscrits.

La 1ère étape de cette approche trouve les "N" meilleurs chemins de segmentation pour une chaîne numérique en utilisant la programmation dynamique afin d'apparier les modèles de Markov cachés (HMMs) contre une chaîne numérique donnée.

Dans ce processus, la segmentation d'une chaîne numérique manuscrite est réalisée lorsque tous ses composants sont identifiés par reconnaissance en tant que chiffres isolés. L'information contextuelle utilisée pendant l'apprentissage du Modèle de Markov caché est la pente des caractères aussi bien que la variation de la taille intra chaîne. La taille intra chaîne est la taille de la fenêtre entourant une chaîne numérique. Cette taille est employée pour effectuer la segmentation implicite. Ainsi, l'extraction des attributs de chaque chaîne d'apprentissage est faite en tenant compte la variation de la taille intra chaîne et la pente estimée de la chaîne originale afin de corriger les chiffres isolés.

Les approches Non_Markov telles que la relaxation probabiliste et le test d'hypothèse ont été employés dans la segmentation par reconnaissance. Hayes a utilisé une méthode de relaxation probabiliste pour lire des mots manuscrits. Le modèle utilise une description hiérarchique des mots dérivés par relaxation d'une représentation squelettique. La relaxation a été exécutée sur le graphe des primitives (sous lettres) et des lettres où toutes les segmentations possibles sont représentées[18].

C. La Segmentation hybride

Les méthodes hybrides combinent des méthodes de dissection et les méthodes de recherche des points de coupure par reconnaissance de manière hybride. Un algorithme de dissection est appliqué à l'image, mais l'intention est de couper l'image suffisamment en plusieurs endroits de telle sorte que les frontières correctes de segmentation sont incluses dans les coupures réalisées.

Une fois que ceci est assuré, la segmentation optimale est cherchée par évaluation des sous ensembles des coupures faites. Chaque sous-ensemble implique une hypothèse de segmentation, et la classification est appliquée pour évaluer les différentes hypothèses et ainsi choisir la segmentation la plus prometteuse .

Un exemple de ces méthodes est l'algorithme de segmentation produit une séquence de chiffres et de symboles, et puis reçoit un feed-back avec le module de reconnaissance pour réajuster les partitions si nécessaire. La segmentation est correcte si tous les chiffres sont reconnus avec une confiance proportionnelle. Le système commence la segmentation par séparations évidentes de caractères, ainsi les chiffres qui ne sont pas correctement identifiés sont considérés comme étant composés de caractères connectés. Les blocs rejetés sont encore divisés et reconnus dans une boucle de feed-back jusqu'à ce que la solution finale soit trouvée avec tous les segments reconnus. Le système fusionne également les fragments des chiffres avec les chiffres voisins pour créer de nouveaux segments. Si ces segments ne sont pas encore identifiés, alors différentes stratégies de fusion seront examinées.

Dans la pratique, plusieurs chiffres touchent leurs voisins, et constituent donc un composant connexe. Dans ce cas, le système applique un algorithme de division de contour pour trouver les chemins possibles permettant la séparation des caractères recouvrés, tels que l'analyse de contour, l'algorithme hybride Drop Falling et l'algorithme étendu de Drop Falling.

Ce système évite les tentatives inutiles de reconnaissance autant que possibles, il emploie donc une approche heuristique pour choisir un chemin et puis sert de la meilleure information venant du module de reconnaissance dans une boucle de feedback. Des chemins sont choisis de manière heuristique en utilisant des attributs structurels. Plusieurs caractéristiques ont été utilisées pour choisir les meilleurs chemins de coupure telles que :

1. *Nombre de coupures faites pour diviser le segment*
2. *Longueur de la coupure*
3. *Bordure de la coupure*

La segmentation est correcte si les deux chiffres sont reconnus. Si aucun caractère n'a été correctement identifié (comme chiffre) avec un niveau de confiance élevé, alors des chemins alternatifs seront examinés. Si l’une des deux parties est identifiée, alors l'autre partie est composée de caractères connectés et le procédé de segmentation sera répété récursivement pour cette partie jusqu'à ce qu'une solution complète soit trouvée. Pendant la reconnaissance du montant chiffre, le système fait de multiples tentatives de fusion et séparation. Lorsque de petits morceaux de caractères sont identifiés en tant que "fragments de caractère", ils seront

fusionnés avec le segment voisin en tenant compte des critères tels que la proximité et le chevauchement de caractères [18].

D. La segmentation holistique

Les approches holistiques essayent d'identifier les mots tout entiers, évitant de ce fait la nécessité de segmenter un mot en caractères. La reconnaissance est basée sur la comparaison d'un ensemble d'attributs simples extraits à partir du mot entier contre un lexique des chaînes représentant la forme théorique des mots possibles. Des mots sont représentés par une liste d'attributs tels que les ascendantes, les descendantes, les boucles, ...etc. Ces techniques sont généralement fondées sur les chaînes de Markov cachées ou la programmation dynamique [18].

2.5.4. Extraction des caractéristiques

Cette étape de la reconnaissance consiste à extraire des caractéristiques permettant de décrire de façon non équivoque les formes appartenant à une même classe de caractères tout en les différenciant des autres classes.

Il existe différentes méthodes d'extraction de primitives citées dans la littérature.

a. Primitives locales

Ces primitives sont des objets de la forme telque les extrémités, les croisements de traits, les boucles et les courbes. L'inconvénient de ces primitives est que leur extraction nécessite une squelettisation préalable du caractère, puisque l'épaisseur du trait ne contient pas d'information. Néanmoins ce sont des primitives très robustes vis à vis de la rotation, translation, homothétie. Dans cette catégorie, il existe 4 familles de caractéristiques : intersections avec des droites, arcs concaves et occlusions, extremas et jonctions [13]

b. Primitives globales:

Ces primitives sont dérivées de la distribution des pixels. Heutte et al suggèrent 3 familles de caractéristiques telles que: les moments invariants, les projections, et les profils. Elles sont extraites en considérant la distribution des pixels noirs de l'objet (caractère, mot, chiffres). Le processus d'identification de la meilleure méthode d'extraction de caractéristiques n'est pas évident. Par exemple, Trier étal .rapportent que les moments de Zernike s'appliquent bien sur des images à niveaux de gris et que la projection s'applique

souvent sur des caractères segmentés pour résoudre leur problème de reconnaissance de l'écriture manuscrite [13].

2.5.5. Apprentissage

Dans le cas d'apprentissage il s'agit en fait de fournir au système un ensemble de formes qui sont déjà connues (on connaît la classe de chacune d'elles). C'est cet ensemble d'apprentissage qui va permettre de « régler » le système de reconnaissance de façon à ce qu'il soit capable de reconnaître ultérieurement des formes de classes inconnues.
Il existe deux types d'apprentissage, supervisé et non supervisé [13].

a. Apprentissage supervisé:

L'apprentissage est dit supervisé si les différentes familles des formes sont connues a priori et si la tâche d'apprentissage est guidée par un superviseur ou professeur, c'est-à-dire le concepteur, indique, pour chaque forme échantillon rentrée, le nom de la famille qui la contient. La tâche d'apprentissage tente de conserver ses liens de parenté en répartissant les familles dans des classes séparées entre elles [13].

b. Apprentissage non supervisé

On l'appelle aussi, suivant l'approche utilisée, classification automatique, inférence ou encore apprentissage sans professeur. Il s'agit, à partir d'échantillon de référence et de règles de regroupement ou de modélisation, de construire automatiquement les classes ou les modèles sans intervention de l'opérateur. Ce mode d'apprentissage nécessite un nombre élevé d'échantillons et des règles de construction précise et non contradictoires pour bien assurer la formation des classes. Il évite l'assistance d'un opérateur mais n'assure pas toujours une classification correspondant à la réalité (celle de l'utilisateur) [13]**.**

2.5.6. Reconnaissance

Le processus de reconnaissance peut toujours se résumer à une décision de classification. Pour cela, il faut choisir une représentation qui permettra une description de l'objet analysé, puis une règle de décision qui s'appuie sur cette description. On peut distinguer trois types de représentation [21].

2.6. Les systèmes de reconnaissance de l'écriture

2.6.1. Les systèmes de reconnaissance de l'écriture imprimée :

Dans le document présenté au système de reconnaissance de l'écriture imprimée, l'écriture est caractérisée par le nombre de fontes (mono, multi ou omni-fonte). Le système de reconnaissance dans ce cas est généralement appelé OCR (pour Optical Caracter Recognition).

- **Un système OCR est dit mono-fonte** s'il ne traite qu'une seule fonte à la fois, c'est-à-dire qu'il ne connaît que le graphisme d'une fonte unique. Ce cas est simple particulièrement au niveau de l'apprentissage car l'alphabet représenté est réduit.
- **Un système OCR est dit multi-fonte** s'il est capable de reconnaître divers types de fontes parmi un ensemble de fontes préalablement apprises. Dans ce cas, le traitement doit réduire les écarts entre même caractère (taille, épaisseur et inclinaison).
- **Un système OCR omni-fonte** est capable de reconnaître toute fonte, généralement sans apprentissage, en se basant sur les règles topologiques et morphologiques de l'écriture [9].

2.6.2. Les systèmes de reconnaissance de l'écriture: Les systèmes de reconnaissance de l'écriture peuvent être classifiés selon deux critères :

- **Systèmes de reconnaissance en ligne**: Dans ce type de systèmes, la reconnaissance est effectuée en temps réel, c'est-à-dire elle est effectuée au fur et à mesure que le caractère est tracé, ce qui permet d'obtenir un large marge de correction et modification selon la réponse donnée à la phase de reconnaissance chevauchée à la phase d'acquisition. Ce mode d'acquisition est réservé généralement à l'écriture manuscrite. C'est une approche « signal » où la reconnaissance est effectuée sur des données à une dimension. L'écriture est représentée comme un ensemble de points dont les cordonnées sont fonction du temps. Où on s'intéresse aux méthodes et techniques de traitement du message tel qu'il est écrit, en prenant en compte les informations relatives au mécanisme d'écriture telles la position des points, la vitesse et l'accélération qui sont des fonctions du temps . Les moyens de saisie en ligne sont nombreux où la tablette graphique avec un stylo électronique et l'écran tactile sont couramment utilisés. Parmi les plus récentes plate-formes disposant d'un système de reconnaissance de l'écriture, nous trouvons le Palm et l'agenda électronique. Ces deux appareils regroupent une tablette à numériser et un programme procédant à la reconnaissance de l'écriture. De ce fait, son utilisation est plus attrayante puisqu'elle épargne à l'utilisateur le besoin de « scanner » a priori son écriture. Ceci a remis la reconnaissance en ligne au centre d'intenses efforts de développement au sein de la communauté de l'écrit et du document [9]

- **Systèmes de reconnaissance hors ligne** L'écriture hors-ligne (ou en différé, ou encore statique) est obtenue par la saisie d'un texte déjà existant, obtenue par un scanner ou une caméra. Dans ce cas, on dispose d'une image binaire ou en niveaux de gris, ayant perdu toute information temporelle sur l'ordre des points. De plus, ce mode introduit une difficulté supplémentaire relative à la variabilité du racé en épaisseur et en connectivité, nécessitant l'application de techniques de prétraitement. Les domaines d'application les plus typiques sont principalement associés au traitement automatique des adresses postales, du montant des chèques, des formulaires, des feuilles de soins [9].

2.7. Les approches de REM

L'objectif de la reconnaissance de l'écriture manuscrite est de développer un système qui se rapproche le plus de l'être humain dans sa capacité de lire. Cependant, le problème de la reconnaissance de l'écriture en général et de l'écriture manuscrite en particulier est si vaste qu'il est nécessaire de l'aborder par des sous-problèmes. Cette reconnaissance regroupe deux thèmes : la reconnaissance de caractères manuscrits isolés (numériques ou alphanumériques) et la reconnaissance de mots.

2.7.1. Reconnaissance de caractères isolés

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 (vue comme une forme globale). Cas études portent sur la reconnaissance de caractères manuscrits provenant de la segmentation de chiffres comme par exemple le montant numérique d'un chèque ou le code postal d'une adresse ou de la segmentation en lettres minuscules et/ou majuscules dans le cas par exemple du montant littéral d'un chèque ou du nom de la ville figurant sur une adresse. Ces caractères isolés présentent de fortes variations principalement provoquées par la position de la lettre dans le mot. Deux types de reconnaissance de caractères numériques manuscrits :

- **La reconnaissance des chiffres isolés**: peut être considéré comme un problème de classification des images des caractères isolés dans un ensemble d'alphabet donné.
- **Reconnaissance des chaînes numériques manuscrites**: il est à noter que ce type de systèmes commence généralement par une phase de segmentation qui consiste à séparer la chaîne numérique à des entités (chiffres) isolées et reconnaître ces dernières par un module de reconnaissance des chiffres isolés [9].

2.7.2. Reconnaissance de mots

Deux approches s'opposent en reconnaissance des mots : globale et analytique

a. L'approche globale

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 approche présente l'avantage de garder le caractère dans son contexte avoisinant, ce qui permet une modélisation plus efficace des variations de l'écriture et des dégradations qu'elle peut subir. Cependant 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.

b. L'approche analytique

Contrairement à l'approche globale, l'approche analytique cherche à identifier les caractères ou sous-caractères (graphèmes) issus de la segmentation (séparation de mots, des caractères) pour reconstituer les mots. La difficulté d'une telle approche a été clairement évoquée par Sayre en 1973 et peut être résumée par le dilemme suivant : "pour reconnaître les lettres, il faut segmenter le tracé et pour segmenter le tracé, il faut reconnaître les lettres". Cette approche est la seule applicable dans le cas de grands vocabulaires. Elle peut s'adapter facilement à un changement de vocabulaire. Elle permet théoriquement 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 [9].

2.8. Particularisation des problèmes

Dans les deux cas, la reconnaissance de l'écriture manuscrite n'a pu progresser que grâce à une particularisation des problèmes à résoudre. Par cette particularisation, le but recherché est de diminuer l'influence de la variabilité sur la reconnaissance. Ainsi, les spécialistes du domaine ont été amenés à s'intéresser à des applications particulières. Pour un type d'application donnée, il est possible d'imposer, à l'écriture à reconnaître, un certain nombre de restrictions et de contraintes :

Le type d'écriture considéré, le style d'écriture, le nombre de scripteurs potentiels, la taille du vocabulaire utilisé [9].

2.8.1. Type d'écriture

Des contraintes plus ou moins fortes concernant le type d'écriture peuvent être imposées au scripteur. On distingue le précasé pour lequel le scripteur doit s'efforcer d'écrire dans des cases prédéfinies (ex. : sur des bordereaux et les formulaires), le zoné où le scripteur écrit dans des zones bien délimitées (**ex**: dans des zones grisées ou colorées), le guidé caractérisé par l'existence d'une ligne support (comme par exemple sur les chèques), et le cas général où l'emplacement de l'écriture est libre [16].

2.8.2. Style d'écriture

À ces dispositions s'ajoutent des contraintes concernant le style d'écriture. Tappert a établi la classification suivante par ordre de difficulté croissante de reconnaissance: caractères précasés de type script, caractères scripts détachés, écriture scripte pure (lettres séparées), écriture cursive (lettres entièrement liées à l'intérieur des mots), écriture mixte mélangeant le script et le cursif [16].

2.8.3. Nombre de scripteurs potentiels

La diminution du nombre de scripteurs permet évidemment une diminution de la variabilité inter-scripteur. On peut distinguer trois types de systèmes de reconnaissance d'écriture, classés par ordre de complexité de fonctionnement croissante;

- ✓ **Mono-scripteur** : un seul scripteur peut utiliser le système de reconnaissance après apprentissage de son écriture ;
- ✓ **Multi-scripteur** : le système peut reconnaître les écritures d'un groupe restreint de personnes, soit après adaptation à l'écriture de chacun, soit sans adaptation;
- ✓ **Omni-scripteur** : le système est censé reconnaître toutes les écritures. Dans ce cas, la variabilité intra-scripteur s'ajoute à la variabilité inter-scripteur.

2.8.4. Taille du vocabulaire à reconnaître

A priori, on peut distinguer trois grandes catégories d'applications:

- o **Applications à vocabulaire très limité:** où le nombre de mots (ou de symboles) à reconnaître constitue un lexique de taille réduite (inférieure à 100 mots), comme par exemple dans le cas de l'ensemble des mots utilisés pour écrire en toutes lettres les montants des chèques. Il est alors possible de confronter chaque mot à reconnaître à l'ensemble des mots du lexique [16].

- **Applications à vocabulaire étendu:** mais pouvant être réduit de façon dynamique, comme l'ensemble des noms de rues associés à un bureau de poste distributeur;
- **Applications à vocabulaire très étendu:** plusieurs milliers ou dizaines de milliers de mots formant un dictionnaire et pour lesquels la confrontation systématique n'est plus possible comme par exemple le dictionnaire des noms de commune [16].

2.9. Conclusion

La lecture automatique de l'écriture manuscrite présente un intérêt indéniable dans l'accomplissement des tâches fastidieuses comme celles que l'on rencontre dans certains domaines : le tri postal, la lecture de chèques bancaires, la lecture des bordereaux, des bons de commande, des feuilles de déclaration **[**16**].**

Chapitre03
Automate pondéré et la matrice PSSM

3.1. Introduction

Dans ce chapitre, On rappellera les concepts de base des automates pondérés. Puis on fait un aperçu sur les matrices qu'on utilisera intensément dans notre modélisation. Enfin on terminera par l'explication du passage entre les matrices et l'automate.

3.2. Automate pondéré

En Informatique théorique, et particulièrement en théorie des automates, un automate fini pondéré est une généralisation des automates finis. Dans un automate fini usuel, qu'il soit déterministe ou non déterministe, les transitions ou flèches portent des étiquettes qui sont des lettres de l'alphabet sous-jacent. Dans un automate pondéré, toute transition porte de plus un certain poids. Ce poids peut être interprété comme le coût pour passer d'un état à un autre lorsque la transition est effectuée [21].

3.2.1. Définition : Un automate pondéré est un quintuplet (Q,∑,E, I, T), similaire à un automate fini où E, l'ensemble des transitions, est un sous-ensemble de $Q \times \sum \times Q \times K$ tel que :

- (K, ⊕,⊗) est un semi-anneau;
- pour tout triplet $(q, \sigma, q') \in Q \times \sum \times Q$, il existe un et un seul élément $k \in K$ tel que $(q, \sigma, q', k) \in E$ [24].

- **Fonction de poids :** La fonction de poids est l'application **w** qui associe à toute paire $(q, \sigma) \in Q \times \sum$ un élément $k \in K$.
- **Poids d'un chemin :** Soit un chemin $\pi = e_1 e_2 \ldots e_n$, le poids du chemin $w(\pi)$ dans l'automate pondéré A = (Q, ∑, E, I, T) dans le semi-anneau (K, ⊕,⊗), et le ⊗-produit des poids de toutes les transitions. On note:

$w(\pi) = w(orig(e_1), symb(e_1)) \otimes w(orig(e_2), symb(e_2)) \otimes \ldots \otimes w(orig(e_n), symb(e_n))$.

- **Poids d'un mot** : Le poids d'un mot **m** pour un automate pondéré A dans le semi-anneau (K, ⊕,⊗) est la ⊕-somme de tous les poids des chemins correspondant à **m** dans l'automate. On note :

$$w(m) = \oplus_{\pi \in \prod m} w(\pi) \qquad (3,1)$$

Notons que si l'automate est déterministe, alors il n'existera qu'un seul chemin pour un mot donné, alors la première opération du semi-anneau ne sera jamais utilisée et le poids d'un mot m sera exactement le poids de son unique chemin.

L'acceptation d'un mot par un automate pondéré se fait par deux critères :

- Le mot est-il dans le langage défini par l'automate fini simple (i.e. un chemin de I vers T) ?
- Le poids du mot a-t-il une valeur donnée ?

Concernant la valeur admissible d'un mot, on définit souvent un sous-ensemble J ∈ K, valeurs du poids pour lesquelles un mot est admissible. Ainsi, un mot m est accepté dans un automate pondéré A dans le semi-anneau (K, ⊕,⊗) si et seulement si :

- m ∈ L(A) ;
- w(m) ∈ J. [24]

Exemple : Un automate pondéré comptant une séquence

Soit le semi-anneau (N, min, ⊕) et l'automate A = (Q, ∑, E, I, T) avec :

• Q = {q_0, q_1, q_2} ;

• ∑= {0, 1} ;

• E = {(q_0, 1, q_1, 0), (q_0, 0, q_2, 0), (q_1, 1, q_1, 0), (q_1, 0, q_2, 0), (q_2, 0, q_2, 0), (q_2, 1, q_0, 1)} ;

• I = {q_0} ;

• T = {q_0} ;

Cet automate est présenté figure 3.1. Tel quel, il accepte tous les mots de ∑* suffixés par 01 et compte le nombre de fois où le motif 01 apparaît dans un mot. Si on pose J = ({2, 3, 4}, min, +) ⊂ (N, min, +), alors l'automate pondéré figure 3.1 accepte tous les mots de ∑* suffixés par 01 et tel que ce motif apparaisse entre 2 et 4 fois inclus.

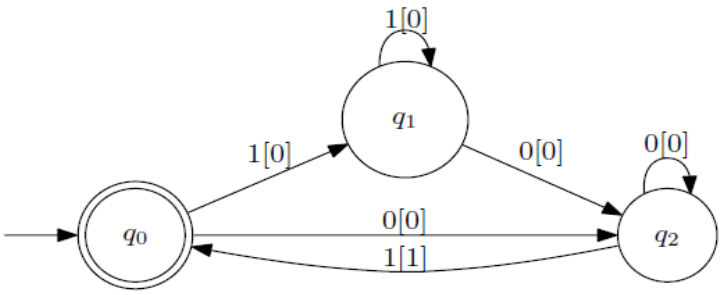


Figure 3.1: *Automate comptant les occurrences du motif 01*

3.2.2. Types des automates pondérés

Les automates pondérés contient plusieurs types parmi de cette méthodes il ya les automates probabilistes et les automates quantiques. Dans un automate probabiliste, les poids représentent des probabilités, et les matrices de la représentation doivent être stochastiques, dans une automate quantique, les matrices sont unitaires.

3.3. Les matrices poids-position, PSSM

Les matrices poids-position PSSM (Position Specific Scoring Matrices), introduit pour la première fois dans (Gribskov et al. , 1987) pour détecter des protéines éloignées, est généré à partir d'un groupe de séquences précédemment alignées par similitude structurelle ou séquentielle.

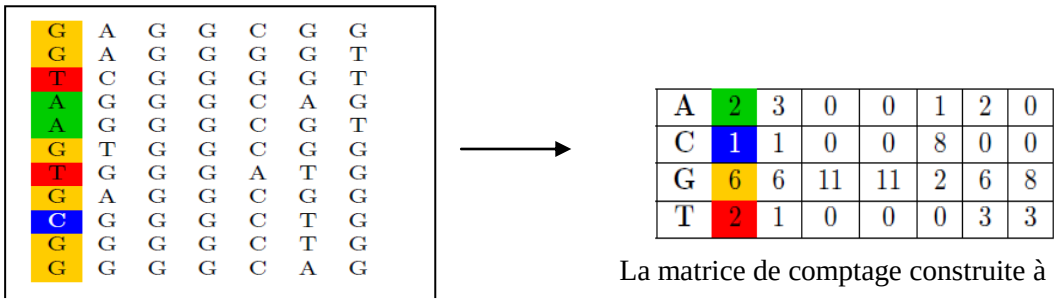
La modélisation de motifs approchés à l'aide de matrices est une caractérisation précise des alignements multiples. En effet, les matrices permettent de restituer l'ensemble des informations contenues dans un alignement si l'on considère les positions indépendantes les unes des autres. Cependant il est possible d'´etendre ce modèle aux dépendances entre les positions, par exemple aux dépendances d'une position [22]. Il y a six types matriciels.

3.3.1. Matrices de comptage

La méthode la plus triviale, pour retranscrire la composition en lettres des positions d'un alignement multiple, est le calcul d'une matrice de comptage. Ce motif associe à chaque lettre de l'alphabet un compte selon sa position dans le motif [23].

▪ **Définition** (La matrice de comptage) Soit A un alignement multiple de longueur L et ∑ un alphabet fini. Une matrice de comptage notée C représentative de l'alignement A est une matrice L × |∑| telle que pour toute lettre l ∈ ∑ et toute position p = 0 … L – 1 L'élément d'indices l et p de C noté C(p, l) est défini par le nombre d'occurrences de la lettre l à la position p de l'alignement multiple A [23].

▪ **Exemple:** la construction d'une matrice de comptage à partir de l'alignement multiple est illustrée par la figure suivant:



Alignement multiple de 11 séquences

A	2	3	0	0	1	2	0
C	1	1	0	0	8	0	0
G	6	6	11	11	2	6	8
T	2	1	0	0	0	3	3

La matrice de comptage construite à partir de l'alignement multiple

Figure 3.2 : *Calcul d'une matrice de comptage*

La matrice de comptage décrit quantitativement l'alignement multiple. Elle caractérise la composition en lettre des positions mais sans la ramener au total.

3.3.2. Matrices de fréquence

La matrice de fréquence normalise la composition en lettres des positions de la matrice de comptage. La matrice de fréquence associe à chaque lettre de l'alphabet une fréquence selon sa position dans le motif [23].

- **Définition:** Soit A un alignement multiple de longueur L et ∑ un alphabet fini. Une matrice de fréquence notée F représentative de l'alignement A est une matrice L × |∑| telle que pour toute lettre l ∈ ∑ et toute position p = 0 … L – 1 l'élément d'indices l et p de F noté F(p, l) est défini par la fréquence de la lettre l à la position p de l'alignement multiple A [23].

$$F(p,l) = \frac{C(p,l)}{\sum_{i\in\sum} C(p,i)} \qquad (3.2)$$

- **Exemple**: la construction d'une matrice de fréquence à partir de la matrice de comptage de la figure 3.2 est illustrée par la figure 3.3.

Ainsi la probabilité de génération d'un mot u = u_0 … u_{L-1}, selon la matrice de fréquence F, se calcule par la suite de multiplications suivantes.

$$P(u\backslash F) = \prod_{p=1}^{L} F(p,u_p) \qquad (3.3)$$

Un tel calcul fait par un ordinateur donnera un résultat très approché voir erroné. La précision des calculs par ordinateur ne permet pas une bonne approximation des suites de multiplications de nombres inférieurs à un.

Une autre lacune du modèle est sa sur-adaptation (overfitting en anglais), lorsqu'on apprend par cœur on ne se trompe jamais. Les fréquences de la matrice collent parfaitement à

A	2	3	0	0	1	2	0
C	1	1	0	0	8	0	0
G	6	6	11	11	2	6	8
T	2	1	0	0	0	3	3

La matrice de comptage de figure 3,2.

A	0,18	0,27	0	0	0,09	0,18	0
C	0,09	0,09	0	0	0,73	0	0
G	0,55	0,55	1	1	0,18	0,55	0,73
T	0,18	0,09	0	0	0	0,27	0,27

La matrice de fréquence construite à partir de la matrice de comptage

Figure 3.3 : *Calcul d'une matrice de fréquence l'ensemble*

L'ensemble de mots de départ alors qu'il ne s'agit que du sous-ensemble connus des données à représenter. Le modèle est donc trop proche des données de départ [23].

3.3.3. Matrices de fréquence corrigée

La correction du modèle des matrices de fréquence se fait par l'ajout d'un pseudocompte et présente un double intérêt. Classiquement, afin de palier aux problèmes d'approximation que posent les suites de multiplications, on convertit les modèles multiplicatifs en modèle additif en passant à l'échelle logarithmique. Cette conversion nécessite d'éliminer les éléments nuls pour le passage au logarithme. De plus un moyen de remédier à la sur adaptation d'un modèle est de lui apporter de la souplesse par l'ajout d'un pseudo-compte. Ce qui revient très précisément à ajouter aux données de départ des données représentatives du contexte. Dans notre cas, le contexte est le texte et la donnée à ajouter est un mot suivant le modèle de texte.

Il existe deux méthodes selon que la donnée de départ soit la matrice de comptage ou la matrice de fréquence.

Partir de la matrice de comptage c'est connaitre le nombre d'observations de départ. On possède plus ou moins d'observations sur les objets que l'on souhaite modéliser. Il parait naturel d'accorder plus de crédit à un modèle basé sur un jeux de données important qu'à un

modèle base extrait de nombreuses observations qu'à un modèle extrait de peu. Dans ce cas le pseudo-compte représente un nombre de mots du contexte à ajouter aux données, ce qui donne plus de souplesse aux modèles calculés sur peu de données [23].

- **Définition**: Soit ∑ un alphabet fini, L un entier naturel, C une matrice de comptage, l une lettre ∈ ∑, p une position = 0…L-1 , c un pseudo-compte et fl la fréquence attendue de la lettre l. La matrice de fréquence corrigée est notée F′ et définie par la formule suivante [23].

$$F'(p,l) = \frac{C(p,l) + c * f_l}{\sum_{l \in \sum} C(p,l) + C} \qquad (3.4)$$

- **Exemple:** le calcul d'une matrice de fréquence corrigée à partir de la matrice de comptage de la figure 3.2 avec des fréquences en lettre du contexte égales à 0, 25 et un pseudo-compte égal à 1 est illustré par la figure 3.4.

A	2	3	0	0	1	2	0
C	1	1	0	0	8	0	0
G	6	6	11	11	2	6	8
T	2	1	0	0	0	3	3

La matrice de comptage de figure 3.3

A	2,25/12	3,25/12	0,25/12	0,25/12	1,25/12	2,25/12	0,25/12
C	1,25/12	1,25/12	0,25/12	0,25/12	8,25/12	0,25/12	0,25/12
G	6,25/12	6,25/12	11,25/12	11,25/12	2,25/12	6,25/12	8,25/12
T	2,25/12	1,25/12	0,25/12	0,25/12	0,25/12	3,25/12	3,25/12

La matrice de fréquence corrigée calculé à partir de la matrice de comptage

Figure 3.4: *Calcul d'une matrice de fréquence corrigée*

Partir de la matrice de fréquence c'est donner un poids, compris entre 0 et 1, au pseudocompte en ignorant le nombre de mots sur lesquels le modèle se base.

- **Définition:** Soit ∑ un alphabet fini, L un entier naturel, F une matrice de fréquence, l une lettre ∈ ∑, p une position = 0 . . .L − 1, c un pseudo-compte et f_l la fréquence attendue de la lettre l avec$\sum_{l \in \partial} f_1 = 1$. La matrice de fréquence corrigée est notée F′ et définie par la formule suivante [23]:

$$F'(p,l) = \frac{F(p,l) + c * f_l}{1 + c} \qquad (3.5)$$

- **Exemple:** le calcul d'une matrice fréquence corrigée score-position à partir de la matrice de fréquence de la figure 3.3 avec des fréquences en lettre du contexte égales à 0, 25 et un pseudo-compte

Les matrices de fréquence corrigée ou non donnent une bonne représentation des observations de départ mais indépendamment du contexte, sans tenir compte du modèle de texte.

A	0,18	0,27	0	0	0,09	0,18	0
C	0,09	0,09	0	0	0,73	0	0
G	0,56	0,56	1	1	0,18	0,56	0,73
T	0,18	0,09	0	0	0	0,27	0,27

La matrice de fréquence de figure 3.3

A	0,19	0,27	0,02	0,02	0,10	0,19	0,02
C	0,10	0,10	0,02	0,02	0,67	0,02	0,02
G	0,53	0,53	0,93	0,93	0,19	0,53	0,67
T	0,19	0,10	0,02	0,02	0,02	0,27	0,27

La matrice de fréquence corrigée calculée à partir de la matrice de fréquence

Figure 3.5 : *Calcul d'une matrice de fréquence corrigée*

3.3.4. Matrices de fréquence relative corrigée

La matrice de fréquence relative corrigée permet de remettre le motif dans son contexte. C'est à dire de rapporter le motif au modèle de texte, de mesurer l'indépendance entre le motif et le modèle de texte [23].

- **Définition:** Soit ∑ un alphabet fini, L un entier naturel, F′ une matrice de fréquence corrigé, l une lettre ∈ ∑, p une position = 0 …L−1 et f_l la fréquence attendue de la lettre l avec $\sum_{l\in\partial} f_l = 1$. La matrice de fréquence corrigée relative est notée F″ et définie par la formule suivante : [23]

$$F''(p,l) = \frac{F'(p,l)}{f_l} \qquad (3.6)$$

- **Exemple:** le calcul d'une matrice de fréquence relative corrigée à partir de la matrice de fréquence de la figure 3.5avec des fréquences en lettre du contexte égales à 0, 25 est illustré par la figure 3.6.

La matrice de fréquence relative corrigée représente les observations de départ en fonction de leur contexte mais ne permettent pas de mesurer le degré d'adéquation entre le mot et le motif.

3.3.5. Matrices d'entropies

Le contenu informationnel d'une matrice peut être mesuré par son entropie. L'*entropie* est une fonction mathématique qui correspond à la quantité d'information contenue ou délivrée par une source d'information, ici la matrice.

Une matrice construite à partir d'un ensemble de mots homogènes, voir identique, apporte beaucoup d'informations sur le motif. Inversement, lorsque les mots sont hétérogènes,

A	0,19	0,27	0,02	0,02	0,10	0,19	0,02
C	0,10	0,10	0,02	0,02	0,67	0,02	0,02
G	0,53	0,53	0,93	0,93	0,19	0,53	0,67
T	0,19	0,10	0,02	0,02	0,02	0,27	0,27

La matrice de fréquence corrigée de figure 3.5

A	0,76	1,08	0,08	0,08	0,4	0,76	0,08
C	0,4	0,4	0,08	0,08	2,68	0,08	0,08
G	2,12	2,12	3,72	3,72	0,76	2,12	2,68
T	0,76	0,4	0,08	0,08	0,08	1,08	1,08

La matrice de fréquence relative corrigée calculée à partir de la matrice de fréquence corrigée

Figure 3.6: *Calcul d'une matrice de fréquence corrigée*

spécialement lorsque les lettres des positions sont équi-réparties, l'information portée par le modéle est faible. Statistiquement la matrice d'entropie est définie comme suit [23].

- **Définition:** Soit $\sum$ un alphabet fini, L un entier naturel, F′ une matrice de fréquence corrigée, l une lettre $\in \sum$ et p une position = 0 … L−1. La matrice d'entropie est notée E et définie par la formule suivante.

$$\boldsymbol{E(p,l) = F'(p,l) \ln F'(p,l)} \qquad (\mathbf{3.7})$$

Cette formule d'entropie est indépendante du modèle de texte, elle donne une bonne estimation de l'adéquation du motif au texte lorsque les lettres sont équi-réparties. Dans le cas contraire il vaut mieux prendre en compte le contexte, le modèle de texte. Une matrice calculée à partir d'un ensemble de mots différents du modèle de texte apporte plus d'information sur le motif qu'une matrice calculée à partir d'un ensemble de mots suivant le

modèle. La formule de l'entropie relative [25] prend en compte le modèle de texte et est définie comme suit.

- **Définition** (**La matrice d'entropie relative**) Soit ∑ un alphabet fini, L un entier naturel, F′ une matrice de fréquence corrigée, F″ la matrice de fréquence relative corrigée de F′, l une lettre ∈ ∑, p une position = 0 . . .L−1 et f_l la fréquence attendue de la lettre l. La matrice d'entropie est notée E et définie par la formule suivante.

$$\boldsymbol{E(p,l) = F'(p,l) \ln F''(p,l) = F'(p,l) ln \frac{F'(p,l)}{f_l}} \qquad (\mathbf{3.8})$$

Chaque lettre contribue selon sa fréquence à l'entropie d'une position. L'entropie d'une position p d'une matrice de fréquence F, notée E_p, est définie comme suit.

$$E_p(F) = \sum_{l \in \sum} E(p,l) \qquad (3.9)$$

Chaque position contribue indépendamment à l'entropie totale. L'entropie d'une matrice, notée E_p, est définie comme suit.

$$E(F) = \sum_{p=0}^{L-1} \sum_{l \in \sum} E(p,l) \qquad (3.10)$$

La matrice d'entropie donne une bonne mesure de l'adéquation entre un mot et le motif [23].

3.3.6. Matrice score-position

La matrice score-position est un modèle très proche de la matrice d'entropie. Elle est obtenue par la méthode du log-ratio qui consiste à faire, pour chaque élément de la matrice, le logarithme du ratio entre la fréquence du motif et la fréquence attendue. Elle est un modèle de score additif. Elle augmente la fiabilité du calcul des probabilités des mots du motif en transformant le modèle multiplicatif de la matrice de fréquence en un modèle additif.

Elle permet la comparaison du motif à son contexte par la création d'un système de score, les éléments positifs de la matrice correspondent aux éléments de fréquence du motif supérieur à la fréquence du contexte, inversement pour les éléments négatifs [23].

- **Définition:** Soit F″ une matrice de fréquence relative corrigée, ∑ un alphabet fini et L un entier naturel. Une matrice score-position notée M est une matrice L×|∑| telle que pour toute lettre l ∈ ∑ et toute position p ∈ {0 . . .L−1} l'élément d'indices l et t de M noté M(p, l) est défini par la formule suivante [23].

$$M(p,l) = \ln F''(p,l) \qquad (3.11)$$

- **Exemple:** le calcul d'une matrice score-position à partir de la matrice de fréquence relative corrigée de la figure 3.6 est illustrée par figure 3.7

Les motifs modélisés par des matrices score-position expriment toute l'ambiguïté et la complexité d'un ensemble de mots. Ils mesurent l'adéquation entre un mot et le motif.

A	0,76	1,08	0,08	0,08	0,4	0,76	0,08
C	0,4	0,4	0,08	0,08	2,68	0,08	0,08
G	2,12	2,12	3,72	3,72	0,76	2,12	2,68
T	0,76	0,4	0,08	0,08	0,08	1,08	1,08

La matrice de fréquence relative corrigée de la figure 3.6

A	-0,29	0,08	-2,49	-2,49	-0,88	-0,29	-2,49
C	-0,88	-0,88	-2,49	-2,49	8	-2,49	-2,49
G	0,73	0,73	1,32	1,32	-0,29	0,73	1,01
T	-0,29	-0,88	-2,49	-2,49	-2,49	0,08	0,08

La matrice score-position calculée à partir de la matrice de fréquence relative corrigée de la figure 3.6

Figure 3.7: *Calcul d'une matrice score-position à partir d'une matrice de fréquence relative Corrigée*

3.4. De PSSM vers automate pondéré

Maintenant que nous avons défini le modèle des matrices score-position à partir d'un alignement multiple nous présentons comment construira leur équivalente automate pondéré alors Soit le quintuplet (Q,∑,E, I, T) défini comme suivant :

Q : l'ensemble des positions possible qui présentent les états union {F} tell que F un état finale

∑ : C'est ensemble les lettres

E : présente tout transition entre deux positions adjacentes

I : c'est la position initiale

T : C'est ensemble {F}

Et pour les opérations sur les automates, on a seulement la fusion définis comme suivants :

Soit l' automates pondéré (Q,∑,E, I, T) la fusion de deux automates pondéré (Q1,∑1,E1, I1, T1) et (Q2,∑2,E2, I2, T2) tel que la :

Q =Q1 U Q2

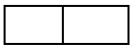
∑ : ∑1

E : E1 U E2

I : I1

T : T1UT2

Exemple:

Soit un motif à deux cases suivant 

Les cas de motif sont A = {0,0}, B = {0,1}, C = {1,0}, D = {1,1}

Il existe deux images qui représentent sur un matrice et détermine à deux classes suivant:

Deux images représentent par des zone, chaque zone représenter par la motif, donc dans ce cas en représente l'image par les matrice suivant:

Img1

0	1
2	3

Img2

0	1
2	3

	0	1	2	3
A	0.4	0.2	0.1	0.1
B	0.2	0.3	0.2	0.1
C	0.1	0.3	0.4	0.1
D	0.3	0.2	0.3	0.7

	0	1	2	3
A	0.5	0.2	0.2	0.2
B	0.1	0.2	0.1	0.2
C	0.1	0.2	0.4	0.5
D	0.3	0.4	0.3	0.1

Dans le premier cas en à**:** Automate de PSSM de la classe 01

- Q1 = {00, 01, 02, 03, f1} ;
- ∑= {A, B, C, D} ;
- E1 = {(00, A, 01, 0.4), (00, B, 01, 0.2), (00, C, 01, 0.1), (00, D, 01, 0.3),
  (01, A, 02, 0.2), (01, B, 02, 0.3), (01, C, 02, 0.3), (01, D, 02, 0.2),
  (02, A, 03, 0.1), (02, B, 03, 0.2), (02, C, 03, 0.4), (01, D, 02, 0.3),
  (03, A, F1, 0.1) , (03, B, F1, 0.1), (03, C, F1, 0.1), (03, D, F1, 0.7)};
- I1 = {00} ;
- T1 = {f1};

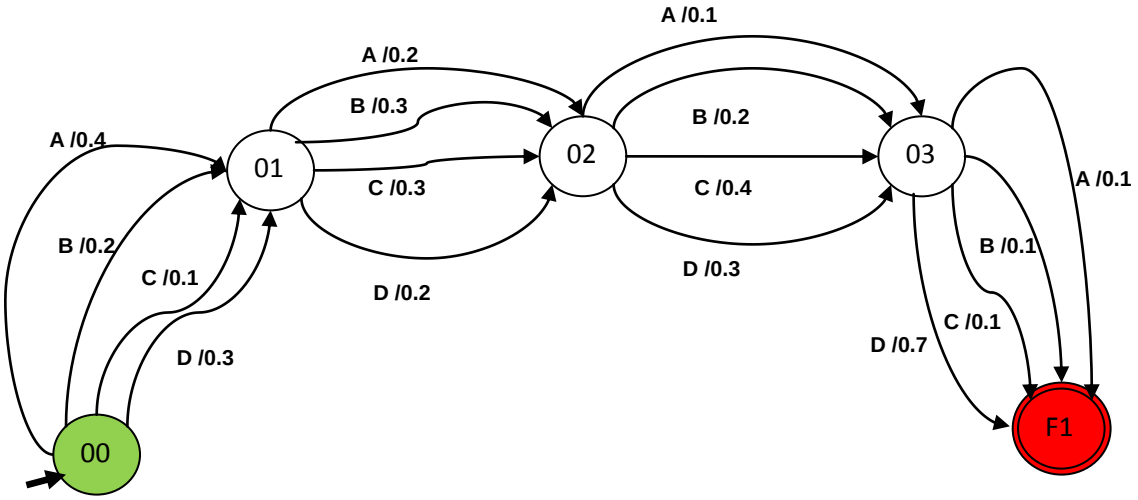


Figure 3.8: *Exemple de représentation de PSSM de classe 1*

Dans le deuxième cas en à**:** Automate de PSSM de la classe 01

- Q2 = {0, 11,12, 13, f2} ;
- ∑= {A, B, C, D} ;
- E 2= {(00, A, 11, 0.5), (00, B, 11, 0.1), (00, C, 11, 0.1), (00, D, 11, 0.3),
  (01, A, 12, 0.2), (01, B, 12, 0.2), (01, C, 12, 0.2), (01, D, 12, 0.4),
  (02, A, 13, 0.2), (02, B, 13, 0.1), (02, C, 13, 0.4), (01, D, 13, 0.3),
  (03, A, f2, 0.2) , (03, B, f2, 0.2), (03, C, f2, 0.5), (03, D, f2, 0.1)};
- I2= {00} ;
- T2 = {f1};

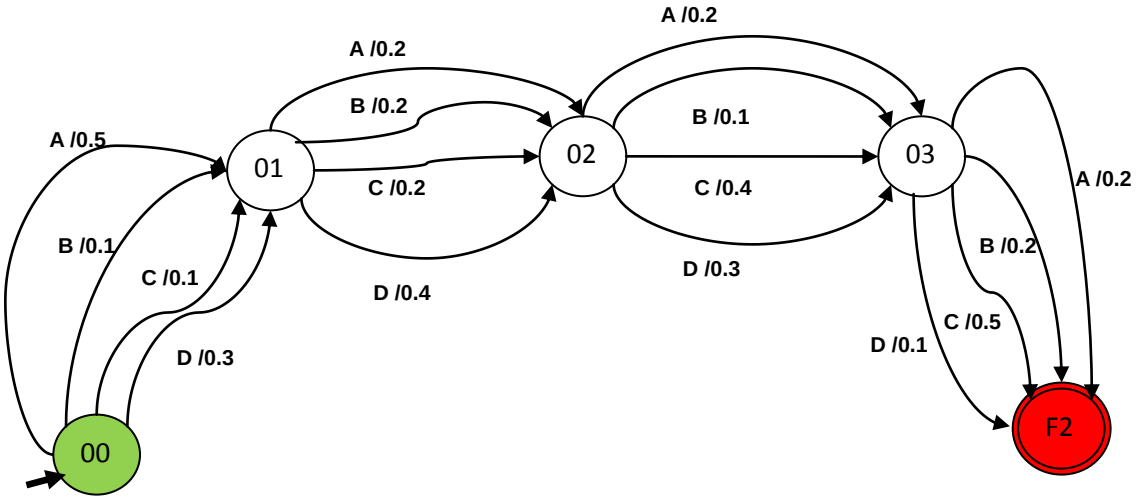


Figure 3.9: *Exemple de représentation de PSSM de classe 2*

La fusion des automates de PSSM de la classe 01 et la classe 02

- Q = Q1 U Q2 ={00, 01,0 2,03, 11, 12, 13, f1, f2} ;
- ∑= {A, B, C, D} ;
- E = E1 U E2= {(00, A, 01, 0.4), (00, B, 01, 0.2), (00, C, 01, 0.1), (00, D, 01, 0.3), (01, A, 02, 0.2), (01, B, 02, 0.3), (01, C, 02, 0.3), (01, D, 02, 0.2), (02, A, 03, 0.1), (02, B, 03, 0.2), (02, C, 03, 0.4), (01, D, 02, 0.3), (03, A, F1, 0.1) , (03, B, F1, 0.1), (03, C, F1, 0.1), (03, D, F1, 0.7), (00, A, 11, 0.5), (00, B, 11, 0.1), (00, C, 11, 0.1), (00, D, 11, 0.3), (01, A, 12, 0.2), (01, B, 12, 0.2), (01, C, 12, 0.2), (01, D, 12, 0.4), (02, A, 13, 0.2), (02, B, 13, 0.1), (02, C, 13, 0.4), (01, D, 13, 0.3), (03, A, f2, 0.2) , (03, B, f2, 0.2), (03, C, f2, 0.5), (03, D, f2, 0.1)};
- I = {00} ;
- T = T1 U T2 ={f1, f2};

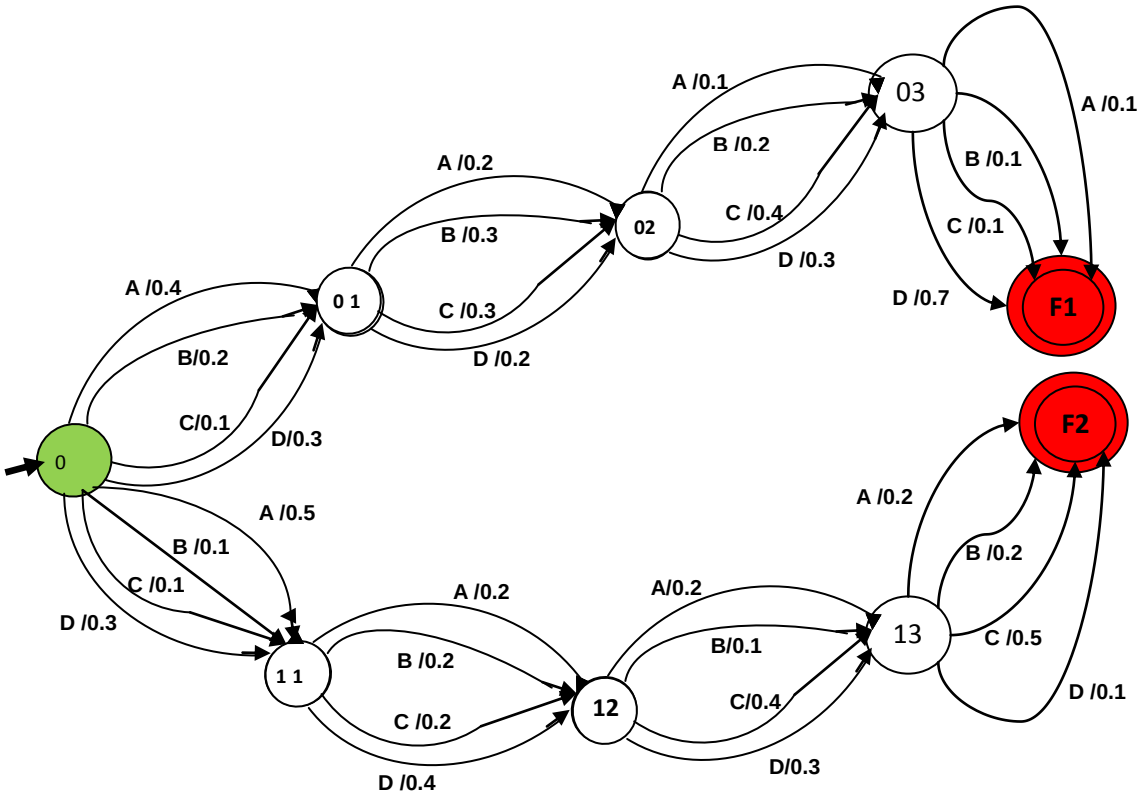


Figure 3.10: *Exemple de représentation de* PSSM *de classe 1 et classe 2*

3.5. Conclusion

Les automates à états finis pondérés constituent un cadre qui permet d'unifier les différents modules qui composent un système de reconnaissance, en les regroupant au sein d'un formalisme unique [26]. Nous avons essayé de trouve un relation entre l'automate pondéré qui est trés utiles pour calculer les noyaux de langages rationnels avec le PSSM qui utilise dans le domaine scientifique

Chapitre 04
Application et résultats

4.1. Introduction

Nous présentons dans ce chapitre une description détaillée de notre application de segmentation de mot manuscrit arabe et de reconnaissance de caractères manuscrits, avec l'évaluation de performance de chaque phase. Nous avons appliqué ce système sur base de données d'images de noms de villes tunisiennes.

4.2. Principales bases de données utilisées

M. Pechwitz et al introduisent la base IFN/ENIT en 2002 .Il s'agit d'une base de données d'images de noms de villes tunisiennes. Outre la séquence de lettres, sont également annotées la forme que prend chacune des lettres au sein du mot, la présence des signes diacritiques secondaires, et une approximation de la ligne de base [26].

4.3. Ressources matérielles et logicielles

4.3.1. Ressources matérielles

Notre système est développé dans un ordinateur dont les caractéristiques techniques, sont les suivantes**:**

Composant	Description
Processeur	Intel « I3 »
RAM	2 Go

Tableau 4.1: Représentation des caractéristiques techniques de l'ordinateur de développement.

4.3.2. Ressources logicielles

- Système d'exploitation : Windows 7 professionnel 32 bit.
- Langage de programmation utilisé : **Matlab** car:

Le logiciel Matlab constitue un système interactif et convivial de calcul numérique et de visualisation graphique. Destiné aux ingénieurs, aux techniciencs et aux scientifiques, c'est un outil trés utilisé, dans les universités comme dans le monde industriel, qui intègre des centaines de fonctions mathématiques et d'analyse numérique (calcul matriciel —le MAT de Matlab—, traitement de signal, traitement d'images, visualisations graphiques, etc.) [27].

4.4. Description de notre système:

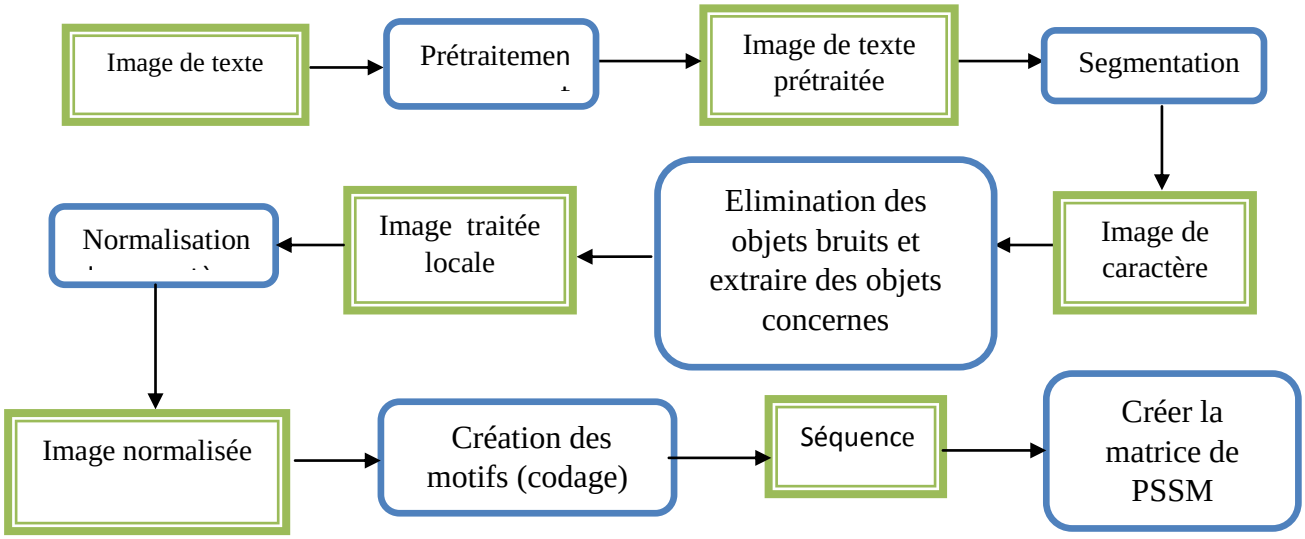


Figure: 4.1 : Organigramme de notre système**.**

4.4.1. Prétraitement:

Nous utilisons deux types de filtrages:"sobel" et "médian".

Par exemple l'image de mot "الشرايع".

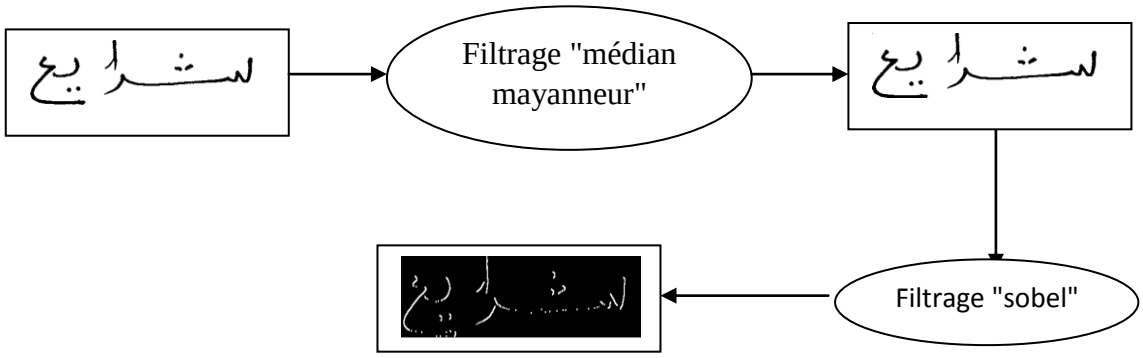


Figure 4.2:Exemple de Prétraitement.

4.4.2. Segmentation:

Pour extraire les caractères on fait la segmentation de chaque image à l'aide des histogrammes (verticale et horizontale).

A. Algorithme de segmentation:

1. Sélectionner une image.
2. A=Filtrer par médian mayaneur de l'image .
3. B=Filtrer par" sobel"de(A).

4. B=Filter par "médian moyaneur" de (B).
5. V_h=Calculer l'histogramme verticale de (B).
6. Calculer le seuil.
7. Possible de segment=0.
8. j=1.
9. Pour i de 1 à taille de (V_h) faire

   Si (V_h <seuil) et (possible de segment=1) alors

   Région (j, 2)=i.

   Possible de segment=0.

   j=j+1.

   Sinon Si (V_h(i)>=seuil) et(possible de segment=0) alors

   Region(j,1)=i.

   Possible de segment=1.

   Finsi.

   Finsi.

10. Fin pour.
11. Pour jj de 1 à taille de region faire

    C=A(region(1,1),region(1,2))
    C=filtre par "sobel"de (C).
    C=filtre par "median moyaneur" de(C)
    H_h=calculer l'histogramme horizentale.
    Clculer la seuil2.
        Pour i de 1 à taille de (H_h) faire
        Si ((H_h)<seuil2) alors
        Si (H_h(i+1)>seuil2)alors ségmenter.
        Finsi.
        Finsi.
        Fin pour.
12. Fin pour.

B. les résultats de segmentation:

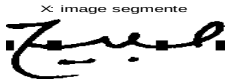
Pourcentage	**Nombres des images**	**Exemple**	
		Code image	**Image**
Inferieur à 25%	45 (7.90%)	ae19_025	
Entre 25 et 50%	115 (20.21%)	ai14_029	
Entre 50 et 75%	246 (43.23%)	aq34_055	
Entre 75 et 100%	163 (28.54%)	ae70_040	

Tableau 4.2: *Les résultats de segmentation.*

C. Les problèmes de segmentation:

☒ Ligatures verticales connectées ou non**.**

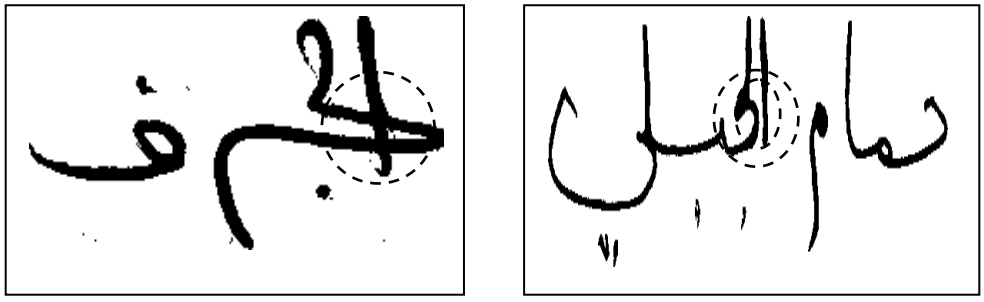


Figure 4.3: Exemple de ligatures verticales connectées ou non.

☒ Liaison indésirable entre caractères.

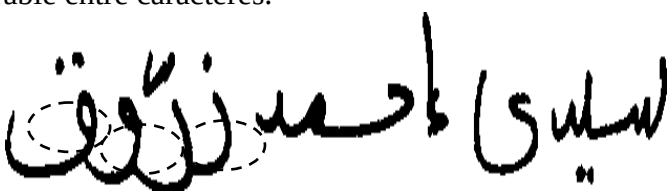


Figure 4.4: *Exemple de liaison indésirable entre caractères.*

☒ Coupure indésirable

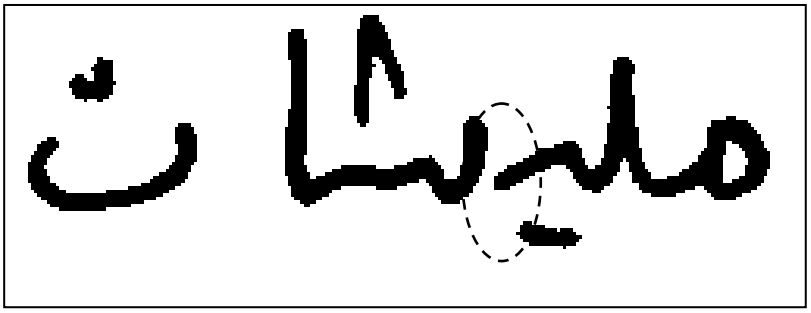


Figure 4.5: Exemple de coupure indésirable.

D. Evaluation des résultats de segmentation :

Malgré les efforts et les travaux intensifs réalisés dans le domaine de la segmentation de mots manuscrits, aucune méthode ne trouve pas des problèmes. Mais au fur et à mesure les auteurs essayent d'améliorer les résultats.

4.4.3. Elimination des objets bruits et extraire des objets concernes:

Dans cette étape nous faisons une élimination des éléments bruis qu'ont des interfaces inferieur au seuil et extraire l'objet concerne.

Exemple:

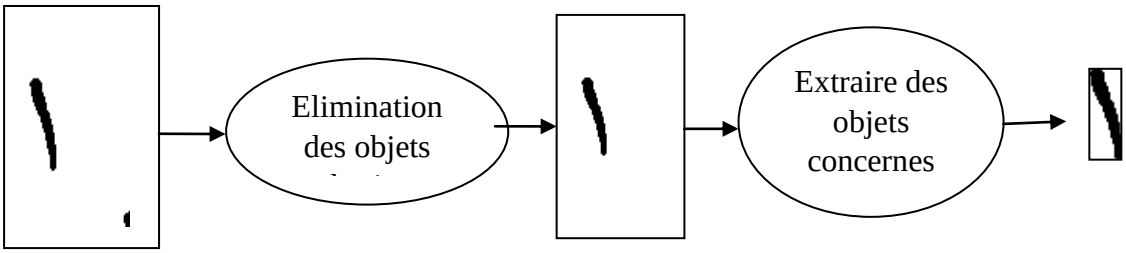


Figure 4.6 : *Exemple d'élimination des objets bruits et extraire des objets concernes.*

4.4.4. Normalisation des images

Nous avons effectué cette opération pour éliminer les conditions qui peuvent fausser les résultats, comme la différence de taille. Après la normalisation de la taille, les images de tous les caractères se retrouvent définies dans une matrice de même taille.

Exemple:

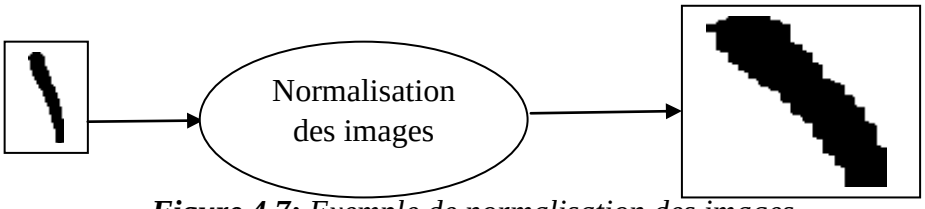
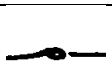


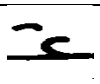
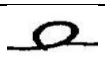
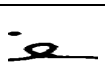
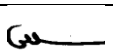
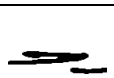
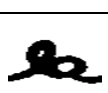
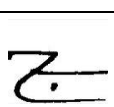
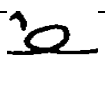
Figure 4.7: *Exemple de normalisation des images.*

4.4.5. Création des motifs (codage):

Dans cette phase :

- Nous utilisons 86 classes de lettres arabes :

Classe	Exemple	Classe	Exemple
Alif_isolé		Mim_début	
Alif_médiale_finale		Mim_finale	
Ayn_début		Mim_isolé	
Ayn_finale		Mim_médiale	
Ayn_isolé		Noun_début	
Ayn_médiale		Noun_finale	
Ba_debut		Noun_isolé	
Ba_finale		Noun_médiale	
Ba_médiale		Qaf_début	
Chin_debut		Qaf_finale	
Chin_finale		Qaf_isolé	
Chin_médiale		Qaf_médiale	
Dâd_début		Ra_finale_médiale	

Dâd_finale		Ra_isolé	
Dâd_médiale		Rayn_début	
Dal_isolé		Sâd_début	
Dal_médiale_finale		Sâd_finale	
Dhal_médiale_finale		Sâd_médiale	
Fa_début		Sin_début	
Fa_médiale		Sin_finale	
Ha_debut		Sin_isolé	
Ha_finale		Sin_médiale	
Ha_isolé		ta_almarbouta	
Ha_isolé		Ta_debut	
Hamza_nabira		Tâ_debut	
HHa_début		Ta_finale	
HHa_finale		Tâ_finale	
HHa_médiale		Tâ_isolé	
Jim_debut		Ta_médiale	
Jim_finale		Tâ_médiale	

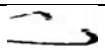
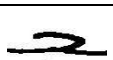
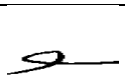
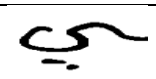
Jim_isolé		Tha_debut	
Jim_médiale		Tha_finale	
Kaf_début		Tha_médiale	
Kaf_médiale		Waw_isolé	
Kha_debut		Waw_médiale_finale	
Kha_finale		Ya_début	
Kha_isolé		Ya_finale	
Kha_médiale		Ya_isolé	
lam_alif		Ya_médiale	
lam_début		Zâ_debut	
lam_finale		Zâ_médiale	
lam_isolé		Zay_finale_médiale	
lam_médiale		Zay_isolé	

Tableau 4.3: *les classes de lettres arabes utilisées.*

❖ Nous découpons chaque image par des petites zones selon le motif:(voir le Tableau 4.4)**.**

a) Si nous posons le motif (1,1) il donne 4 classes**.**

b) Si nous posons le motif (2,2) il donne 16 classes**.**

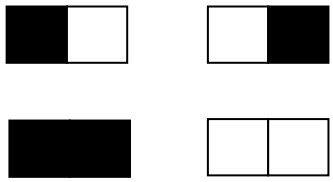
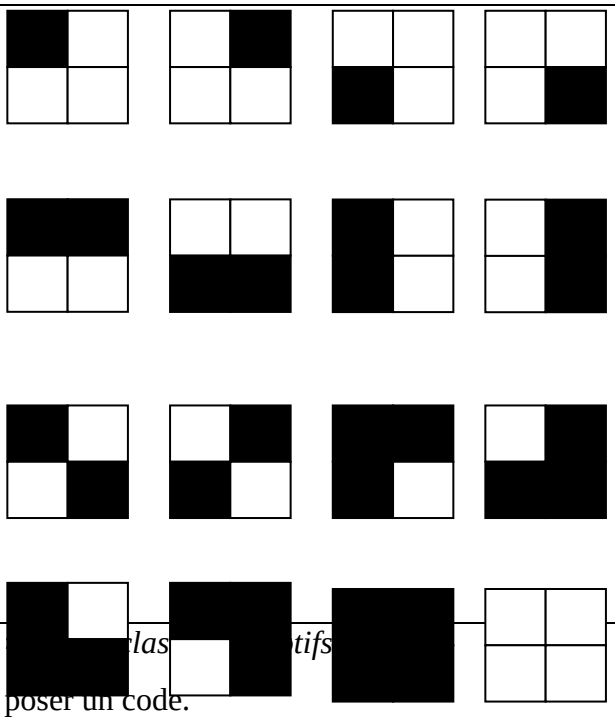
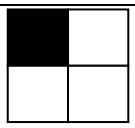
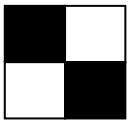
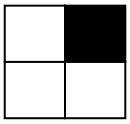
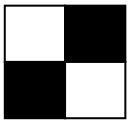
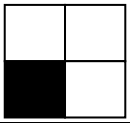
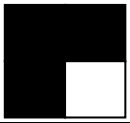
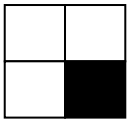
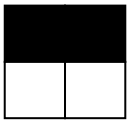
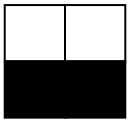
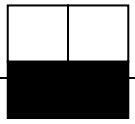
Les classes de motif(1,1)	Les classes de motif(2,2)
	

***Tableau** Les classes de motifs*

- Pour créer la séquence il faut poser un code.
- Par exemple si nous choisissons motif(2,2)

Les classes de motif	code	Les classes de motif	code
	A		I
	B		J
	C		K
	D		L
	E		M
	F		N
	G		O

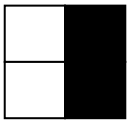
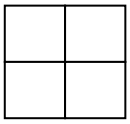
	H		P

Tableau 4.5: *code de motif(2,2).*

Exemple: séquence de d'image.

L'image	Séquence
	PPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPDJJOOOO OOOOONOOLOONOOIOOIOOIOOIOOIOKIOMIOKIOGIOGIO GIOPIOPOKPBGPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPP PPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPP

Tableau 4.6:*exemple de séquence.*

4.4.6. Création de PSSM: Dans cette phase nous créons une matrice que associé à chaque code une fréquence selon sa position.

- Exemple des séquences:

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Sq1	P	P	A	E	F	D	P	N	O	K	M	N	J	P
Sq2	B	C	A	P	P	P	M	D	F	L	N	E	C	D
Sq3	E	P	P	H	I	E	D	C	G	L	N	H	P	H
Sq4	D	C	I	G	J	O	C	M	P	O	L	E	I	J
Sq5	O	O	I	D	C	B	A	N	M	L	H	D	N	L

Tableau 4.7:*exemple de séquences.*

- Exemple de PSSM:

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**

A	0	0	0.14	0	0	0	0.07	0	0	0	0	0	0	0
B	0.07	0	0	0	0	0.07	0	0	0	0	0	0	0	0
C	0	01.4	0	0	0.07	0	0.07	0.07	0	0	0	0	0.07	0
D	0.07	0	0	0.07	0	0.07	0.07	0.07	0	0	0	0.07	0	0.07
E	0.07	0	0	0.07	0	0.07	0	0	0	0	0	0.14	0	0
F	0	0	0	0	0.07	0	0	0	0.07	0	0	0	0	0
G	0	0	0	0.07	0	0	0	0	0.07	0	0	0	0	0
H	0	0	0	0.07	0	0	0	0	0.07	0	0.07	0.07	0	0.07
I	0	0	0.14	0	0.07	0	0	0	0	0	0	0	0.07	0
J	0	0	0	0	0.07	0	0	0	0	0	0	0	0.07	0
K	0	0	0	0	0	0	0	0	0	0.07	0	0	0	0
L	0	0	0	0	0	0	0	0	0	0.21	0.07	0	0	0.7
M	0	0	0	0	0	0	0	0.07	0.07	0	0.07	0	0	0
N	0	0	0	0	0	0	0	0.14	0	0	0.14	0.07	0.07	0
O	0.07	0.07	0	0	0	0.07	0	0	0.07	0.07	0	0	0	0
P	0.07	0.07	0.07	0.07	0.07	0.07	0.07	0	0.07	0	0	0	0.07	0.07

Tableau 4.8: *exemple de PSSM.*

❖ **Exemple2:** PSSM d'Alif_médiale_finale par motif(3,3).

	1	**2**	**3**	**4**	**5**	**6**	**……**	**900**
1	0.0004	0.0013	0.0022	0.0016	0.0019	0.0019	……	0.0021
2	0	0	0	0	0	0	……	0
3	0	0	0	0	0	0	……	0
4	0	0	0	0	0	0	……	0
5	0	0.0065	0.0130	0.0130	0	0	……	0
512	0.0009	0.0006	0.0006	0.0006	0.0005	0.0006	……	0.0004

Tableau 4.9:*exemple de PSSM d'Alif_médiale_finale.*

4.5. Reconnaissance

Pour identifier un caractère il faut passer à des étapes:

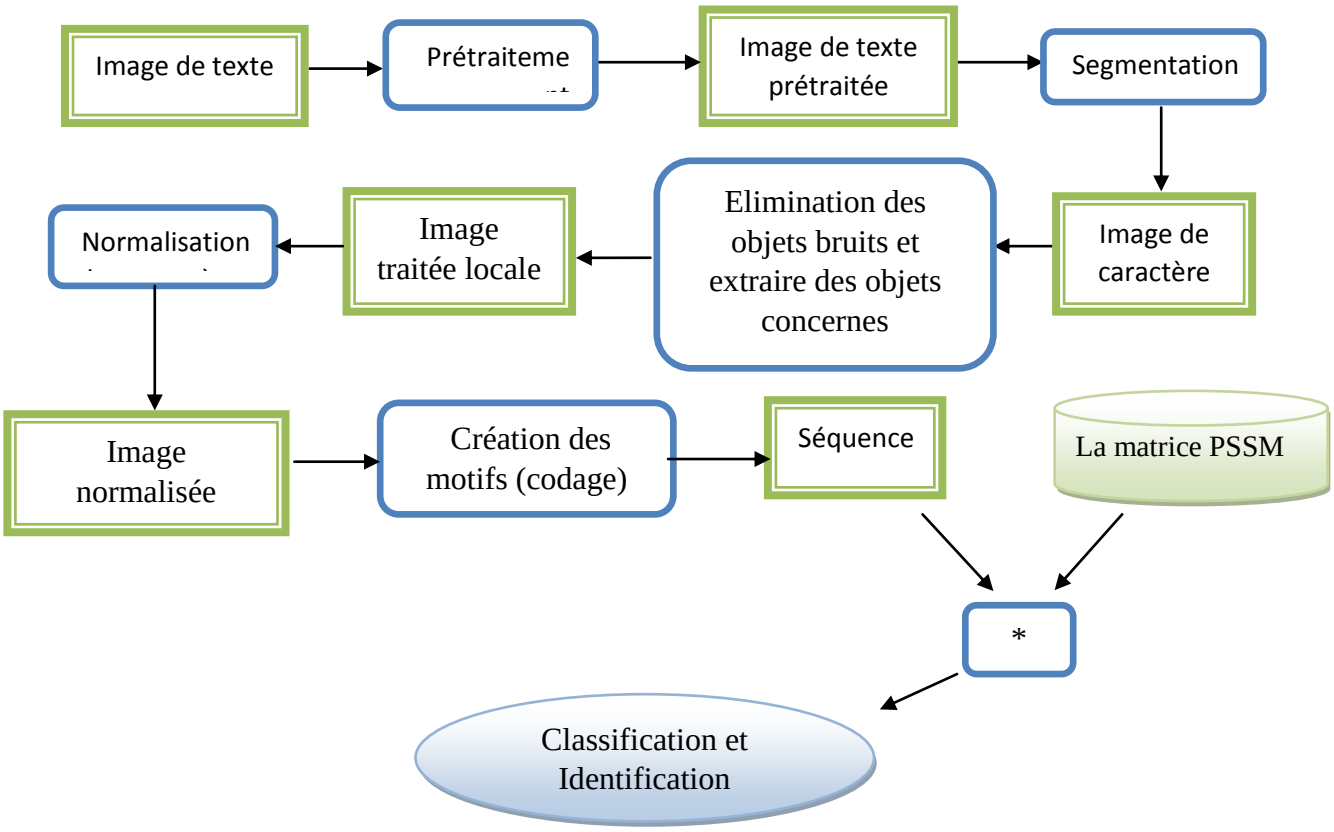


Figure 4.8: *Organigramme de reconnaissance.*

4.5.1. Classification et identification

❖ Nous préparons des matrices de PSSM déjà. Alors nous comparons la séquence avec ces matrices et nous calculons la probabilité.

❖ **Exemple:**

On pose :

- une séquence d'une image :(BCDA).
- Deux MSSP de deux classes.

$PSSM_1$ de classe C_1 et $PSSM_2$ de classe C_2.

$PSSM_1$:

	1	2	3	4
A	0.4	0.2	0.1	0.1
B	0.2	0.3	0.2	0.1

$PSSM_2$:

	1	2	3	4
A	0.5	0.2	0.2	0.2
B	0.1	0.2	0.1	0.2

C	0.1	0.3	0.4	0.1		C	0.1	0.2	0.4	0.5
D	0.3	0.2	0.3	0.7		D	0.3	0.4	0.3	0.1

Tableau 4.10:PSSM de classeC_1 et PSSM de classeC_2.

On va calculer la probabilité de chaque symbole (BCDA).

PSSM1:

P1=P (B) +P(C) +P (D) +P (A).

P1=0.2+0.3+0.3+0.1.

P1=0.9.

MSSP2:

P2=P (B) +P(C) +P (D) +P (A).

P2=01+02+03+02

P2=0.8.

P2>P1 Alors l'image de séquence (BCDA) appartienne à la classe C_1.

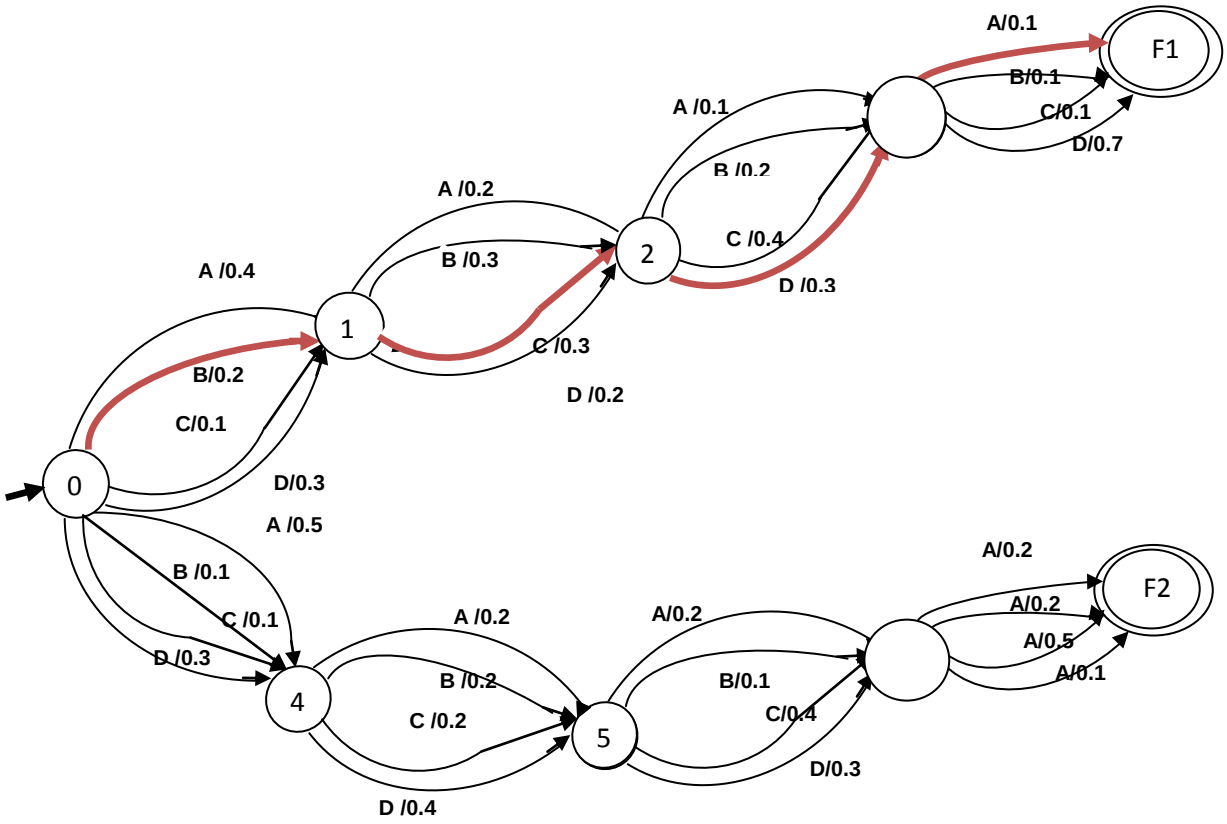


Figure 4.9: *représentation de la route de séquence(BCDA) dans l'automate.*

4.5.2. Analyse les résultats:

A. Tester les démentions:

- Pour faire l'analyse on va fixe le nombre motif et faire varier les démentions.
- On a des résultats semblables comme moyenne entre 99.73 et 100.

Démentions	**Moyen de taux final**
(60,60)	99.73%
(90,90)	99.91%
(120,120)	100%
(150,150)	100%

Tableau 4.11 : *les données de tester.*

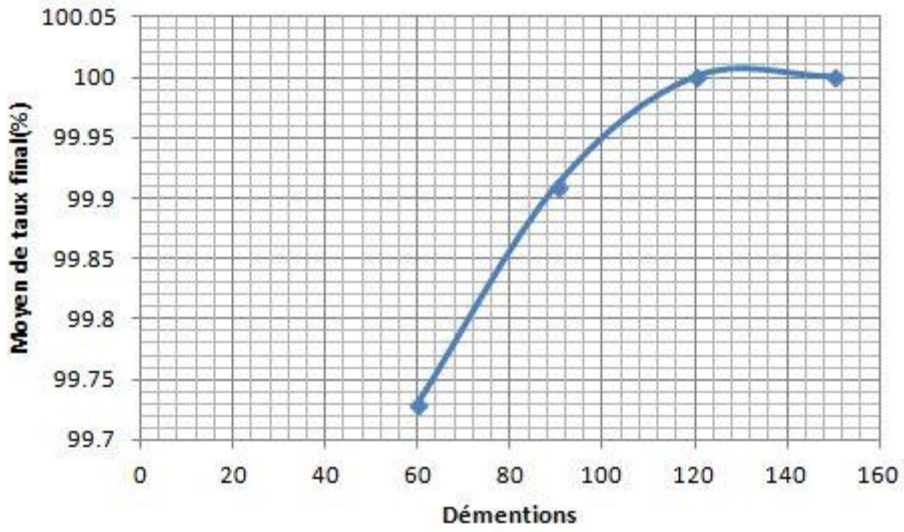


Figure 4.10 : *graphe de taux final.*

- Nous observation une relation directe entre le nombre de motif d'image et le taux de reconnaissance.

B. Tester des motifs:

Pour faire l'analyse on va fixe les démentions et varie les motifs.

On a des résultats semblables comme moyenne entre 98.61et 99.91.

Motif	Taux final
(3,3)	99.91
(2,2)	98.61

Tableau 4.12: *les données de tester le motif.*

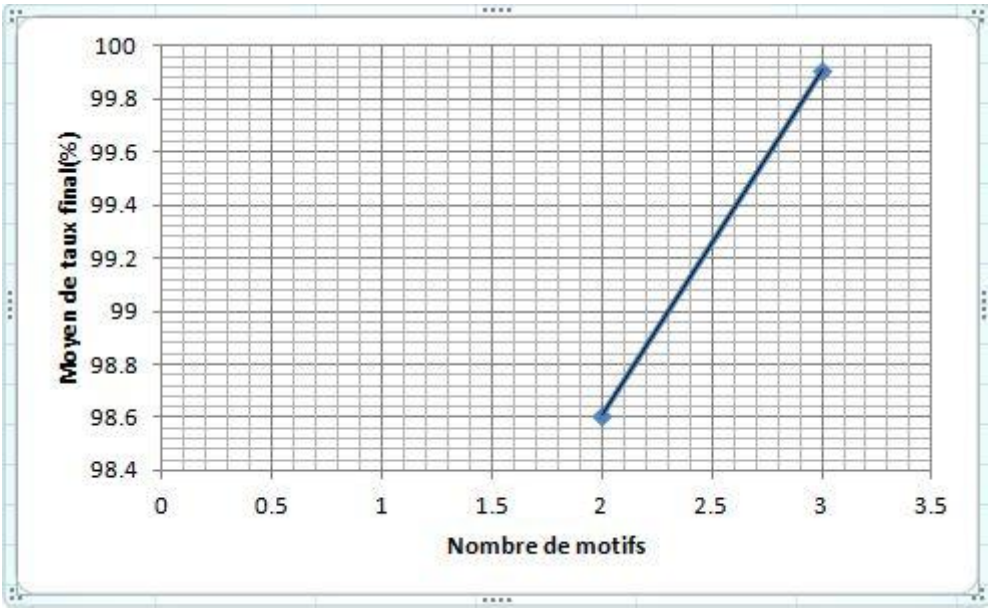


Figure 4.11: *graphe de taux final.*

- Nous observons une relation directe entre le motif et le taux de reconnaissance.

4.6. Conclusion

Nous avons présenté dans ce chapitre une description détaillée de notre système de reconnaissance de caractères manuscrits par la présentation de chaque phase de processus de reconnaissance**.** On a réalisé une nouvelle méthode utilisant l'automate pondéré basé sur une modélisation matricielle sous la matrice PSSM et qui nous a donné des bons résultats.

Conclusion générale

Conclusion générale

Nous avons présenté une approche du problème de la reconnaissance d'écriture manuscrite et son application à une tâche de reconnaissance de caractères manuscrits arabe. Les apports principaux de ce système sont l'utilisation d'automate pondéré et fondée sur une modélisation par les matrices poids-position PSSM (*position specific scoring matrices*) qui permettent une vraie modélisation bidimensionnelle de l'écriture. L'originalité de ce travail réside aussi dans l'extraction de primitives issues d'une analyse par fenêtrage. Ces deux principes sont valides par des résultats satisfaisants, et des performances intéressantes ont été obtenues sur le traitement des caractères malgré un nombre d'images d'apprentissage très restreint.

Lors de nos travaux, nous avons étudie l'influence des différents paramètres sur le taux de reconnaissance. Cependant, un paramétrage peut être plus efficace pour une partie des échantillons de la base par exemple la taille de fenêtre d'analyse ou la dimension d'image normalisé.

Les perspectives ouvertes par ce travail sont très nombreuses et nécessiteront, pour entre explorées, de très grandes bases de données pour l'apprentissage et le test des modèles. À propos de paramétrage c'est mettre en place une stratégie à plusieurs classifieurs, dont les paramètres sont différents, utilisés en parallèle puis déterminer la classe en combinant les résultats obtenus. Ainsi qu'une proposition pour l'adaptation de cette approche à une tâche de reconnaissance de mots manuscrits.

Référence

Référence:

[1]: CHIKH Mohammed Tahar. Amélioration des images par un modèle de réseau de neurones (Comparaison avec les filtres de base). Mémoire de fin d'études pour l'obtention du diplôme de Master en Informatique. Université Abou-Bakr Belkaid -Tlemcen- 2011.

[2]: BENBA Zeyneb BENBA Fatouma. L'utilisation la transformée de Hough pour la Segmentation des images d'Iris. Mémoire de fin d'études pour l'obtention du diplôme de Licence en Informatique, Option**(**Ingénierie des systèmes d'information et de logiciels). Université d'Adrar 2013/2014.

[3]: Melle MEDJAHED Fatiha, Détection et Suivi d'Objets en Mouvement Dans Une Séquence d'Images. Mémoire en vue de l'obtention du diplôme de Magister, Spécialité (Electronique) Option (Signaux et Systèmes), Université des Sciences et de la Technologie d'Oran U. S. T. O.

[4]: Melle BELAROUCI Sara & Melle BENMOKHTAR Sara. Méthode coopérative pour la segmentation d'images IRM cérébrales basée sur les techniques FCM et Level Set. Mémoire Pour l'obtention du diplôme de MASTER en Génie Biomédical. Université Abou-Bakr Belkaid –Tlemcen-2012

[5]: HOCINI Lotfi, conception de méthaheuristiques d'optimisation pour la segmentation des images de télédétection, Mémoire Pour l'obtention du diplôme de Magister en Electronique, Université MOULOUD MAMMERI, TIZI-OUZOU 2012

[6]: Houaria ABED, Lynda ZAOUI, système D'Indexation et de Recherche d'Images par le contenu, Université des Sciences et de la Technologie d'Oran - Mohamed Boudiaf –

[7]: BENMOHAMED Abderrahim, Une Approche semi-automatique pour l'indexation de documents anciens, MEMOIRE Présenté de l'obtention du diplôme de MAGISTER, UNIVERSITE BADJI MOKHTAR – ANNABA

[8]: Ricardo da Silva Torres, Alexandre Xavier Falcão, "Content-Based Image Retrieval: Theory and Applications"

[9] **BOUGAMOUZA Fateh**, Contribution à la reconnaissance automatique de l'écriture manuscrite arabe, application sur les montants littéraux des chèques**.** 2008-2009.

[10] Mr DJEDDI Chawki. Contribution à l'analyse et la caractérisation de l'écriture manuscrite. Année 2013-2014.

[11] A. BOUCENNA, "ON THE ORIGIN OF THE ARABIC NUMERALS" , Département de Physique, Faculté des Sciences, Université Ferhat Abbas 19000 Sétif, Algérie.

[12] Najoua Ben Amara1, Abdel Belaïd2 et Noureddine Ellouze ,"Utilisation des modèles markoviens en reconnaissance de l'écriture arabe" : Etat de l'art..

[13] BOUKHAROUBA Abdelhak. "Contribution à la segmentation et à la reconnaissance de l'écriture arabe manuscrite". 2011.

[14] HAITAAMAR. Schahrazed ,Segmentation de textes en caracteres pour la reconnaissance optique de l'ecriture arabe., 2007.

[15] DJEDDI Chawki, SOUICI-MESLATI Labiba ", Identification de scripteurs pour l'écriture arabe par une approche locale". 2010.

[16] Mr : Azizi Rebiai. "Une approche hybride pour la reconnissance d'ecriture arabe manuscrite". 2006/2007.

[17] Leila Chergui, Maamar Kef, Mohammed Benmohammed ,"La Théorie de la Résonance Adaptative et les Moments de Zernike pour la Reconnaissance de Mots Arabes Manuscrits".

[18] Salima Nebti**,** Reconnaissance de Caractères Manuscrits par Intelligence Collective, these pour l'obtention du diplôme de Doctorat en science, Option : Informatique, Soutenu le : O7/mars/2013

[19]: C. Fang, "Deciphering Algorithms for Degraded Document Recognition", Ph.D thesis. 1997.

[20]: T. Saba, G. Sulong, A. Rehman, "A Survey on Methods and Strategies on Touched Characters Segmentation", International Journal of Research and Reviews in Computer Science (IJRRCS), Vol. 1, No. 2, June 2010.

[21]: https://fr.wikipedia.org/wiki/Automate_pondéré , 2 novembre 2016

[22]: I.-P. Ioshikhes N.-I. Gershenzon, G.-D. Stormo. Computational technique for improvement of the position-weight matrices for the dna/protein binding sites. *Nucleic Acids Research*, 33(7) :2290–2301, 2005.

[23]: Aude LIEFOOGHE, " Matrices score-position, algorithmes et propriétés", thése pour pour l'obtention du Doctorat de l'Université des Sciences et Technologies de Lille, 2008

[24]: Julien Menana," Automates et programmation par contraintes pour la planification de personnel", THÈSE DE DOCTORAT Informatique, UNIVERSITÉ DE NANTES, *2011.*

[25]: O.G. Berg and P.H. von Hippel. Selection of DNA binding sites by regulatory proteins. Statistical-mechanical theory and application to operators and promoters. Journal of Molecular Biology, 193(4) :723–50, 1987.

[26]: Farès Menasri, "Contributions à la reconnaissance de l'écriture arabe manuscrite", thèse pour l'obtention de Doctorat de l' Université Paris Descartes juin 2008

[27]: Florent Krzakala.Premiers pas en Matlab

[28]: A.D.S.B. Jr, R.Sabourin, , F. Bortolozzi, C.Y. Suen, "A Two-Stage HMM-Based System for Recognizing Handwritten Numeral Strings". ICDAR 2001: 396-400. 2001.