République Algérienne Démocratique et Populaire

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Université de Tébessa
Faculté des Sciences Exactes et des Sciences de la Nature et de la Vie
Département:Mathématiques et Informatique

MEMOIRE DE MASTER

Domaine: Mathématiques et Informatique

Filière: Informatique

Option: Systèmes Multimédia

Thème

Analyse automatique de l'écriture manuscrite pour la détermination du sexe d'un scripteur

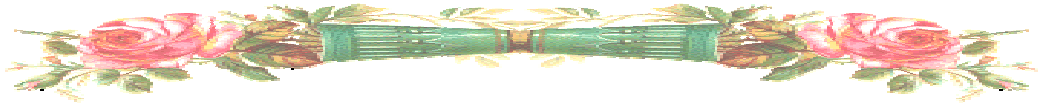
Présenté par:
Belghit Abdelhalim
Bougherara Abdelaziz

Encadrer par:
Dr: Bennour Akram
Dr: Djeddi Chawki

Soutenue le 30 Mai, devant le jury composé de

Mr. Ahmim Ahmed	MAA	Université de Tébessa	Président
Mr.Zeggari Ahmed	MAA	Université de Tébessa	Examinateur
Dr. Bennour Akram	MCB	Université de Tébessa	Encadreur
Dr. Djeddi Chawki	MCB	Université de Tébessa	Co- Encadreur

Anne universitaire 2016



Remerciement

Après avoir remercie allah le tout le

Tout puissant

Nous remercions infiniment en fortement notre encadreur

DR. BENNOUR AKRAM

DR. DJEDDI Chawki

Pour son conseil.

Aussi nous remercions tout personne qui nous a donnés l'aide et le

courage chacune avec son nom

ABDELAZIZ
Abd El Halim

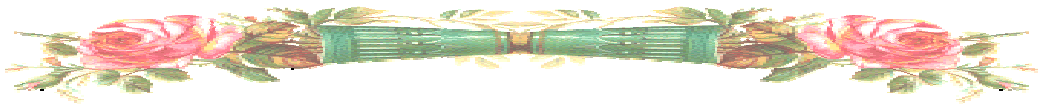


TABLE DES MATIERES

.

LISTE DES FIGURES

LISTE DES TABLEAUX

Chapitre I

Introduction et Motivation

Chapitre I
Introduction et Motivation

Introduction générale

Dans notre vie quotidienne le nombre des documents que nous utilisons va sans cesse croissant. Même si aujourd'hui ces documents prennent le plus souvent une forme électronique, il est encore bien des cas où un document, par exemple un formulaire, est en partie manuscrit. Malgré les progrès quotidiens de la technologie dans le domaine de la miniaturisation, il demeure que le stylo reste toujours le moyen d'écriture le plus mobile et qui ne nécessite aucune source d'énergie extérieure au scripteur pour être opérationnel.

D'autre part le bon usage des documents suppose que ceux-ci se trouvent à disposition au moment où ils doivent servir. Le bon aiguillage des documents pour la réalisation de la tâche les concernant et leur rangement devient de plus en plus délicat et dans bien des cas une automatisation de la classification de ces documents serait bénéfique. Dans cet objectif, une reconnaissance de la nature des documents et de leur contenu est nécessaire [CLA 01] [KEB 98].

Le problème de l'authentification d'un document est né en même temps que naissait la notion même de document. L'identification du scripteur représente une phase cruciale de cette authentification et remonte pratiquement à la naissance de l'écriture. De nos jours et dans notre environnement, la nécessité de valider un texte écrit est un problème récurrent, il n'est pas limité au monde juridique dans lequel il est fondamental de pouvoir authentifier l'auteur d'un document, d'un testament, ou d'un acte de vente par exemple. Il est beaucoup d'autres circonstances où l'authentification est indispensable. A ce titre, le sceau, la signature ou le mot de passe a été introduit pour particulariser les documents. Ces moyens représentent une première phase permettant une authentification ultérieure, mais qui n'est pas toujours effectuée au moment de la réalisation du document. Dans tous les cas, c'est à partir d'une caractérisation du scripteur que l'authentification peut être réalisée.

On fait alors appel à des experts en graphologie ou en écriture qui examinent un ensemble de points caractéristiques de l'écriture et d'écritures de références. Ces comparaisons les conduisent à décider ou non quelle est l'origine de l'écriture.

Depuis quelques années, il est un autre domaine où l'identification du scripteur est devenue un élément important, c'est celui de la reconnaissance automatique de l'écriture [LEC 94] [CRE 96] [NOS 02]. En effet, la variabilité importante des écritures, d'un scripteur à l'autre, rend le problème de la reconnaissance particulièrement difficile. Dans le cas de la reconnaissance d'écriture manuscrite dite « off line » qui nous intéresse ici, des solutions existent dans des cas bien particuliers. On peut en noter trois principalement. Tout d'abord le cas d'une approche mono-scripteur où la variabilité, si elle est encore notablement plus grande que dans le cas omni-scripteur, est plus réduite. De même, dans le cas où l'on se limite à un vocabulaire réduit, de bons résultats ont été obtenus comme pour la lecture des montants littéraux sur les chèques par exemple. Enfin si l'on impose au scripteur des contraintes de forme des lettres à respecter on peut également obtenir des taux de reconnaissance tout à fait utilisables dans les applications. En revanche, si l'on considère le cas omni-scripteur, le problème n'a toujours pas reçu de solution satisfaisante. Une des voies de recherche consiste à limiter la phase de reconnaissance pure à des sous-ensembles définis après réduction de la variabilité des formes qui interviennent dans l'écriture. L'utilisation d'un processus identifiant le scripteur concerné, ou son style, permettrait d'effectuer de manière globale une approche multi-scripteurs tout en s'adaptant automatiquement à chaque individu ou aux caractéristiques du style reconnu pour l'écriture. Ainsi, la résolution de divers problèmes identiques mais plus simples permettrait d'apporter une solution générale.

Plusieurs études ont été menées pour apporter une solution à ce problème. De manière implicite, les styles sont très naturellement repérables par le lecteur humain qui sans lire un document, identifie facilement l'identité de son auteur si ce dernier fait partie de ceux dont il connaît l'écriture. « Connaître une écriture », c'est en quelque sorte savoir identifier son scripteur. Le style, quant à lui, recouvre une notion plus globale qui concerne un ensemble de scripteurs pour lesquels on peut constater une ressemblance plus ou moins importante des écritures. Le niveau d'observation est assez global pour que l'on ne rentre pas précisément dans les détails de l'écriture. Les styles peuvent être définis par des caractéristiques de formes locales fréquentes dans l'écriture concernée [CRE 95]. J.-C. Simon désigne par invariants ces caractéristiques de l'écriture.

Ces éléments invariants, que l'on peut observer au sein de l'écriture, sont de natures variées [KHA 96]. Ils peuvent être géométriques (boucles, lignes droites verticales, etc.) et/ou topologiques (points de croisement, points extrêmes, etc.). Durant la dernière décennie, plusieurs

études ont été menées, parmi lesquelles il convient de citer celles de J.-P. Crettez [CRE94] et de L. Heutte [HEU 00]. Ces auteurs ont extrait les propriétés spécifiques de chaque écriture en utilisant des attributs comme la pente ou le nombre de composants connexes dans le contour du texte. N. Vincent dans [VIN 94] a développé une approche plus globale montrant que les images d'écriture ont un comportement fractal. La dimension fractale apparaît comme étant un paramètre robuste, constant pour chaque scripteur et qui varie continûment avec certaines caractéristiques du style.

L'extraction de formes invariantes d'une écriture [NOS 99] constitue une avancée majeure dans le processus d'identification d'un scripteur. L'auteur établit que les invariants sont les caractéristiques propres à chaque scripteur. Cette technique s'apparente à l'expertise graphologique.

Dans notre étude, nous nous intéressons à un système global de traitement d'une chaîne d'utilisation de documents manuscrits dans un contexte général. Si nous tentons d'utiliser autant que possible les particularités de l'application pour atteindre une efficacité maximum, nous gardons néanmoins toujours à l'esprit la volonté d'assurer la généralisation de nos développements pour des problèmes voisins.

Motivation

Dans le domaine de l'analyse des écritures, on distingue différents objectifs. On peut s'intéresser à l'identification d'un individu que l'on nomme le scripteur, en s'intéressant aux échantillons d'écritures présentant de fortes similarités de contenus. Il s'agit dans ce cas de porter l'attention sur des particularités internes spécifiques d'une main.

L'identification du scripteur cherche à tirer profit de la variabilité des autres écritures avec lesquelles il faut produire la comparaison. De manière complémentaire aux tâches d'identification, on peut également s'intéresser aux approches de vérification du scripteur. Alors que l'identification d'un scripteur consiste à identifier un individu parmi un ensemble de scripteurs connus du système, la tâche de vérification consiste à déterminer si deux échantillons d'écriture sont ou non le produit d'une même main. Dans ce travail, nous nous sommes intéressé à l'analyse des écritures au sens du style (à la différence du scripteur qui est perçu comme une empreinte de la personnalité individuelle) dans un contexte de corpus d'images. Nous nous intéressons ici à l'appartenance d'une écriture à une famille présentant des propriétés morphologiques communes et qui se basent sur l'estimation de critères décrivant ce qui, dans

l'écriture, est invariant (stable et fréquent) plutôt que spécifique et rare. La notion de style d'écriture est donc centrale dans notre travail.

Le style d'une écriture est caractérisé par la présence de formes redondantes distribuées sur la surface totale de l'échantillon d'écriture, et qu'il est possible de classifier selon des critères perceptuels graphométriques spécifiques (rondeur, cursivité, linéarité…) rendant compte de leur fréquence d'apparition. Deux textes de styles différents peuvent ainsi être décrits avec le même vocabulaire de base contenant l'ensemble des occurrences des formes élémentaires (sacs de mots) mais conduisant à un rendu très différent en fonction à la fois des combinaisons locales des mots de ce vocabulaire et de leur fréquence d'apparitions sur la page.

Le style est une notion très perceptive associée à l'apparence de l'écriture dans sa globalité. Deux échantillons d'écriture de styles similaires seront donc perceptuellement proches et devront pouvoir être associés par le classifieur dans une même classe de style. Dans ce contexte, l'objectif du travail consiste à proposer des classements paramétrables des écritures selon leur style. La notion de style paléographique est une notion qui a été laissée volontairement floue par les experts paléographes associés à cette étude afin de ne pas influencer le système automatique d'extraction de descripteurs et le fonctionnement du classifieur. Il s'agit pour nous de trouver des réponses possibles à la question : « Existe-t-il une description robuste des écritures, capable de conduire à une classification des styles d'écritures paléographiques sans connaissance a priori sur les spécificités grapho-morphologiques des écritures à travers les époques ? » Cette question soulève naturellement la question de la pertinence des descripteurs, de l'élaboration de mesures de similarités consistantes et de la mise en place d'un outil de classification capable de produire des jeux de classes discriminants. Les enjeux de travail touchent à ces trois étapes fondamentales que tout classifieur met en jeu.

Parmi les applications de l'analyse des styles des écritures paléographiques, nous avons attaché une attention particulière à la recherche d'information par le contenu (RI) dans les corpus d'images de textes, à la classification des styles en groupes homogènes et à l'identification d'un style au sein d'un ensemble de styles connus a priori. Ces différents aspects applicatifs seront présentés dans le dernier chapitre de la thèse à travers l'analyse de plusieurs bases d'images de manuscrits de la période médiévale. Nous démontrerons enfin la généralisation de notre méthode de classification en styles sur des images de textes manuscrits contemporains.

Concrètement il s'agit pour nous :

- De produire une décomposition de l'écriture en graphèmes cohérents, en évitant notamment de produire des graphèmes qui correspondraient à certains gestes de rebroussement (retour en arrière du mouvement de la plume), qui sont considérés comme des mouvements incompatibles avec la nature des plumes (le plus souvent des calames) et celle des supports (nature du papier) à ces époques.
- De produire une classification de l'ensemble des graphèmes produisant un dictionnaire de formes (nommé également codebook), considéré comme un dictionnaire des graphèmes triés par similarité. Cette classification est destinée à un usage paléographique : elle intègre la possibilité pour les experts en Sciences Humaines de proposer simplement plusieurs solutions de classification des écritures par la saisie d'une seule valeur de seuil et offre la possibilité d'un rendu visuel exploitable par les experts par reprojection des étiquettes (ou couleurs) des graphèmes sur les pages d'écritures. Les similarités rendues ainsi visibles facilitent la saisie des fragments de lettres fréquents et des formes redondantes.
- De classifier les manuscrits par style d'écriture ou par main en se basant sur l'exploitation des dictionnaires de formes, dictionnaires considérés comme des signatures propres à chaque manuscrit.

L'objectif de notre travail

L'objectif de notre travail est de déterminer le sexe d'un individu d'une manière fiable à partir de son écriture manuscrite. En d'autres termes, est-ce qu'une écriture manuscrite (texte au document manuscrit) peut être attribuée spécifiquement à l'une ou l'autre des deux catégories du genre (masculin/féminin).

Organisation du mémoire

Le présent document est structuré en trois chapitres , Le première chapitre est consacré à la présentation d'introduction générale et la motivation. Dans le deuxième chapitre la présentation des principaux travaux de recherche est dans ce domaine (Un état de l'art), dans la troisième chapitre nous abordons de manière détaillée nos choix conceptuels, la mise en œuvre ainsi que les résultats obtenus par les systèmes proposés pour la détermination de sexe d'un scripteur.

Chapitre I : Introduction et motivation

Ce premier chapitre présente en bref notre démarche de travail visé qui a pour titre la motivation.

Chapitre II : Approches de classification de scripteurs: un état de l'art

Ce deuxième chapitre est dédié à l'état de l'art dans le domaine de l'analyse de l'écriture manuscrite pour la classification de scripteurs. Nous nous concentrons sur la présentation des principaux travaux de recherche dans le domaine de la reconnaissance de scripteurs. Nous distinguons entre les approches classique et les approches basées sur des caractéristiques locales, globales . Ensuite, nous discutons les diverses compétitions dans le domaine de la classification de scripteurs. et enfin nous terminons le chapitre par une comparaison des différents travaux du domaine, comme critères de comparaison, sur la taille de la base de données utilisée, les caractéristiques choisies, la taille des échantillons ainsi que le script considéré.

Chapitre III : Une méthode locale pour la détermination du sexe à partir de textes manuscrits.

Ce troisième et dernier chapitre concerne la proposition d'une approche de détermination du sexe des individus à partir d'images scannées de leurs traces écrites. Notons que la reconnaissance du sexe pourrait, éventuellement, rehausser les résultats de la reconnaissance de scripteurs. Le chapitre est dédié à la description de certains attributs des écritures manuscrites qui servent à la distinction entre les individus de sexe masculin et féminin, celle des caractéristiques proposées ainsi que les techniques de classification utilisées. Les expérimentations effectuées ainsi que les évaluations exposées sont discutées a la fin de ce chapitre.

A la fin de ce mémoire, nous émettons nos C**onclusions** sur les recherches que nous avons entreprises dans le domaine de la classification de scripteurs.

Chapitre II

Approches de classification de scripteurs: un état de l'art

Chapitre II

Approches de classification de scripteurs Un état de l'art

1. Introduction

La classification des scripteurs est un domaine de recherche attractif et très ouvert, il a attiré l'attention des chercheurs depuis quatre décennies. En revanche beaucoup de méthodes utilisées dans le domaine de la reconnaissance des formes et du traitement d'images ont été appliquées à la classification de scripteurs. En plus, les capacités grandissantes des moyens informatiques et la création de bases de données publiques de taille considérable, ont permis de donner naissance à de nouvelles méthodes de plus en plus complexes et par conséquent, les performances des techniques de classification se sont trouvées améliorées.

Jusqu'à l'an 2000, le but de la recherche dans ce domaine se concentrait sur la classification de scripteurs en mode dependant du texte et à partir de bases de données de tailles relativement petites. Ce n'est qu'au milieu des années 2000 que les chercheurs ont commencé l'utilisation de bases de données de plus en plus volumineuses et sans aucune restriction en matière de contenu textuel des échantillons. Durant la dernière décennie, les travaux dans ce domaine se sont multipliés, les types de caractéristiques proposées ainsi que les scripts considérés se sont diversifiés et les bases de données utilisées pour l'évaluation des systèmes développés se sont élargies.

Dans ce chapitre, nous allons voire quelque base de données utilisée par la comunauté dans ce contexte. Nous nous concentrons dans la deuxieme sections sur la présentation des principaux travaux de recherche dans ce domaine, en les classifions suivant les caractéristiques utilisées (les approuches classique, locales, globales ou la combinaison des deux). Ensuite, les différents compétitions relatives a ce domaine sont hexibées.

2. Ensembles de données :

Cette section décrit les ensembles de données les plus populaires pour l'identification de scripteur, qui sont librement disponibles.

2.1. La base de données CVL

La base CVL (Computer Vision Laboratory data base) est une base de données publique crée en 2013 et présenté par Kleber et Al. dans [FSM 13], elle peut être utilisée pour la recherche de scripteurs (writer retrieval), la reconnaissance de scripteurs ainsi que le repérage de mots (word--spotting). La base de données est composée 1609 textes issus de 311 scripteurs différents, 27 scripteurs ont contribués par 5 documents chacun alors que les 284 scripteurs restants ont contribué par 7 documents chacun. Pour chaque texte, une image couleur RGB (300 ppp) comprenant un texte manuscrit ainsi d'un échantillon imprimé du même texte est disponible. La base de données CVL se compose d'images avec des textes manuscrits cursifs allemands et anglais qui ont été choisis parmi des oeuvres littéraires. Cette base de données a été utilisée pour la reconnaissance de scripteurs par Fiel et Sablatnig [FIE 13].

La figure 2.1 présente deux échantillons du CVL :

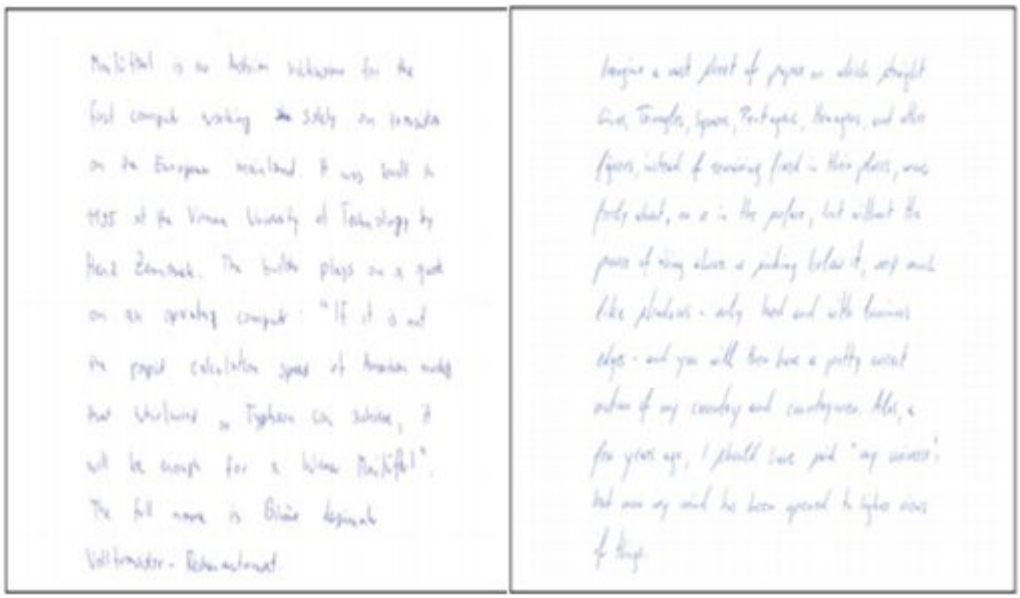


Figure 2.1 Echantillons de la base CVL.

2.2. Base de données IAM :

La base de données IAM (Institut für Informatik und angewandte Mathematik) est présenté par Marti et Bunke dans [**MAR 02**], elle est constituée de pages manuscrites correspondant à des textes anglais extraits du corpus ''Lancaster-Oslo/Bergen'' (LOB). Le corpus est une collection de textes qui se compose d'environ un million d'instances de mots [JOH 78]. Elle comprenait, dans sa première version, 556 images de textes produits par environ

250 scripteurs différents [MAR 99]. Ensuite, elle a été étendue pour contenir 1539 images de textes produites par 657 scripteurs différents. Au vu de sa disponibilité au publique, sa structure flexible, et le grand nombre de scripteurs qu'elle contient, la base de données IAM a été couramment utilisée pour la reconnaissance de scripteurs latins [SID 10, BER 13] ainsi que pour la reconnaissance de l'écriture manuscrite [SCL 04b].

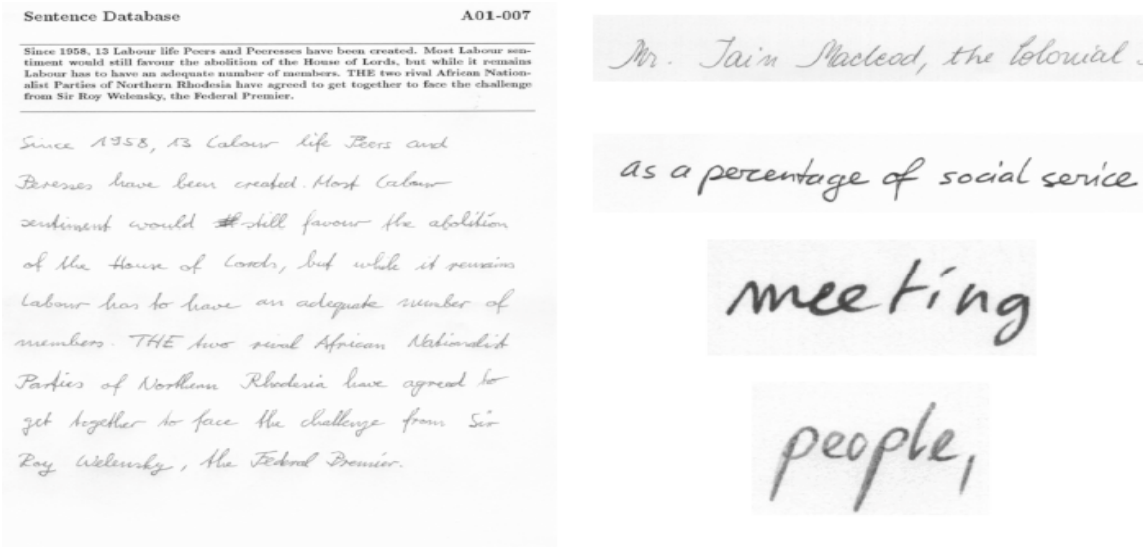


Figure 2.2 Texte de la base de données IAM

2.3. Base de données CEDAR:

La base de données CEDAR a été développée à l'Université de Buffalo (l'université d'État de New Yor) [SRI 02], elle est considérée comme l'une des premières grandes bases de données développées pour la classification des écritures manuscrites latines et plus particulièrement, pour la reconnaissance de scripteurs. Elle est composée de 4701 images de textes manuscrits écrits par 1567 scripteurs différents qui ont été sélectionnés pour être représentatifs de la population des Etats--Unis d'Amérique. Chaque scripteur a recopié trois exemplaires de la lettre CEDAR [SRI 02]. Cette lettre est un document qui contient 156 mots, à partir d'un lexique de 124, qui inclut tous les caractères (lettres et chiffres).

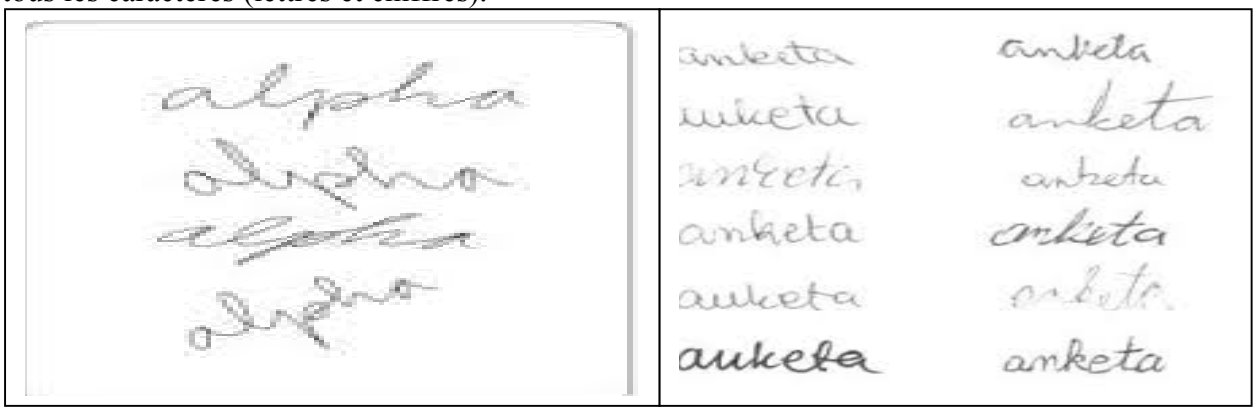


Figure 2.3 Echantillons de la base CEDAR.

2.4. Base de données KHATT :

KHATT [MAH 12] est une nouvelle base de données qui contient des images de textes arabes manuscrits, elle peut être utilisée pour la reconnaissance de scripteurs, la segmentation des textes en lignes ainsi que la reconnaissance de textes manuscrits. La base de données KHATT contient 4000 images de paragraphes en niveaux de gris, ces images contiennent des textes scannées à différentes résolutions (200, 300 et 600 ppp). 1000 Scripteurs de différents âges et origines et provenant de 18 pays différents ont participé à la collecte de cette base de données. Sur les 1000 scripteurs, 677 étaient de sexe masculin tandis que les 323 restants étaient de sexe féminin. 928 Scripteurs étaient des droitiers, tandis que 72 étaient des gauchers. 2000 Images sur les 4000 de la base contiennent un texte similaire couvrant tous les caractères et chiffres arabes alors que les 2000 images restantes contiennent des textes libres écrits par les scripteurs sur un sujet de leur choix.

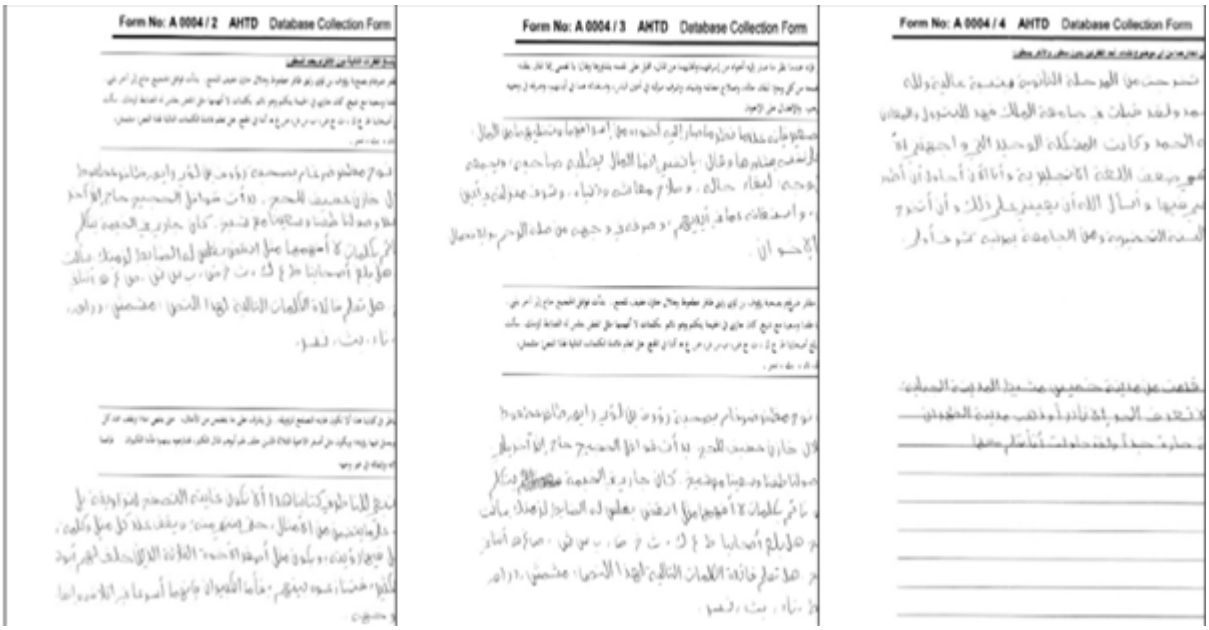


Figure 2.4 Echantillons de la base KHATT.

2.5. Base de données QUWI :

QUWI [**ALM 12**] est une nouvelle base de documents manuscrits hors-ligne créée par une équipe de recherche à l'Université du Qatar. Cette base de données contient des documents écrits en arabe et en anglais, elle peut être utilisée pour évaluer les performances des systems de reconnaissance de scripteurs ainsi que de ceux de reconnaissance du genre. Elle se compose de documents manuscrits de 1017 bénévoles de différents âges, nationalités, sexes et niveaux d'éducation. Les scripteurs ont été invités à copier un texte spécifique et à générer un texte

aléatoire ce qui permet à la base de données d'être utilisée aussi bien en mode dépendant du texte qu'en mode indépendant du texte. Il est à mentionner qu'une partie de cette base de données (475 scripteur) a fait l'objet d'etude sur la prédiction du genre à partir de documents manuscrits qui s'est déroulée dans le cadre de la conférence ICDAR 2013 [**HAS 13**].

2.6. Base de données RIMES:

La base de données RIMES comprend 5600 courriers manuscrits écrits en langue française tels que ceux envoyés par des particuliers à des entreprises ou administrations. Chaque courrier contenant 2 à 3 pages, la base de données représente 12600 pages originales (au format A4). 1300 Scripteurs bénévoles ont participé à la constitution de la base de données RIMES en rédigeant les lettres avec leur propre formulation. Cette base de données peut donc être considérée comme réaliste, car la formulation du contenu des documents était libre. Il est à noter que la base de données RIMES a été utilisée pour la reconnaissance de scripteurs par Siddiqi et Vincent [SID 10] ainsi que pour la reconnaissance de l'écriture française manuscrite lors des compétitions qui se sont déroulées dans le cadre des conférences ICDAR 2009 [GRO 09] et ICDAR 2011 [GRO 11].

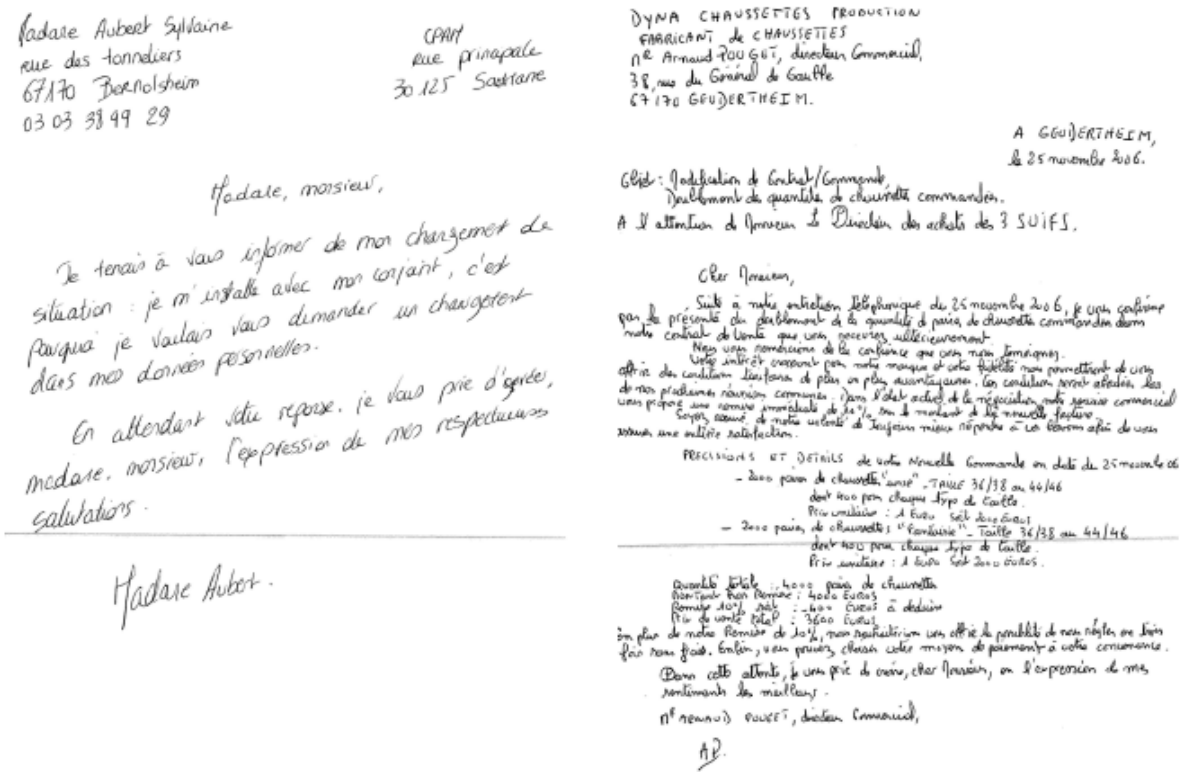


Figure 2.5 Echantillons de la base RIMES.

3.État de l'art sur les méthodes de classification de scripteur

L'identification du scripteur est un domaine de recherche très actif durant ces deux dernières décennies. Une grande variété de systèmes est basée sur l'utilisation du traitement d'image par ordinateur et des techniques de reconnaissance de forme ont été proposées pour résoudre les problèmes rencontrés dans l'analyse automatique de l'écriture, nous nous concentrons sur les principaux travaux du domaine, effectués ces dernières années. Afin de mieux décrire les différentes méthodes proposées pour la classification de scripteurs, nous avons rassemblé les méthodes existantes en trois sections principales.

3.1. Approches classiques

Ces approches sont basées sur les techniques de traitement d'images et de reconnaissance de formes. Les solutions apportées permettent de résoudre les problèmes généraux rencontrés usuellement dans les différentes étapes du traitement d'images. Elles concernent les prétraitements la sélection de caractéristiques, la comparaison des écritures et des scripteurs ou l'évaluation de la performance globale. Pour cela, sont distinguées les phases d'apprentissage et d'identification. En ce qui concerne l'apprentissage, on suppose que le texte considéré comme exemple de l'écriture est assez long pour contenir toutes les informations utiles sur le scripteur et on s'accorde généralement pour considérer qu'une longueur de texte est suffisante lorsque le texte comporte entre 3 et 5 lignes [SER 03].

Nous distinguerons deux types de méthodes classiques, les approches contextuelles et les approches non contextuelles.

3.1.1 Approches contextuelles

Les approches contextuelles ne considèrent pas seulement l'image du texte comme une forme mais se servent aussi du contenu sémantique du texte. Ces approches s'appuient donc sur une localisation interactive et une segmentation manuelle du texte en mots et lettres dont la connaissance fait partie intégrante des données. L'enregistrement des scripteurs peut être réalisé dans une phase initiale où un protocole approprié est défini. On peut par exemple demander aux scripteurs d'écrire un texte fixé au préalable en écrivant chaque lettre dans une boîte. L'association entre l'image et le contenu sémantique peut alors être réalisée de manière automatique.

Mihelic et al. [MIH77] utilisent en 1977 la transformée de Walsh-Hadamard appliquée aux lettres, aux mots ou au texte global. Ils ne gardent que les coefficients qui sont les plus

porteurs d'information comme caractéristiques du scripteur. Le classifieur retenu repose sur le principe du maximum de vraisemblance.

Naske [NAS 82] considère la forme particulière de chaque écriture comme une distorsion spécifique au scripteur à partir d'un prototype, *a priori* celui qui est appris durant la scolarité de tout écolier. Le prototype de chaque lettre est choisi commun à tous ceux qui utilisent le même alphabet. Ainsi des caractéristiques sont calculées en prenant en compte les différences existantes entre le caractère écrit et le prototype. Deux méthodes sont appliquées, l'une consiste à approximer le caractère écrit à partir d'un prototype par une matrice de déplacement et l'autre procède à partir de l'image du caractère qui est déformée de manière à retrouver au mieux un des prototypes. Une description non-paramétrique d'une fonction de transfert est recherchée. Pour chacune de ces deux méthodes, 50 caractéristiques sont extraites.

L'ensemble des caractéristiques est ensuite réduit grâce au critère de Fisher [KLE 80] après avoir décorrélé les caractéristiques par une transformée de blanchiment. Ainsi sur les 50 caractéristiques extraites, les 20 plus significatives sont retenues. La méthode d'identification repose sur le critère de Bayes avec un classifieur de calcul des distances minimums entre d'une part le prototype transformé et le caractère à reconnaître pour la première méthode et d'autre part entre le caractère transformé et les prototypes. Les tests sont effectués sur le mot allemand "DREIHUNDERT" (trois cent) écrit 10 fois en caractères bâtons par 100 scripteurs. L'auteur obtient plus de 98% d'identification sur une base de 9 caractères. Ces méthodes sont assez contraignantes puisqu'elles ne peuvent pas être utilisées si on considère un seul caractère en majuscule, c'est pour cela que les scripteurs doivent écrire le mot sous forme de plusieurs combinaisons (DRE, DRE.HUND ou DRE.HUNDER).

3.1.2. Approches non contextuelles

Elles sont beaucoup plus nombreuses que les approches contextuelles et permettent de laisser aux scripteurs une plus grande liberté dans leur comportement. Nous présentons ici des systèmes d'identification de scripteur avec une approche non contextuelle selon l'ordre chronologique.

Benjamin arazi montre dans [ARA 77] la similarité existante entre les histogrammes vertical et horizontal de deux échantillons d'écriture d'une même personne. Après avoir scanné en noir et blanc un texte de 300 lignes de 13 scripteurs différents, de façon horizontale (i.e. parallèle aux lignes écrites) et verticale, l'auteur enregistre les histogrammes correspondants. De ces deux histogrammes enregistrés pour chaque scripteur, il en ressort plusieurs mesures. Sur

l'histogramme horizontal, la localisation, la largeur et la valeur du pic d'une part, le calcul de similarité entre deux textes identiques écrits par deux scripteurs différents d'autre part. La similarité est la valeur absolue de la différence calculée entre le nombre de *runs* dans les deux histogrammes et la somme de ces valeurs pour toutes les longueurs de *runs* possibles.

L'histogramme vertical possède deux pics dont le premier informe sur la largeur des lettres et le deuxième correspond à l'espacement entre deux lignes. Les mesures précédentes appliquées à l'histogramme vertical ne donnant pas assez de résultats concluants, une seule mesure est retenue. Cette mesure est celle de la similarité, précédemment décrite, elle est obtenue comme la somme de toutes les valeurs élevées au carré. Cela permet d'accentuer les écarts entre scripteurs.

Klement décrit dans [KLE 81] un système complet d'identification de scripteur et dans [KLE 83] comment ce système est intégré au réseau informatique. C'est un regroupement de plusieurs modules réalisés avec d'autres chercheurs dans le cadre d'un projet de recherche appelé « FISH » (Système de l'Information Experte de l'Ecriture) financé par le Ministère de la Recherche Allemand. Ce système est composé d'une partie utilisant des techniques d'analyse d'image classiques telles que le critère des moindres carrés pondérés [NAS 80] avec extraction de caractéristiques non-contextuelles, et une autre partie utilisant des caractéristiques relatives à un seul caractère [NAS 82]. La première étude a été faite par Kuckuck et *al*.

Kuckuck et al. décrivent dans [KUC 79] un processus d'identification de scripteur basé sur la méthode d'évaluation « hold-one-out ». Cette méthode consiste à retirer un échantillon de la base d'apprentissage, à entraîner le système, et à tester le système avec l'échantillon retiré. La méthode est utilisée lorsque l'on ne dispose pas d'un ensemble d'apprentissage de taille suffisante. Une étape d'extraction de caractéristiques est d'abord appliquée sur les écritures binaires. Elle comporte trois traitements distincts appliqués en parallèle. Le premier est une description analytique de l'écriture La seconde méthode pour extraire des caractéristiques repose sur le calcul de la distribution fréquentielle sur 6 directions de toute l'image.

Et la troisième méthode considère à nouveau la distribution fréquentielle mais elle est ici calculée sur 8 directions. Après application de ces 3 méthodes, 128 caractéristiques sont extraites. Les auteurs appliquent alors une étape de compression en approximant les distributions par les moindres carrés pour obtenir au final 24 caractéristiques par scripteur.

Deux classifieurs sont mis en place, l'un a recours à la méthode des plus proches voisins quant à l'autre, c'est le critère de Bayes en faisant l'hypothèse d'une distribution gaussienne des caractéristiques. Les auteurs obtiennent 96% d'identification sur une base de 20 scripteurs.

Une nouvelle méthode d'extraction de caractéristiques a été apportée à ce processus par W. Kuckuck [KUC 80]. Il discute de quelques méthodes basées sur la transformée de Fourier ou sur la fonction d'autocorrélation, qui sont apparentées à l'analyse spectrale. Des caractéristiques sont extraites par calcul de coefficients dérivés du spectre de puissance de la transformée de Fourier. Il applique par ailleurs la fonction d'autocorrélation sur ce spectre pour extraire d'autres caractéristiques. L'auteur obtient aux alentours de 90% d'identification avec ces deux nouvelles méthodes et plus de 95% avec la fonction d'autocorrélation qui demande moins de temps de calcul.

Zois et Anastossopoulos [ZOI 00] identifient des scripteurs en utilisant une transformation morphologique pour extraire un vecteur de caractéristiques sur un mot. Une projection horizontale est effectuée sur le mot squelettisé. Cette projection est alors segmentée à l'aide d'une opération morphologique, la granulométrie. Deux types de fenêtres sont appliqués sur les segments de la projection pour contrôler le flot d'information. Les espaces entre les lettres sont aussi pris en compte dans la formation du vecteur de caractéristiques. L'étape d'identification dépend de la dimension du vecteur qui dépend elle-même de la longueur du mot. La base de données a été constituée de 50 scripteurs qui ont écrit deux mots de même longueur en Anglais et en Grec. Les classifieurs utilisés sont celui de Bayes et un réseau de neurones. Les auteurs obtiennent 95% d'identification.

Marti et al. [MAR 01] extraient des caractéristiques par ligne souvent visibles par le lecteur et les classent selon deux classifieurs tels que les réseaux de neurones et l'algorithme des plus proches voisins pour identifier les scripteurs. Leur processus de découpe en trois étapes : prétraitements extraction de caractéristiques et classification. L'étape de prétraitement consiste à segmenter une page de texte par ligne à l'aide d'une projection puis à binariser ces lignes. Ensuite, pour chaque ligne, un ensemble de caractéristiques est extrait. Cet ensemble se compose de 12 caractéristiques : 6 correspondent aux zones du texte telles que la partie supérieure et inférieure du corps du texte, 2 correspondent à la taille de l'écriture (la distance entre deux mots), les 2 suivantes prennent en compte la pente de l'écriture et les 2 dernières calculent les pentes extraites d'un graphe tracé par le calcul de la dimension fractale. Les tests sont effectués sur 20 scripteurs avec pour chacun, 482 lignes de textes pour l'apprentissage et 182 lignes pour les tests. Ils obtiennent 87,8% d'identification lorsqu'ils utilisent 7 des 12 caractéristiques sur

l'algorithme des plus proches voisins et 90,7% en utilisant toutes les caractéristiques sur un réseau de neurones de niveau 20.

Ces méthodes classiques recherchent des caractéristiques qui permettraient de discriminer de manière objective les écritures, ou plutôt les scripteurs. Partant de la constatation que nous savons reconnaître l'écriture des gens qui nous sont proches sans l'analyser, mais en la regardant globalement sans chercher à lire le contenu du texte considéré, différentes approches ont cherché à identifier les scripteurs par leur style d'écriture. Le style peut aussi être considéré comme une caractéristique commune à une famille de scripteurs dont les écritures se ressemblent.

2.2. Méthodes locales

Dans cette section, nous présentons les différents travaux proposés dans le domaine de la classification de scripteurs utilisant des caractéristiques locales. Nous avons essayé de détailler les travaux que nous avons jugés les plus intéressants, les autres sont brièvement discutés.

Dans ces méthodes, l'accent est mis généralement sur les caractères individuels d'allogreffes et la façon dont une personne particulière serait les dessiner. Au lieu de prendre tout l'ensemble de l'écriture, Dans notre discussion, nous allons d'abord présenter brièvement l'identification actuelle d'un scripteur à partir d'un faible niveau de caractéristiques suivi par la représentation d'une classification des styles d'écriture à l'aide de la modélisation d'allogreffe.

2.2.1. Méthodes basées sur les caractéristiques marphologique

Dans [MAR 01] l'auteur extrait un ensemble de 12 caractéristiques (qui correspondent principalement à des caractéristiques visibles de l'écriture) à partir de lignes manuscrites de texte qui sont ensuite classés en utilisant le k-plus proche voisin pour la reconnaissance du scripteur. En utilisant une projection en profils, l'image de texte manuscrit est d'abord segmentée en lignes qui sont ensuite binarisés. Pour chaque ligne, un ensemble de caractéristiques est extrait comprenant principalement la hauteur des trois principales zones d'écriture -ainsi que leurs rapports-, la largeur des caractères et les distances inter-mots. En outre, les caractéristiques en fonction du comportement fractal de l'écriture, qui sont corrélés avec l'écriture de la lisibilité, sont également utilisées.

Le système a été évalué en 100 pages de 20 scripteurs différents avec un total de 912 lignes de textes divisé en cinq sous-ensembles disjoints (4 utilisés dans la formation et 1 dans les tests). Une identification moyenne d'un taux de 87,8% est réalisée en utilisant 7 des 12 caractéristiques

avec le plus proche-k classement voisin et 90,7% est mesurée par une classification sur les réseaux de neurone en utilisant l'ensemble complet de caractéristiques.

Les taux d'identification du scripteur sont obtenues avec ces caractéristiques sont prometteurs mais ils sont basés sur un protocole d'évaluation où quatre exemples d'images par scripteur ont été utilisés pour le test.

2.2.2. Méthodes basée sur le codebook:

Le concept des invariants d'un scripteur introduit dans [NOS 99] a été employé par **Bensefia et al**. décrit dans [BEN 02], qui ont proposé un système d'identification de scripteur basé sur un modèle de correspondance de graphèmes extraits des textes à comparer. La redondance morphologique de l'individu, définit comme étant les invariants de l'écrivain (Codebook), est un ensemble de motifs ou graphèmes similaires extrait de la segmentation du texte manuscrit [BEN 04]. Ceci permet la compression du texte manuscrit tout en maintenant un bon taux d'identification. Les composantes connexes du document sont d'abord extraites puis segmentés en graphèmes qui sont en réalité des motifs élémentaires de l'écriture manuscrite et qui sont produits par une segmentation basée sur l'analyse des minima sur le contour supérieur [NOS 99] comme il est indiqué sur la figure 2.6. Dans des études ultérieures, les auteurs ont également introduit la concept de bi-grammes (tri-grammes) obtenus à la suite de la concaténation de deux (trois) graphèmes voisines (graphèmes i et i + 1) [NOS 02].

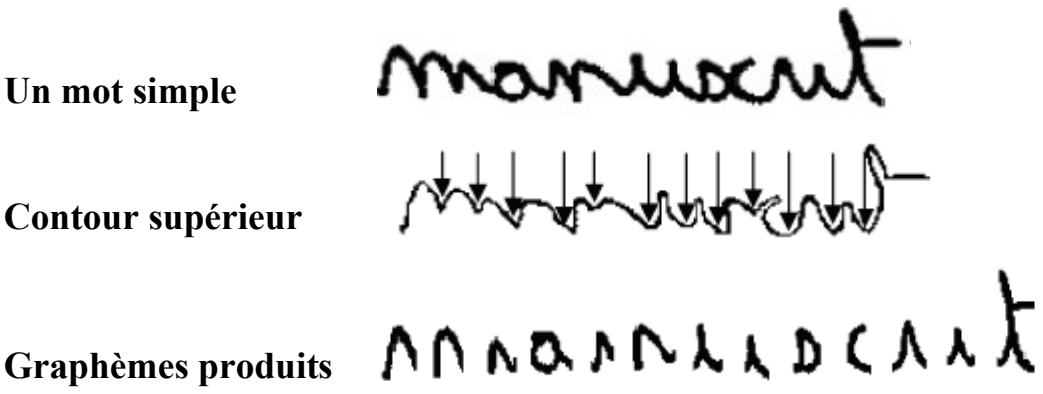


Figure 2.6. Segmentation en graphèmes (Image: [BEN 05])

Une fois que les graphèmes ont été segmentés, ceux qui sont morphologiquement similaires seront regroupés en utilisant un algorithme de clustering séquentielle [FRI 99] où deux graphèmes sont comparés à l'aide d'une mesure de corrélation de similarité. Les auteurs proposent également d'effectuer plusieurs regroupements séquentiels avec une sélection aléatoire

des éléments moins sensible à l'ordre dans lequel ils sont présentés. Enfin, seuls les éléments qui sont toujours regroupés, lors de l'itération les procédures de regroupement, sont retenus pour constituer un cluster appelé cluster invariant (Figure 2.7).

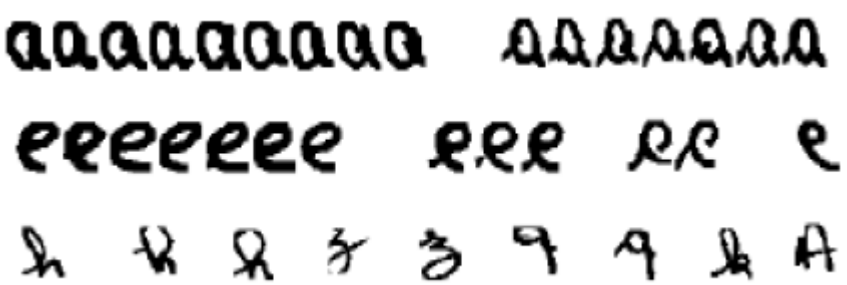


Figure 2.7: Des échantillons de clusters invariants extraits d'une page manuscrite [BEN 02]

Jain et al. [JAI 11] adaptent les descripteurs k-AS (k-Adjacent Segments) introduit initialement par Ferrari et al [FER 08] afin de modéliser les écritures manuscrites de différents individus. Les auteurs procèdent d'abord au traitement de l'image binaire du document manuscrit par le détecteur de Canny afin de détecter les contours qui caractérisent mieux les formes des caractères ainsi que les courbures. Un algorithme d'ajustement de ligne est ensuite utilisé pour réduire les contours obtenus en un ensemble de segments (voir figure 2.8). Puis les segments ayant leurs extrémités proches (segments adjacents) sont reliés ensemble. Le descripteur k-AS est alors calculé à partir de chaque sousensemble connexe de 2, 3 et 4 points successivement.

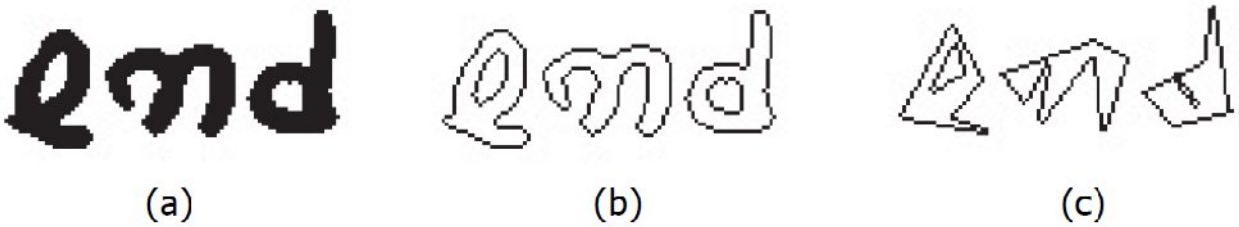


Figure 2.8. Illustration de l'extraction de segments à partir de l'image du mot "end" (a) : Image originale, (b) : Contour de l'image et (c) : Segments extraits à partir des contours.

Les auteurs utilisent la technique de propagation affine [HE 08] pour le regroupement des k--AS obtenus à partir de la base d'apprentissage afin de construire un codebook. Une fois le

codebook obtenu, il est utilisé pour la caractérisation des documents manuscrits des différents scripteurs. La figure 2.9 illustre un codebook généré à partir d'un ensemble de 3-AS.

Figure 2.9 Codebook de 3-AS.

Les auteurs ont évalué leur méthode sur deux bases de données différentes : la base IAM contenant des documents écrits en anglais et issus de 650 scripteurs différents, ainsi que la base MADCAT [STR 09] comprenant des manuscrits arabes parvenant de 302 scripteurs différents. La méthode permet d'atteindre des taux d'identification de l'ordre de 93% et de 90% sur les bases IAM et MADCAT, respectivement. Les résultats obtenus montrent que les performances de la méthode augmentent à mesure que le nombre d'échantillons d'apprentissage augmente et que le codebook est générique, indépendant des langues et des scripteurs de sorte qu'il n'ait pas besoin d'être recréé en fonction du script considéré (arabe ou latin).

Siddiqi et al. [SID 10] ont présenté une méthode pour la reconnaissance de scripteurs combinant des caractéristiques allographiques et structurelles. Les caractéristiques allographiques sont basées sur de petits fragments d'écriture, ces fragments sont extraits par un découpage adaptatif de l'écriture en imagettes de taille pixels. Ces imagettes sont ensuite regroupées en utilisant l'algorithme de regroupement k-means (la valeur de k est fixée à 100) pour avoir un codebook (voir figure 2.10).

Les caractéristiques allographiques atteignent des taux d'identification de l'ordre de 84% pour la base IAM (qui contient 650 scripteurs) et 74% pour la base RIMES (qui comporte 375 scripteurs). Les erreurs de vérification enregistrées sont de l'ordre de 4.49% et de 10.57% pour les bases IAM et RIME successivement.

En ce qui concerne les caractéristiques structurelles, elles sont basées sur les contours qui encapsulent le style d'écriture de l'auteur et permettent de préserver des variations (qui dépendent du scripteur) entre les formes de caractères.

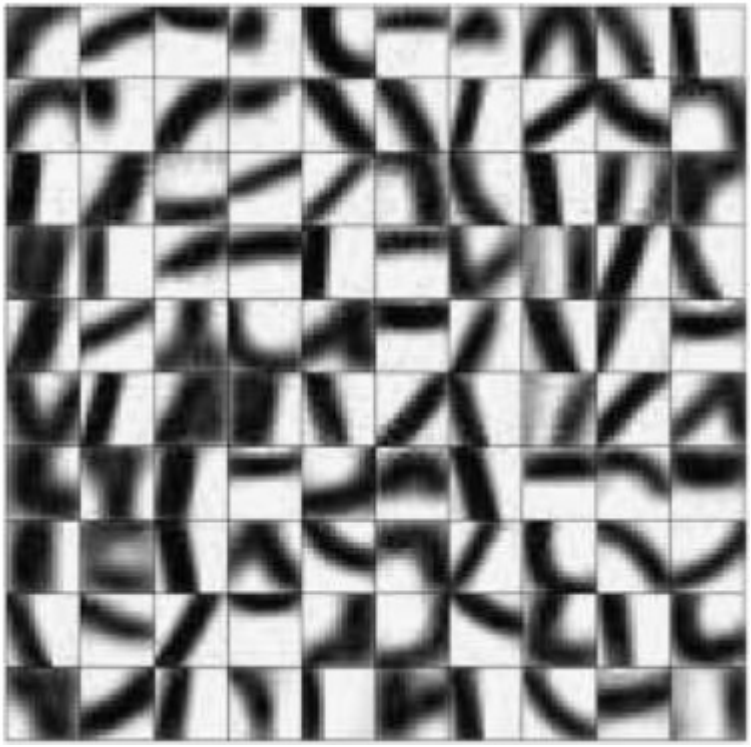


Figure 2.10 Un codebook universel de taille 100 obtenu à partir d'échantillons de la base RIMES [SID 10].

Deux représentations différentes des contours correspondant à deux échelles d'observation et à deux niveaux de détails différents sont envisagées : la première représentation est basée sur les chaînes de Freeman, alors que la deuxième représentation est basée sur un ensemble de polygones approximant les contours. Par l'utilisation de ces deux représentations, un ensemble de 14 caractéristiques est défini. Ces caractéristiques sont les distributions des codes de Freeman, les distributions des différences du 1er et 2ème ordre de codes de Freeman, les distributions de paires et de triplets de codes de Freeman, les distributions des indices de courbure, les distributions des directions de trait, les distributions des pentes de segments, les distributions des courbures ainsi que les distributions des longueurs de segments.

Lorsque les deux types de caractéristiques structurelles sont combinés, des taux d'identification de l'ordre de 89% et de 85% ont été enregistrés sur les bases IAM et RIMES respectivement. Dans ce cas, les erreurs de vérification atteignent 2.46% pour la base IAM et 4.87% pour la base RIMES.

Djeddi et al. [DJE 10] ont Proposé une approche locale en mode dépendant du texte où on cherche les formes invariantes et propres à l'écriture de chaque scripteur. Ces formes sont extraites par un découpage de l'écriture suivi d'une classification des formes obtenues qui sont ensuite organisées dans une base de référence pour permettre d'identifier l'auteur d'un document inconnu dans un processus de Pattern Matching. Les résultats obtenus sont de l'ordre de 93.33% en Top 1 et de 100% en Top 2 sur une base de 30 scripteurs.

3.3. Méthodes globales

Ces méthodes identifient le scripteur d'un document basé sur l'aspect et la convivialité de l'écriture.

3.3.1. Analyse de Texture des textes manuscrits

[SAI 00] présente un algorithme pour l'identification automatique d'un scripteur indépendant du texte manuscrit en considérant l'écriture de chaque individu comme une texture différente. Le système comprend trois principales étapes: la normalisation, l'extraction de caractéristiques et l'identification du scripteur. Au cours de la phase normalisation, la détection et la correction des mots détectées dans les images représentant les textes manuscrits sont réalisés en utilisant le montage sur ligne des composants connectés. Ensuite, l'espace entre les lignes / mots et les marges sont réglées sur une taille prédéfinie pour produire un motif bien défini utilisé pour l'analyse de texture. Les caractéristiques de la texture sont obtenues par deux procédés principaux à savoir le filtrage à canaux multiples de Gabor [PEA 97] et la matrice de gris en co-occurrence (GLCM) [TAN 96], implémentés sur des blocs aléatoires non chevauchés (de 128x128 pixels) extraits de l'image normalisée.
Les deux filtres de Gabor sont de symétries opposées. Ils sont donnés par:

$$h_e(x,y;f,\theta) = g(x,y)\cos\big(2\pi f(x\cos\theta + y\sin\theta)\big)$$

$$h_o(x,y;f,\theta) = g(x,y)\sin\big(2\pi f(x\cos\theta + y\sin\theta)\big)$$

Où g est une fonction gaussienne 2-D, f et sont la fréquence et l'orientation radiale qui définissent l'emplacement du canal dans le plan de fréquence. Quatre fréquences de 4, 8, 16 et 32 cycles / degré ont été utilisés et pour chacune de ces fréquences, le filtrage est réalisée pour 4 orientations (0°, 45°, 90° et 135°). Ceci a donné un total de 16 images de sortie (4 pour chaque

fréquence), à partir de laquelle les caractéristiques de l'écrivain sont extraites. Ces caractéristiques sont la moyenne et l'écart type de chaque image de sortie. Par conséquent, 32 caractéristiques par image d'entrée sont calculées.

La matrice de covariance est une matrice carrée M de taille N (nombre de niveaux de gris) où chaque élément M (i, j) de la matrice représente le nombre de paires de pixels séparés par une distance d pour un angle , ayant les valeurs de gris i et j respectivement. Les auteurs ont examiné 5 distances (1, 2, 3, 4 et 5) et 4 directions (0°, 45°, 90° et 135°) sur les images d'écriture binaires donnant un total de 20 matrices. Pour chaque matrice 2 x 2 il y a seulement 3 valeurs indépendantes en raison de la symétrie diagonale. Elles sont utilisées directement en tant que caractéristiques donnant ainsi un total de 60 (20 x 3) caractéristiques pour l'image de l'écriture manuscrite.

3.3.2. Analyse fractale de l'écriture manuscrite

La dimension fractale telle que définie par B. Mandelbrot [MAN 75], est un "nombre qui mesure le degré d'irrégularité ou de la fragmentation d'un ensemble », ou la mesure de la complexité de l'ensemble étudié. Le comportement fractal des écritures a été prouvé par N. Vincent [VIN 95]. Des études ultérieures montrent que sous certaines conditions d'observation, les paramètres fractales sont stables et une discrimination suffisante pour établir une classification de l'écriture manuscrite selon la [BOU 98]. Le calcul de la dimension fractale appliquée était basée sur la mesure de la dimension de Minkowski-Bouligand qui est donné pour un ensemble X comme :

$$D(X) = \lim_{r \to 0}[2 - \frac{\log(A(X_r))}{\log(r)}]$$

Où A(Xr) est la surface de la couverture optimale de X par des sphères de rayon r. Pour une courbe fractale, le comportement de log [A (Xr) / r] en fonction de log (r) est linéaire et le graphe correspondant est appelé comme graphe d'évolution. Il est déterminée par le calcul de la superficie de la surface couvrante de X par des sphères de rayon r qui est mis en œuvre par dilatation de X (comme indiqué sur la figure 2.11), r fois par une sphère de rayon 1.

- Trois zones ayant des pentes différentes peuvent être identifiée dans le graphe; chaque une caractérise un comportement particulier et correspond à une échelle différente de l'observation.

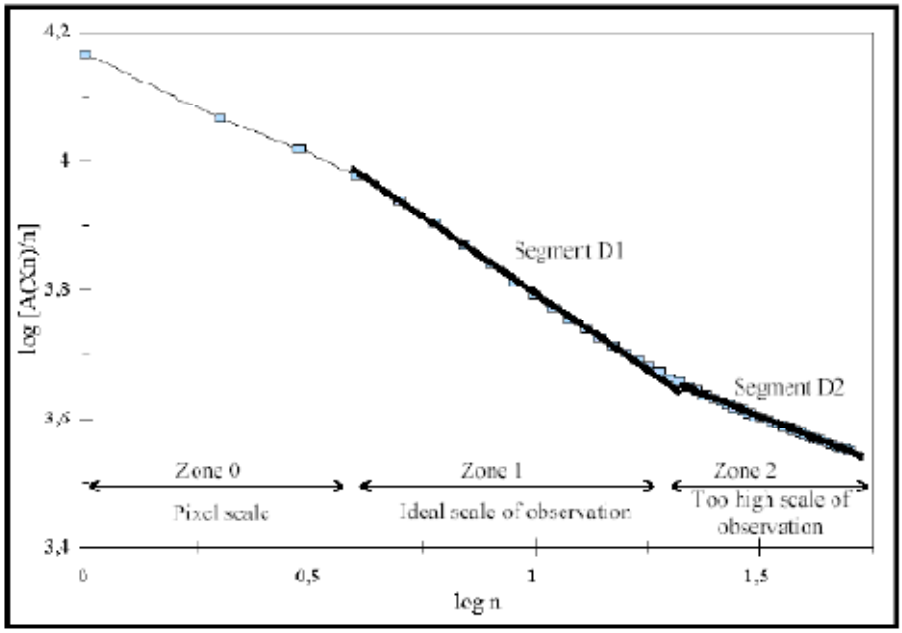


Figure 2.11 : Graphe d'évolution [BOU 98].

Les auteurs définissent alors deux paramètres : la dimension fractale et la dimension secondaire de l'écriture manuscrite. La dimension fractale (FD) est calculée à partir de la pente de la zone 1, tandis que la dimension secondaire (D2) est calculée à partir de la pente de la zone 2. Ces deux paramètres correspondent respectivement à une perception de l'écriture d'une adaptation et une distance éloignée de l'observation.

Les auteurs ont testé la stabilité de la dimension fractale sur des contraintes physiques liées à la fois à l'écriture et l'acquisition [BOU 95]. Ils concluent que l'utilisation différent instruments d''écriture ont une influence uniquement sur les premiers points du graphe d'évolution (zone 0) qui ne sont pas pris en considération pour le calcul de la FD. Il a également été démontré [BOU 97] que les changements de résolution produisent une un changement des différentes zones du graphe sans affecter les pentes de la zone 1 et 2 utilisés dans le calcul des paramètres proposés. Représentant les écritures dans le plan, FD vs D2, les auteurs suggèrent que la distribution est liée à la lisibilité nommé graphe de lisibilité et (figure 2.23).

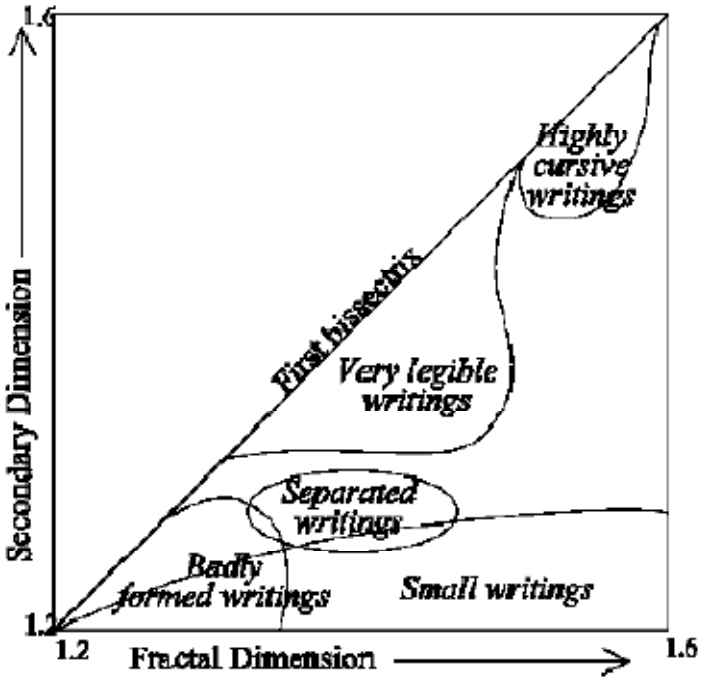


Figure 2.12 : Graphe de lisibilité [BOU 98].

Le graphe de la lisibilité basé sur les dimensions fractales représente une écriture dans un espace à deux dimensions qui est assez bon pour classer les styles d'écriture, mais pourrait ne pas être suffisante pour des problèmes comme l'identification du scripteur, où l'attribution d'une écriture à une classe particulière (Scripteur) nécessite plus de précision.

3.3.3. Méthode basée sur la loi de Zipf

Dans une tentative d'indexer et identifier les manuscrits, [PAR 06] présente une méthode basée sur la loi de Zipf [ZIP 49] qui modélise l'occurrence d'objets distincts dans des collections triées et utilisé à l'origine dans les domaines mono-dimensionnelle. La loi stipule que quand un ensemble donné de symboles (modèles) sont triées en respectant la diminution de la fréquence d'occurrence La relation suivante peut être observé :

$$N_{\sigma(i)} = k \times i^{a}$$

Où $N_{\sigma(i)}$ représente le nombre d'occurrences du symbole de rang i, et k et sont des constantes.

La puissance de cette loi est caractérisée par la valeur de l'exposant **a** tant que k est plus liée à la longueur de la séquence de symboles étudiés. Cette relation n'est pas linéaire, mais une simple transformation peut la conduire conduit à une forme linéaire relation entre le logarithme de N et le logarithme du rang.

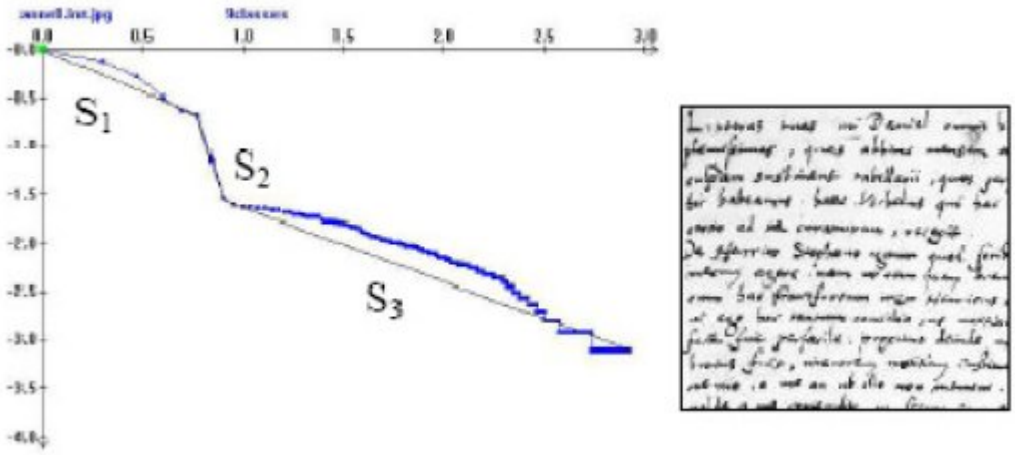


Figure 2.13 : Exemple d'un manuscrit et de son représentation Zipf [PAR 06]

L'application de cette loi aux images nécessite un certain type de codage. Les auteurs ont choisi de quantifier les valeurs de gris à k-niveaux k, k étant fixé à 9, puis à 3. Un masque 3x3 (pour k = 9) et un masque de connectivité de 4 (k = 3) créant respectivement 99 et 35 motifs possibles. Bien que loi de Zipf ne tient pas pour l'ensemble de l'image, les courbes Zipf sont construits et estimés par certains segments de ligne droite. Les auteurs ont choisi de considérer dans chaque courbe jusqu'à trois zones linéaires différents (figure 2.13). Le point fractionnement sur un segment de courbe est défini comme le point le plus éloigné de la ligne droite reliant les deux points extrêmes de la courbe à diviser. Les trois pistes et abscisses des extrémités de chaque segment sont considérées comme caractéristiques du document de représentation et deux images sont comparées par la distance de Hamming dans l'espace de fonction. La méthode est testée sur une collection de documents du 16ème siècle par 20 auteurs différents rapportant un taux d'identification du scripteur jusqu'à 80%.

3.3.4. Méthodes d'identification par styles d'écriture:

L'objectif de ces méthodes est d'extraire des caractéristiques propres au scripteur indépendamment du contenu sémantique du texte analysé. Ces caractéristiques permettraient aussi d'adapter la reconnaissance d'écriture au scripteur par le fait que les écritures de même style présentent une variabilité moindre que l'ensemble de toutes les écritures possibles. Il existe

deux axes : la modélisation d'allographes qui permet d'absorber la variabilité de l'écriture et une méthode globale par styles qui permet de regrouper les scripteurs par familles.

➢ Modélisation d'allographes

Les modélisations d'allographes consistent à représenter les lettres sous forme d'un ensemble de pseudo-lettre. Elles permettent de s'affranchir de la trop grande variabilité de l'écriture.

Plusieurs méthodes ont été utilisées. M. Gilloux décrit dans [GIL 94] une méthode améliorant le taux de reconnaissance d'écriture en incluant le style d'écriture. Pour ce faire, il propose une méthode d'adaptation au scripteur basée sur plusieurs modèles d'écritures estimés par les Modèles de Markov Cachés (MMC).

Chaque modèle est entraîné pour prendre en compte un style en particulier qui est connu au préalable. Les modèles de Markov supposent que l'image peut être représentée comme une séquence d'observations. L'auteur utilise 3 façons de classer les styles pour s'assurer que les observations sont indépendantes mutuellement quand les états cachés des MMC sont connus. La première étape est la normalisation de l'écriture ce qui permet de garantir une meilleure généralisation du modèle markovien. Les autres aspects du style de l'écriture sont pris en compte dans le MMC par l'utilisation de plusieurs sous modèles.

La détection du style se fait pendant l'étape d'extraction de caractéristiques. Cette étape consiste à segmenter les composantes connexes des mots en contours supérieur et inférieur puis à rechercher les boucles et les extensions supérieures et inférieures de chaque segment de lettres. En tenant compte des espaces, ces caractéristiques peuvent se résumer en un ensemble de 27 symboles. C'est l'agencement de ces symboles associé à une corrélation avec le mot qui permet de classer les différents styles de scripteurs.

➢ Méthode global des styles d'écriture:

Cette méthode consiste à considérer plusieurs familles de styles et à essayer d'inclure chaque scripteur dans l'une d'elles.

J.P. Crettez [CRE 95] propose une caractérisation des écritures dans le but d'améliorer le module de reconnaissance des chèques postaux par processus de Markov cachés développés au Service de Recherche Technique de la Poste. Le principe consiste à s'adapter à la variabilité de

l'écriture et non à l'occulter. Cette caractérisation permet de définir à quelle famille de styles appartient une écriture. L'écriture est une succession d'unités graphémiques appelées allographes. Ces allographes sont constitués par un enchaînement de traits de plume appelés les allotraits. Ils sont les éléments de base d'une écriture. L'auteur définit deux degrés de caractérisation. Le premier est une analyse non-supervisée qui consiste à détecter les allotraits et le deuxième est une analyse supervisée qui consiste à détecter des allographes.

Onze observations non-sémantiques sont retenues pour classer les écritures. Trois d'entre elles sont relatives à la structure du mot. Elles représentent l'épaisseur du tracé, le corps du texte et la densité spatiale de l'écriture. Huit autres sont relatives aux orientations du tracé.

Elles sont détectées sur les différentes parties rectilignes et sont obtenues à partir d'un diagramme décomposé en 4 lobes d'amplitudes et d'orientations différentes qui sont à leur tour projetés selon l'axe vertical. Cette projection permet de sélectionner 4 groupes d'allotraits. Ainsi le premier groupe représente l'ossature des hampes et des jambages, tandis que son orientation correspond à l'inclinaison de l'écriture. Le deuxième groupe renferme les liaisons naturelles inter- et intra-lettres et ainsi de suite jusqu'à obtenir les 8 observations.

4. Compitition Relative:

La classification de scripteurs et les méthodes d'évaluation de ces systèmes de classification ont évolué massivement les dernières années. En raison de l'importance du domaine et afin de le promouvoir, des chercheurs de différentes équipes appartenant à la communauté de l'analyse et la reconnaissance de documents ont organisé plusieurs compétitions depuis 2011 et ce dans le cadre des conférences très spécialisées telles que International Conference on Document Analysis and Recognition (**ICDAR**) et International Conference on Frontiers in Handwriting Recognition (**ICFHR**).

Le but des compétitions organisées en marge d'ICDAR 2011 [LOU 11, HAS 11, FOR 11] ICFHR 2012 [LOU 12, HAS 12] et ICDAR 2013 [LOU 13, MAL 13, HAS 13] était de fournir une plate-forme pour l'évaluation comparative des méthodes développées par des chercheurs appartenant à des institutions scientifiques et commerciales. L'analyse comparative des algorithmes de cette manière permet d'évaluer objectivement les performances des systèmes participants et met en évidence les points forts et points faibles de ces systèmes.

Il est important de noter également que toutes ces compétitions se sont déroulées sur des bases de données de tailles différentes ayant des contenus et des scripts différents, les protocoles d'évaluation aussi sont complètement différents, les bases de données utilisées pour l'évaluations de différents systèmes participants ont été rendues publiques juste après la fin desconférences ICDAR et ICFHR. Une récapitulation des différentes compétitions est présentée dans le tableau 2.1.

Compétition	Tâche	Script	Nombre de scripteurs	Docs par Scripteur	Systèmes participants	Type d'écriture
ICDAR2011 Arabic Writer Identification Competition [HAS 11]	Identification de scripteurs	Arabe	54	3	30	Hors-ligne
CDAR 2011 Writer Identification Contest [LOU 11]	Identification de scripteurs	Latin et Grec	26	8	8	Hors-ligne
The ICDAR 2011 Writer Identification on Music Scores Competition [FOR 11]	Identification de scripteurs	Partitions musicales	50	20	8	Hors-ligne
The ICFHR2012 Competition on Writer Identification - Challenge 2: Arabic Scripts [HAS 12]	Identification de scripteurs	Arabe	206	3	43	Hors-ligne
The ICFHR2012 Competition on Writer Identification Challenge 1: Latin/Greek Documents [LOU 12].	Identification de scripteurs	Latin et Grec	126	4	7	Hors-ligne
ICDAR 2013 Competition on Writer Identification [LOU 13]	Identification de scripteurs	Latin et Grec	350	4	12	Hors-ligne
ICDAR 2013 Competitions on Signature Verification and Writer Identification for On- and Offline Skilled Forgeries [MA 13]	Identification de scripteurs	Latin	50	6	8	Hors-ligne
ICDAR2013 - Competition on Gender Prediction from Handwriting [HAS 13]	Identification de scripteurs	Arabe et Latin	475	4	194	Hors-ligne

Tableau 2.1 : Aperçu des compétitions organisées dans le cadre des conférences ICDAR 2011 ICFHR 2012 et ICDAR 2013. [DJE 12]

4.1. Compétition pour l'identification de scripteur ICDAR 2011 :

En 2011, la premiere compétition sur l'identification de scripteur au sein de la Conférence internationale sur l'analyse de documents et de reconnaissance (ICDAR) a été réalisée. Elle a été organisée par Louloudis et al. [**LOU 11**]. Pour le concours de 26 scripteur copiés 08 pages, résultant en 208 images de documents.

Les textes sont également répartis en quatre langues (anglais, allemand, français et grec). En outre, un deuxième ensemble de données a été créé en découpant les deux premières lignes de chaque document, et ce qui rend la tâche d'identification d'écriture plus difficile, étant donné que moins de texte est présent sur ces images. Les participants doivent soumettre leur méthode sachant qu'un petit ensemble des pages de l'échantillon (qui ne faisaient pas partie de l'ensemble de données d'évaluation) et leurs méthodes ont été évaluées en fonction de critères de certain qui est décrit plus loin.

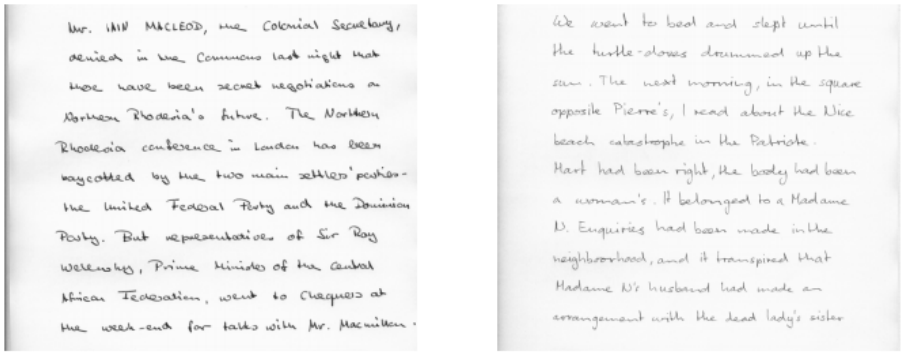


Figure 2.14 Deux images de documents des échantillons provenant de la I AM-DB. L'écriture Gauche Id 0 et l'écriture droite Id 671.

Huit méthodes différentes ont été présentées par sept institutions différentes et leurs résultats sont présentés dans [**LOU 11**]. Depuis la langue grecque utilise un alphabet différentes, la performance de toutes les méthodes sur ces images est bien pire que sur les documents écrits en alphabet latin. La figure 2.15 présente quatre échantillons de la version recadrée de ICDAR 2011 data set (Identification de scripteur: 1, Textes: 1-4).

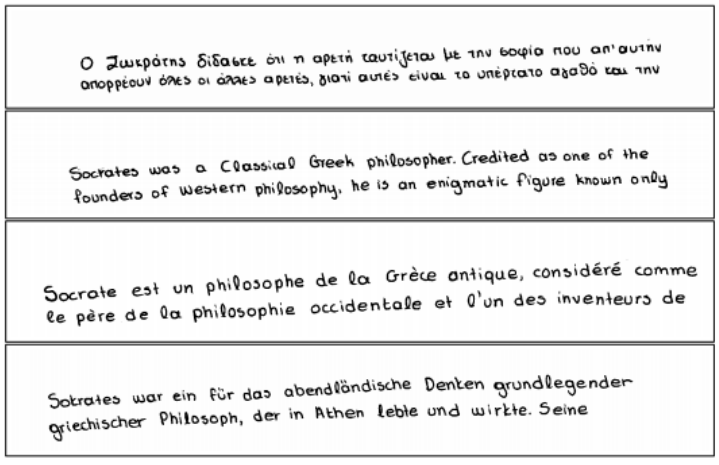


Figure 2.15 Quatre exemples d'images du base de donnée ICDAR 2011 Tous les échantillons sont Du scripteur 1.

4.2. Compitition pour l'identification de scripteur ICFHR 2012 :

Lors de la Conférence internationale sur les frontières de la reconnaissance d'écriture (ICFHR) 2012, une compétition sur l'identification de scripteur [**LOU 12**] a eu lieu et pour cette compétition une nouvelle base de données a été créée avec 400 images. 100 scripteurs copiés quatre textes en deux langues (anglais et grec).

Chaque image du document contient environ 4 lignes de texte. 4 Instituts ont participé à cette concrétion en soumettant 7 méthodes différentes. Deux exemples d'images peut être vu dans la figure 2.16.

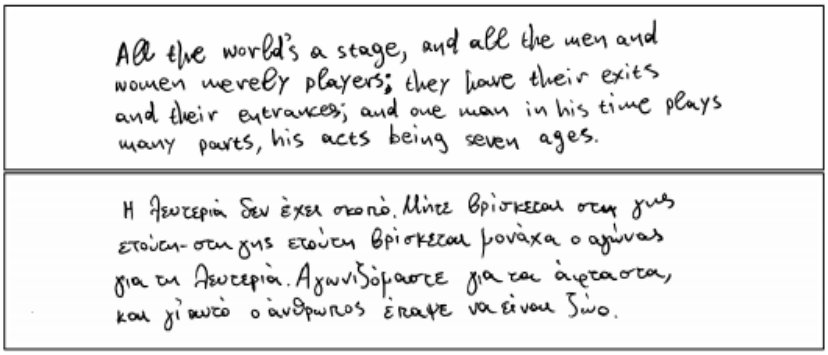


Figure 2.16: Deux échantillons du document des images de la compitition ICFHR 2012. Texte 1 et 3 de scripteur 36.

4.3. Compétition sur la prédiction du sexe du scripteur ICDAR2013 :

ICDAR 2013 Competition en Genre Prediction de l'ecriture maniscrite [HAS 13] est la première compétition dans le domaine de la détermination du sexe d'un scripteur à partir de son écriture manuscrite hors-ligne. Cette compétition a été organisée dans le cadre de la conférence ICDAR 2013 et a été organisée par une équipe de recherche du département d'informatique et d'ingénierie de l'université de Qatar, Doha, Qatar. Cette compétition a attiré l'attention de 194 équipes de recherche à travers le monde parmi lesquelles dix-neuf équipes ont accepté de partager leurs méthodes et identités.

La base de données utilisée dans cette compétition est un sous-ensemble de la base de données QUWI, un total de 475 scripteurs ont produit, chacun, 4 documents manuscrits: le premier document contient un texte manuscrit arabe qui varie d'un scripteur à l'autre, le deuxième document contient un texte manuscrit arabe qui est le même pour tous les scripteurs, le troisième document contient un texte manuscrit anglais qui varie d'un scripteur à l'autre et le quatrième document contient un texte manuscrit anglais qui est le même pour tous les scripteurs.

Les meilleures performances sont atteintes par le système soumis par Anil Thomas de Cisco Systems des Etats-unis (tablaux 2.2). Ce système applique une sélection de caractéristiques sur les caractéristiques fournies par les organisateurs de la compétition par les arbres de décision en utilisant la méthode du gradient boosting [FRI 01, FRI 02]. Les 80 caractéristiques les plus discriminantes ont été utilisées pour entraîner le système sur les écritures arabes et anglaises séparément en utilisant la même technique employée pour la sélection de caractéristiques.

Performances	Male	Female
AnilThomas	0.369	0.523
Elliot	0.436	0.53
AlexanderLarko	0.414	0.575
Megasoft	0.43	0.524
BugsBunny	0.448	0.548
BenoitPlante	0.549	0.476
Ryank	0.576	0.435
Razgon	0.494	0.546
Ihar	0.504	0.553
RandyC	0.473	0.554
Willkurt	0.46	0.554
ChaoticExperiments	0.465	0.566
Conquistator	0.493	0.552
ECPInstructors	0.45	0.587
Shiggles	0.474	0.538
JustinFister	0.454	0.597
AlejandroDubrovsky	0.495	0.532
VanZeidt	0.425	0.634
ClassifiedICMC-USP	0.636	0.597
AlKhwarizmi	0.516	0.557
Jajo	0.5	0.534
DanB	0.72	0.631
NicodeVos	0.491	0.76
RandomForest	0.722	0.632
LogisticRegression	0.382	1.008
RobustFittingofLinearModels	0.726	0.632
Hilbert	0.346	0.588
TebessaUniversity	1.673	1.61

Tablaux 2.2: Apreçu des resultat d'une compétition «gender classification » ICDAR 2013.

Le tableau 2.3 donne un aperçu des résultats des différentes méthodes sur différentes bases de données. On peut voir dans ce tableau, que, souvent, un sous-ensemble de l'ensemble de données est utilisé pour l'évaluation et les résultats ainsi, même sur le même ensemble de données ne peuvent être comparés les unes aux autres. Parmi les méthodes présentées dans le travail, seulement Djeddi et al. [DJE 12] gérer la différence entre les deux alphabets de très bons en obtenant un taux d'identification de 63,2% dans le critère Haut-2.

Auteur	Base de données	Scripteur	Pages	La langue	Top 1
Marti et al. [**MAR 01**]	IAM-DB	20	100	English	**90.7%**
Bulacu et al. [**BUL 03**]	Firemaker	250	500	Dutch	**75%**
van der Maaten et al. [**VAN 05**]	Firemaker	150	300	Dutch	**86%**
Siddiqi et al. **[SID 07]**	IAM-DB	50	100	English	**94%**
Li et al. [**LIA 09]**	HIT-MW	240	480	Chinese	**95%**
Xu et al. [**XIA 11**]	ICDAR 2011	26	208	English, Greek, German, French	**99.5%**
	ICDAR 2011 cropped	26	208	English, Greek, German, French	**79.8%**
	CVL-DB	310	1604	English (4 texts), German (1 text)	**97.7%**
Jain et al. **[JAI 11]**	IAM-DB	300	600	English	**93.3%**
	MADCAT	302	3020	Arabic	**78%**
	CVL-DB	310	1604	English (4 texts), German (1 text)	**97.9%**
	ICDAR 2013	250	1000	English, Greek	**85.5%**
Djeddi et al. **[DJE 12]**	IFN/ENIT	275	1374	Arabic	**93.5%**
	ICFHR 2012	100	400	English, Greek	**94.5%**
	ICDAR 2013	250	1000	English, Greek	**93.4%**
	CVL-DB	310	1604	English (4 texts), German (1 text)	**97.6%**
Jain et al. [**JAI 13]**	ICDAR 2013	250	1000	English, Greek	**95.1%**
	IAM-DB	301	602	English	**96.5%**
	ICFHR 2012	100	200	English	**98%**
	ICFHR 2012	100	200	Greek	**97.5%**
	MADCAT	316	632	Arabic	**87.5%**
Du et al.**[YOU 10]**	-	50	100	Chinese	**68%**
Hiremath et al. **[HIR 10]**	-	30	750	English	**87.94%**
	-	30	750	Kannada	**91.45%**
Christlein et al. **[CHR 14]**	CVL-DB	310	1604	English (4 texts), German (1 text)	**99.2%**
	ICDAR 2013	250	1000	English, Greek	**97.1%**
Jain et al. [**JAI 14**]	IAM DB	657	1314	English	**94.7%**
	CVL DB	310	1604	English (4 texts), German (1 text)	**98.3%**
		100	400	Kannada	**82.75%**
		100	400	Devanagari	**85.25%**
		100	1200	English, Kannada, Devanagari	**82.19%**

Tableau 2.3 : Résultats des différentes méthodes sur différentes bases de données.

Conclusion

Nous venons de voir dans ce chapitre les méthodes qui nous ont semblé les plus caractéristiques depuis ces 30 dernières années dans le domaine de l'identification de scripteurs. Ce sujet fait partie du domaine de la reconnaissance de formes et de l'analyse de l'écrit. La recherche a beaucoup évolué durant cette période.

Au début on pouvait classer les études selon deux approches : l'approche contextuelle et l'approche non contextuelle. Toutes les méthodes décrites obtiennent un taux d'identification aux alentours de 98% [SEP 03]. Mais ces taux sont souvent obtenus sur peu de scripteurs, sous contraintes d'écriture ou avec un choix du nombre des caractéristiques les mieux adaptées pour le processus.

Ensuite, Les modélisations d'allographes ont permis de s'affranchir de la trop grande variabilité de l'écriture. Les méthodes globales par style obtiennent de très bons taux d'identification de façon plus robuste et surtout sans contrainte pour le scripteur. Pour l'instant, elles consistent à créer plusieurs familles de styles d'écriture et à essayer de rattacher une écriture inconnue à une des familles. Elles sont surtout utilisées pour l'aide à la reconnaissance à cause de leur application industrielle mais pas pour faire de l'identification un but fini.

Ainsi après avoir passé en revue des méthodes sur l'identification de scripteurs, une méthode très intéressante en est ressortie ; celle de l'équipe de T. Paquet [BEN 02] sur l'extraction d'invariants. Cette technique consiste à déterminer des portions extraites dans l'écriture qui possèdent des propriétés d'invariance au sein de l'écriture d'un même scripteur.

L'extraction des invariants associés à un scripteur peut être utilisée aussi bien comme aide à la reconnaissance du texte que pour l'identification du scripteur.

Les études menées par N. Vincent ont permis de séparer les écritures par familles de styles grâce au calcul de la dimension fractale. Ainsi il en ressort que les écritures ont un comportement fractal et qu'il est possible d'utiliser la « *fractalité* » de l'écriture pour en extraire des caractéristiques propres à chaque scripteur. Cette étude laisse penser qu'avec une technique de

compression fractale, il serait également possible de caractériser un scripteur et donc des propriétés inhérentes à celui-ci.

Associer les invariants au comportement fractal de l'écriture nous est apparu comme étant un axe de recherche prometteur. Il permettrait ainsi, non plus de relier un scripteur à un groupe mais d'extraire des caractéristiques inhérentes à chaque scripteur.

Chapitre III

Une méthode locale pour la détermination du sexe à partir de textes manuscrits

Chapitre III

Une méthode locale pour la détermination du sexe à partir de textes manuscrits

1. Introduction:

L'écriture est un moyen essentiel de communication dans notre civilisation. Elle s'est développée et a évolué au fil du temps. Comme toute production humaine, l'écriture est soumise à de nombreuses variations d'origines très diverses qui pourraient être historique, géographique, ethnique ou sociale. Toutefois, elle a aussi un lien fort avec des caractéristiques innées d'une personne qui expliquent la grande variabilité observée entre les écrits de différentes personnes, même s'ils appartiennent à des communautés voisines. Contrairement à la version électronique ou imprimée, le texte manuscrit comporte des informations supplémentaires sur la personnalité de la personne qui a écrit. Il existe un certain degré de stabilité dans le style d'écriture d'un individu ou d'une population d'individus [BOU 08], ce qui rend possible le processus de classification des scripteurs selon plusieurs attributs démographiques (identité, sexe, âge, origine ethnique,…… , etc).

Bien qu'un nombre important d'organisations emploient l'analyse de l'écriture manuscrite pour le profilage de la personnalité [ROY 00, SHA 94], la corrélation entre la personnalité et l'écriture manuscrite reste discutable [RIC 83, ROB 97, EFR 89, ADR 87] et doit encore être validée sur des bases scientifiques. La seule corrélation significative et qui a été validée expérimentalement existe entre l'écriture et le sexe du scripteur [JAM 91, HAM 96, BEE 05, WIL 96, GOO 45, SOK 12, BUR 02].

La détermination automatique du sexe d'un individu à partir de son écriture manuscrite, cependant, a été un domaine relativement peu exploré avec seulement quelques contributions significatives. L'identification des classes démographiques d'un individu telles que l'origine ethnique, le sexe, la main dont il se sert pour écrire ainsi que l'âge à partir de documents manuscrits a été étudiée dans certains travaux [DJE 10, SUN 01, KAR 05, HAM 96, LIW 06, LIW 11, SOK 12]. Plusieurs études ont montré que le sexe d'un scripteur peut être détecté à partir de l'écriture manuscrite [HAM 96, JAM 91, WIL 96] avec divers degrés de succès. Ceci est soutenu par l'observation que les individus interagissent avec les documents manuscrits, par exemple, les enseignants, apprennent à distinguer entre écritures

des filles et des garçons avec le temps. Les examinateurs humains formés sont également en mesure de prédire le sexe un scripteur d'un document manuscrit avec une précision suffisante pour s'affranchir du hasard [SOK 12]. Les psychologues attribuent les différences dans l'écriture des scripteurs masculins et féminins à des différences dans la coordination motrice [JAM 91] ou les différents types d'hormones qu'ils produisent [WIL 96]. En tout cas, les chercheurs sont en désaccord en ce qui concerne la corrélation entre le genre et l'écriture manuscrite.

Dans ce chapitre, nous présentons notre proposition d'une méthode de classification automatique de documents manuscrits par le sexe de leur scripteur en utilisant un ensemble de descripteurs locaux de l'écriture [SID 10] qui sont inhérents à la manière dont un scripteur écrit spécifiquement des caractères. Chaque échantillon d'écriture est représenté par un ensemble de caractéristiques qui sont utilisées par un classificateur afin qu'il puisse apprendre à distinguer entre les deux catégories d'écritures (masculine et féminine). La classification est effectuée à l'aide des séparateurs à vaste marge (SVMs). La méthode développée est en utilisant la base de données QUWI [ALM 12] et les résultats enregistrés sont assez intéressants.

Dans la deuxième section de ce chapitre, nous présentons la base de données utilisée pour évaluer le système développé. Dans la section trois, nous décrivons la méthode proposée pour la génération du codebook suivie d'une description de la méthode utilisée pour la caractérisation des scripteurs, ainsi nous présentons Les résultats expérimentaux et leur analyse.

2. Base de données utilisée:

Pour évaluer les performances de la méthode proposée, nous avons utilisé des échantillons d'écriture issus de 475 scripteurs différents de la base QUWI qui a été développée par une équipe de recherche de l'université de Qatar [ALM 12].

Dans la base de données QUWI, chaque scripteur a contribué à la production de 4 pages, deux en arabe et deux en anglais (voir figure 3.1). Le premier document contient un texte manuscrit arabe qui varie d'un scripteur à l'autre, le deuxième document contient un texte manuscrit arabe qui est le même pour tous les scripteurs, le troisième document contient un texte manuscrit anglais qui varie d'un scripteur à l'autre et le quatrième document contient un texte manuscrit anglais qui est le même pour tous les scripteurs. Cette variabilité du

contenu textuel des documents écrits par une même personne permet d'utiliser la base de données en mode dépendant du texte ainsi qu'en mode indépendant du texte.

Dans nos expérimentations, nous avons divisé la base de données QUWI en 3 parties, des échantillons provenant de 82 scripteurs sont utilisés pour générer les différents codebooks, ceux de 201 scripteurs constituent l'ensemble d'apprentissage (d'entraînement) tandis que ceux de 192 scripteurs sont utilisés comme ensemble de test. La répartition des scripteurs sur les trois parties est résumée dans le tableau 3.1. La répartition des scripteurs reste la même dans les différentes expérimentations.

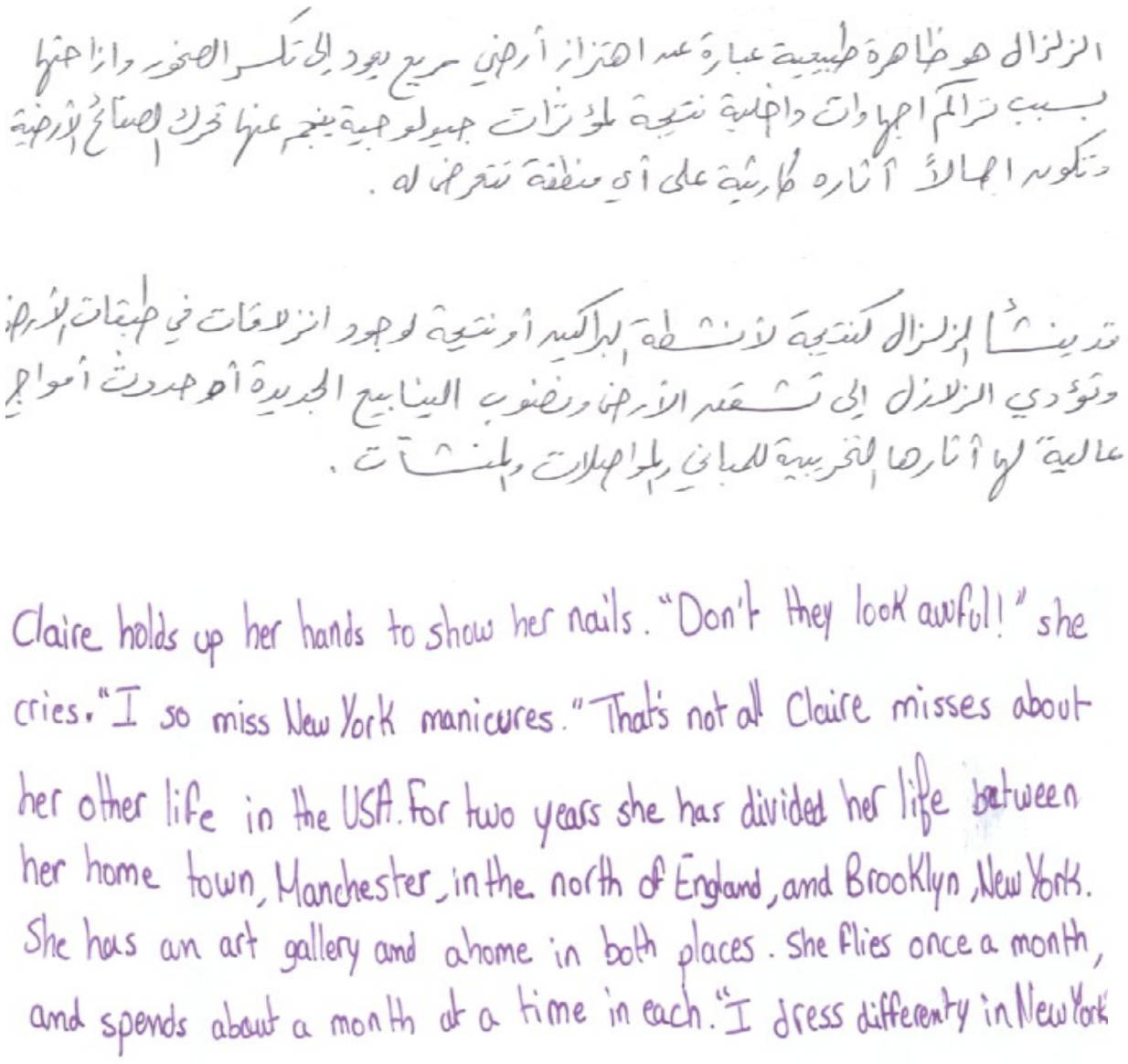


Figure 3.1. Échantillons d'écriture manuscrite de la base de données QUWI, en arabe et en anglais.

	Codebook	Apprentissage	Test
Scripteur	82	201	192
Echantillons	328	804	768

Tableau 3.1. Répartition de la base de données QUWI.

3. Méthode proposée:

Depuis quelques années, la tendance des recherches sur la classification de scripteurs, s'est focalisée vers les méthodes basées sur les codebooks où l'écriture est segmentée en graphèmes ou en allotrais (petits fragments d'écriture) qui sont ensuite comparés avec des éléments d'un codebook soit propre au scripteur [BEN 02] soit un codebook universel [BEN 05] [SCH 04] [BUL 05]. Ces méthodes ont été très développées au cours des dernières années et elles ont démontré une grande performance pour l'identification et la vérification de scripteurs. Dans cette étude, nous avons opté pour l'évaluation de la méthode proposée par Siddiqi et Vincent [SID 10], qui avait été proposée pour l'identification et la vérification de scripteurs, en reconnaissance du sexe d'une personne en se basant sur son écriture manuscrite. Cette méthode est basée sur l'extraction de petits fragments d'écriture, ces fragments sont extraits par un découpage adaptatif de l'écriture en imagettes de taille *n*n* pixels qui seront ensuite regroupées en utilisant l'algorithme de regroupement K-means et ce afin d'avoir un codebook universel (voir figure 3.2) contenant les formes redondantes et les plus fréquente dans script quelconque.

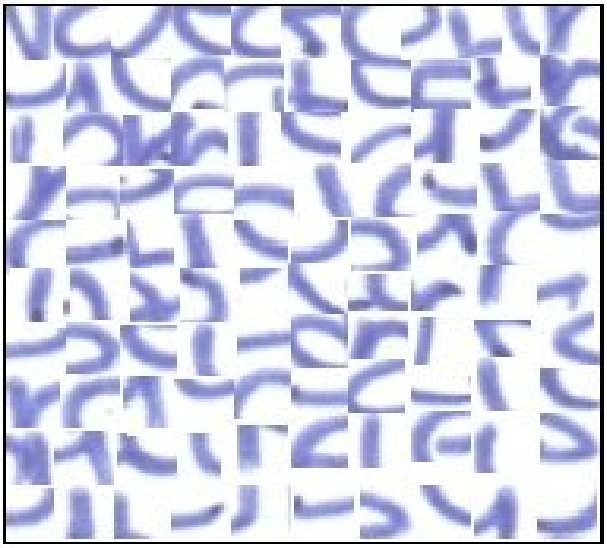


Figure 3.2. Codebook universel de size 100 avec découpage 19*19 obtenu à partir d'échantillons de la base QUWI.

3.1. Architecture du système proposé:

Notre système détermination de sexe d'un scripteur est basé sur trois étapes principales : génération de codebook , extraction de caractéristiques et décision. Une base de référence des scripteurs est créée en effectuant l'extraction des caractéristiques de l'écriture de chaque scripteur. Le sexe d'un scripteur d'un document inconnu est alors identifié ultérieurement pendant une étape de décision. L'architecture générale de notre système est présentée sur la Figure 3.2 et les différentes étapes sont décrites dans les prochaines sections.

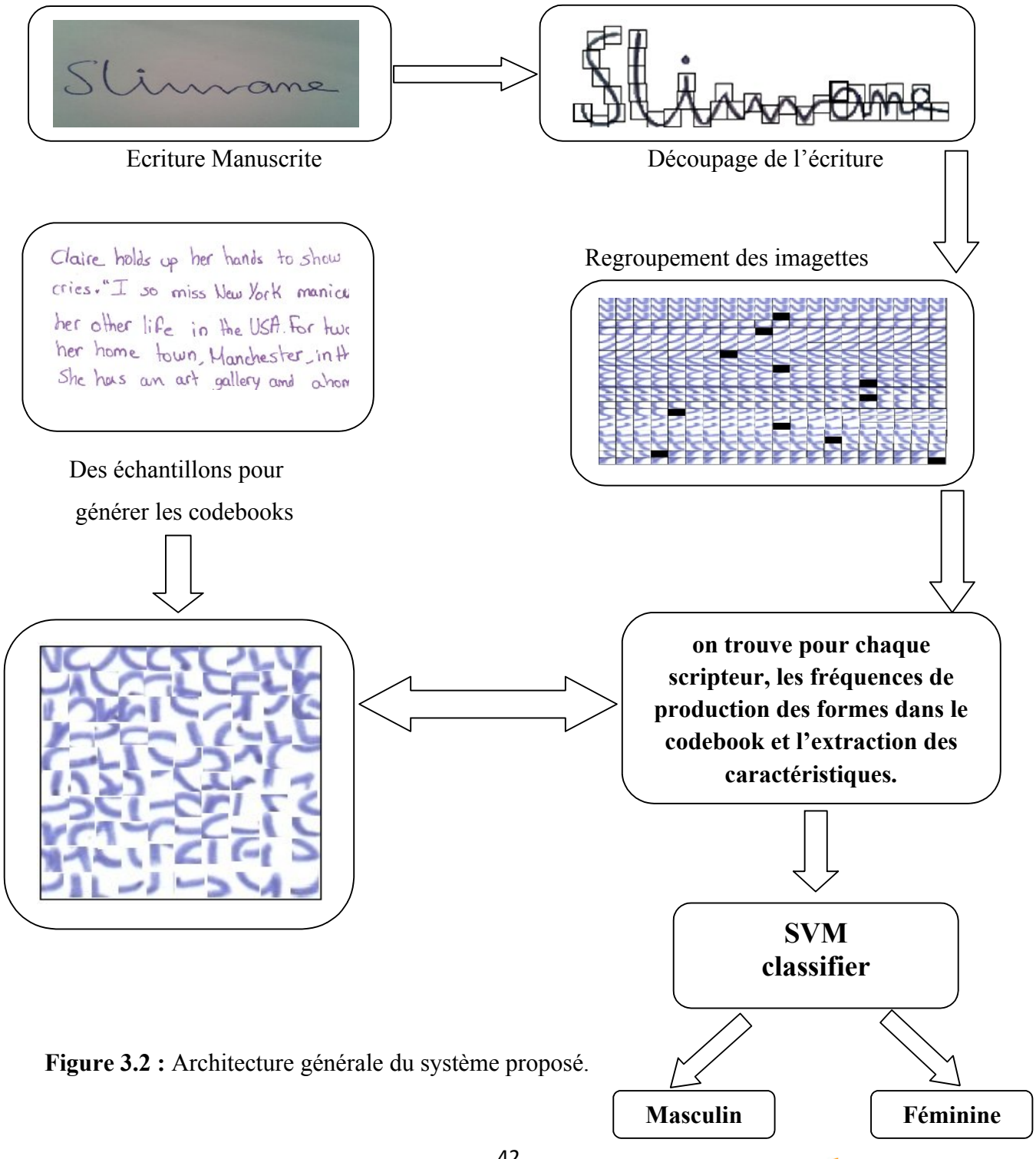


Figure 3.2 : Architecture générale du système proposé.

3.2. Découpage de l'écriture:

Le découpage de l'écriture est utilisé pour extraire des éléments inhérents au scripteur, donc c'est une partie importante du processus. Il doit être dépendant du tracé pour que les contenus puissent être comparables. On a choisi un découpage en carrés de taille *nxn* où la taille *n* a été variée du 11 au 19. En utilisant un algorithme adaptatif de positionnement, ces fenêtres sont placées sur le texte divisant ainsi l'écriture en un grand nombre de petites imagettes [SID 08]. Les fenêtres sont consécutivement positionnées en suivant le squelette de l'écriture. Sans chercher à reconstituer l'ordre du tracé, nous réalisons un suivi du trait. C'est pour cela que nous avons opté pour un découpage lié à la notion d'écriture et plus précisément de trait, c'est-à dire en privilégiant la direction verticale. Nous utilisons un découpage en carrés en ce qui concerne les ranges qui contiennent une portion de trait. Evidemment ces carrés sont sans recouvrement pour réaliser la partition. Nous ne porterons que peu d'attention aux zones ne contenant pas de portion de trait d'écriture (zones blanches).

La technique la plus simple consisterait à découper la totalité de l'image régulièrement de gauche à droite et de bas en haut par exemple. Toutefois, comme nous ne traitons pas les imagettes vides, ou blanches, nous avons préféré mettre en œuvre un découpage mot par mot ou plus exactement composantes connexes par composantes non connexes. L'origine verticale étant ainsi propre à chaque mot, on ajuste le quadrillage horizontalement pour recouvrir le mot avec un minimum d'imagettes. Ainsi l'image reste découpée de façon régulière selon les lignes et sur chaque ligne un carré peut être décalé vers la gauche ou vers la droite pour optimiser le partitionnement de l'écriture.

Nous avons utilisé la méthode proposée par Siddiqi et Vincent [SID 10] qui suit la direction naturelle de la main et du tracé. Il ne s'agit pas de retrouver l'ordre exact du tracé mais de positionner les imagettes par rapport aux éléments caractéristiques fondamentaux. De manière à trouver les extrémités du trait, nous nous référons au squelette des différentes composantes connexes, et positionnons une première fenêtre sur un point extrême du squelette.

Pour chaque fenêtre nous définissons quatre drapeaux : est, ouest, nord et sud ; le drapeau étant affiché si le trait sort de l'imagette par ce coté, si le squelette sort de E (ou de O), nous plaçons la prochaine fenêtre vers la droite (gauche respectivement) de la position présente (sur l'image originale), et déplaçons la fenêtre dans la direction verticale (en haut ou en bas) pour trouver le meilleur placement suivant. par contre, s'il sort de N (ou de S), nous plaçons la prochaine fenêtre au-dessus (en dessous respectivement), et déplaçons la fenêtre

dans la direction horizontale (gauche ou droite) de sorte qu'elle soit bien placée. Le processus est illustré dans **la figure 3.3**.

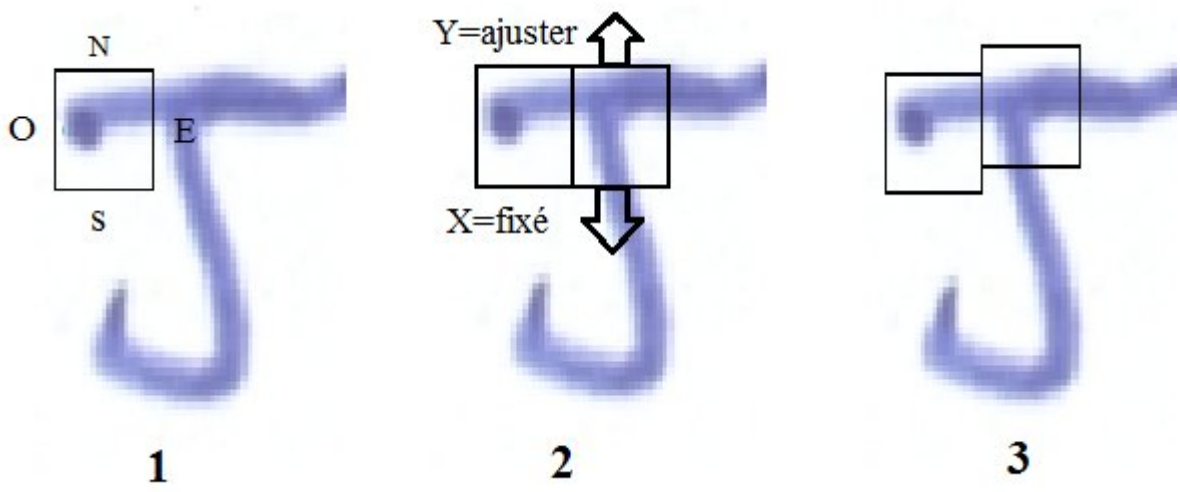


Figure 3.3 : placement de fenêtre (1) début d'un trait, (2) positionnement initial de la fenêtre suivante, (3) glissement de la fenêtre par rapport au trait.

Dans le cas ou le squelette sortirait de plus d'une direction, nous traitons séparément chacune des directions concernées. L'algorithme du placement des fenêtres est appliqué à chacune des composantes et par conséquent nous avons un découpage de l'écriture en petites imagettes n*n = 15*15; illustré pour le nom, SLIMANE, dans la figure 3.4.

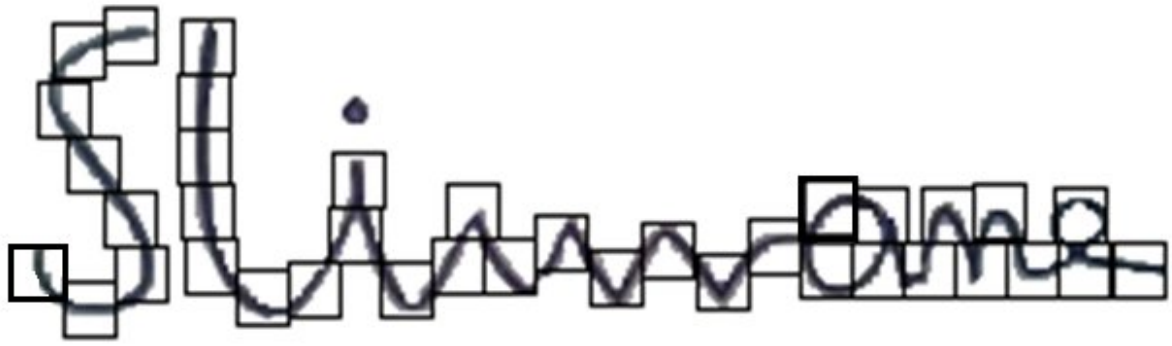


Figure 3.4. Illustration de la méthode du découpage de l'écriture.

La taille des fenêtres utilisées pour le découpage de l'écriture est un paramètre important de la méthode proposée, elle est choisie en fonction de deux critères. Le premier

est que la taille doit être assez grande pour contenir assez d'informations sur le style de scripteur et le second est que cette même taille doit être assez petite pour assurer une bonne performance de reconnaissance. Pour nos évaluations nous avons testé différentes tailles de fenêtres à savoir 13x13, 15x15, 17x17 et 19 x19.

3.3. Caractérisation des imagettes:

Une fois que le texte est divisé en imagettes, nous procédons à l'extraction de certains descripteurs de forme sur chaque imagette. Ces descripteurs comprennent les histogrammes horizontaux et verticaux ainsi que les profils supérieur et inférieur. Chaque imagette est alors représentée par un vecteur de dimension $d = 4n$, où n est la taille de la fenêtre. En représentant chaque imagette par un vecteur, on procède ensuite à leur regroupement de manière à réduire la quantité de données et à rendre le résultat indépendant de la quantité de texte étudié.

3.4.Regroupement des imagettes:

La méthodologie de classification dépendra de codebook universel.

- Représentant chaque imagette par un vecteur.
- On procède ensuite à leur regroupement de manière à réduire la quantité de données et à rendre le résultat indépendant de la quantité de texte étudié.
- L'objectif est de grouper les formes produites par le même geste de la main dans les mêmes classes.

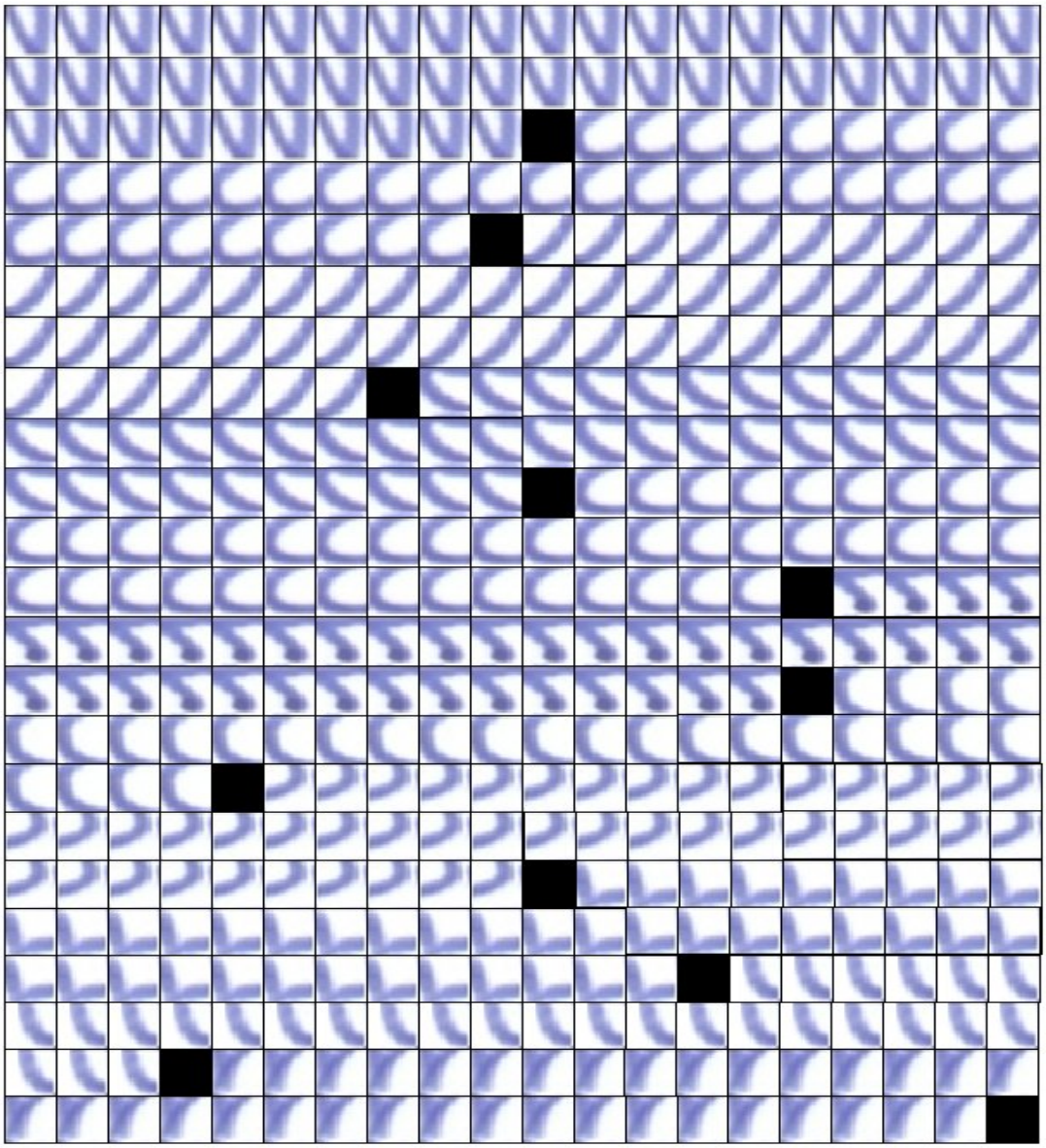


Figure 3.5. Illustration du regroupement des imagettes obtenu à partir d'échantillons de la base QUWI et de taille d'imagette 19*19.

Pour un codebook universel, généré à partir d'une partie d'échantillons manuscrits de la base QUWI Un total de 164 échantillons arabe et 164 échantillons anglais d'écriture sont utilisés pour produire les codebooks , nous avons évalué un certain nombre d'algorithmes qui ne demandent pas un choix préalable du nombre de classes car ce paramètre va varier d'une écriture à une autre. Après une série d'expériences sur l'ensemble de validation, nous avons choisi une classification hiérarchique pour extraire les formes redondantes de l'écriture et générer les codebooks . Pour chaque classe dans le codebook, on estime sa probabilité d'apparition et aussi on calcule le vecteur moyen représentant la classe. L'ensemble de ces probabilités peuvent être considérées comme une distribution hD, où chaque cardinal de

classe de *hD* représenterait la probabilité d'émission de la forme respective par l'auteur du document *D*. Cette distribution est ensuite utilisée pour caractériser l'auteur d'un échantillon donné. tandis que les fragments sont regroupés en utilisant l'algorithme *k-means* dans l'espace de caractéristiques, Une fois le codebook généré, on trouve pour chaque scripteur, les fréquences de production des formes dans le codebook, la répartition étant caractéristique du scripteur.

Cette méthode consiste à construire une partition en k classes en sélectionnant k individus comme centres des classes, ils sont tirés au hasard dans l'ensemble des scripteurs . Après cette sélection, chaque scripteur est associé au centre le plus proche ce qui donne une partition en k classes, les centres des classes seront actualisés et de nouvelles classes seront formées suivant le même principe. L'entier k désigne le nombre maximum de classes désiré. Généralement la partition obtenue est localement optimale car elle dépend du choix initial des centres. Pour cela les résultats entre deux exécutions de l'algorithme peuvent varier de façon significative. Cette méthode est simple, compréhensible et est applicable à des données dans un espace de grande dimension. La contrainte principale est que le nombre de classes doit être fixé au départ, Figure 3.6. K-Means Exemple.

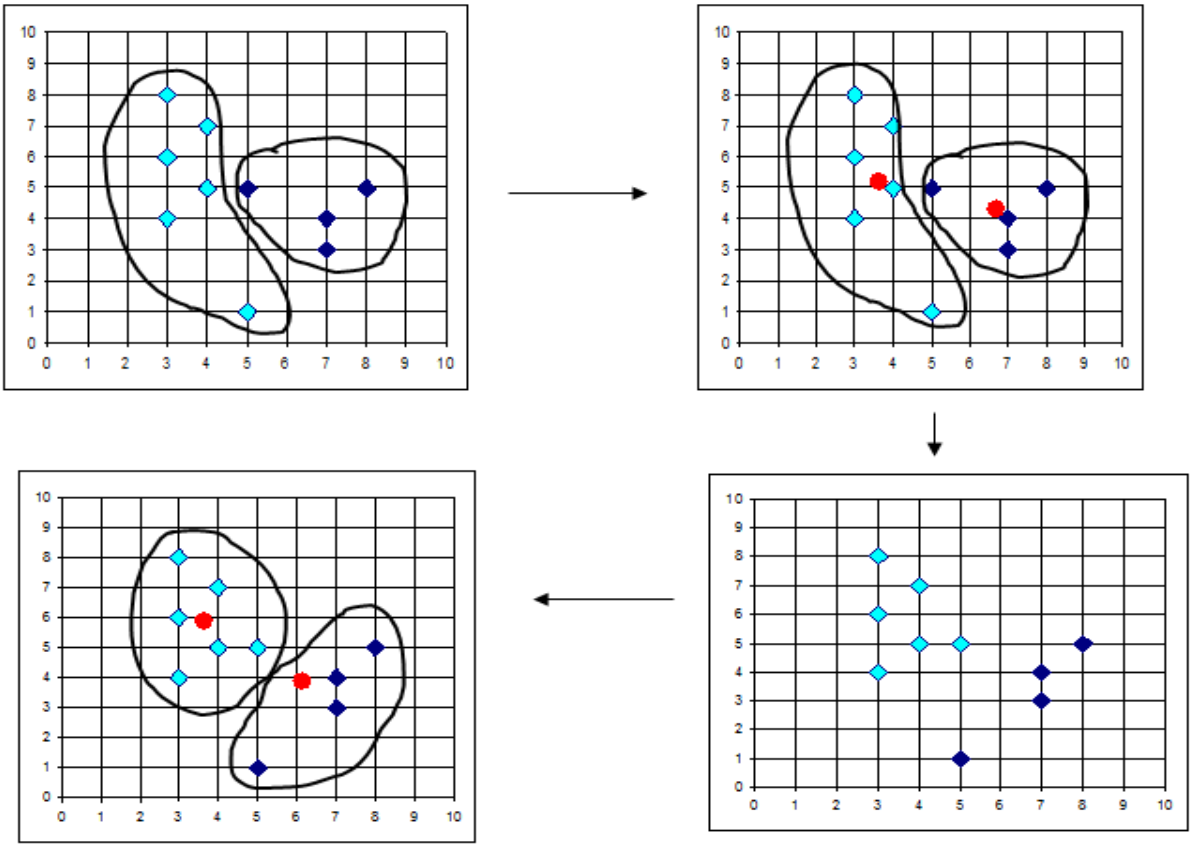


Figure 3.6. K-Means Exemple.

3.5. Détermination du sexe d'un scripteur:

Lorsqu'on utilise un codebook universel, chacun des fragments de l'image requête est attribué à l'une des classes dans le document de référence. Pour comparer le document en question *Q* de sexe 0 ou 1(0 pour femelle, 1 pour male) avec un document de référence *D* de sexe 0 ou 1, pour chaque forme dans le document de test, on retrouve la plus proche classe dans le document *D*. Nous trouvons par conséquent comment les entrées dans le codebook de *D* sont distribuées dans le document *Q*. Ainsi, en fait, le sexe d'un scripteur de document en question est représenté dans l'espace du document de référence et les deux écritures sont comparées en calculant la distance entre les distributions respectives *hD* et *hDQ*.

Pour le codebook, on compare les imagettes extraites du document en question aux formes contenues dans le codebook et donc on trouve les fréquences d'occurrence des formes du codebook pour un scripteur particulier. Deux écritures sont ensuite comparées en calculant la distance entre les distributions de probabilité respectives *hD* et *hQ* de produire les formes du codebook.

La détermination de sexe d'un scripteur est effectuée en calculant la distance entre l'image requête *Q* et toutes les images dans la base d'apprentissage en utilisant une caractéristique sélectionnée, le sexe d'un scripteur de *Q* étant identifié comme le sexe d'un scripteur du document qui donne la distance minimale. Cela correspond à la classification du plus proche voisin (*knn* avec $k = 1$).

3.6. Classification:

Le système de classification de sexe d'un scripteur proposé est conçu pour classer automatiquement les écrivains en deux catégories qui sont «homme» ou «femme». des images hors ligne du texte manuscrit sont utilisés pour calculer les caractéristiques locales. Le classificateur SVM (les séparateurs à vaste marge), reçoit des fonctions de données et décide si le texte manuscrit a été écrit par un homme ou une femme. la classification de sexe d'un scripteur a été basée sur certaines fonctionnalités hors ligne classiques. Étant donné qu'aucune des enquêtes exhaustives ont été menées pour décrire les données hors ligne pour la classification de sexe d'un scripteur, la principale contribution de ce travail consiste à utiliser plus puissantes fonctionnalités hors ligne. Plus précisément, notre attention se concentre sur les caractéristiques locales, qui sont généralement plus approprié pour l'écriture de caractérisation.

La classification est réalisée en utilisant classifieur (SVM). Le classifieur est entraîné à l'aide d'ensembles de caractéristiques extraites de l'ensemble de données d'apprentissage (d'entraînement) tandis que les différents paramètres de classifieur sont déterminés empiriquement sur l'ensemble des données de validation.

la couche de sortie comprend deux classes (masculine et féminine) est déterminé en fonction de la dimension du vecteur de caractéristiques (à l'aide l'ensemble de données de validation).

Le classifieur SVM est basé sur un noyau polynomial. les paramètres du SVM sont déterminés de façon empirique sur l'ensemble de données de validation. Pour la mise en oeuvre, nous avons utilisé la boîte à outils Matlab des SVM décrite dans [CAN 05].

Les performances de ce classifieur ainsi que les différentes évaluations effectuées sont discutées en détails dans la section suivante.

4. Résultats expérimentaux et discussion:

Cette section présente les séries d'expérimentations que nous avons effectuées afin d'évaluer l'efficacité des caractéristiques proposées pour prédire le sexe du scripteur de l'écriture considérée. Les différentes évaluations sont menées sur la base de données QUWI. Dans toutes les expérimentations, nous nous assurons qu'il n'y a pas des échantillons du même scripteur appartenant aux ensembles de test et d'apprentissage en même temps. Cela pourrait conduire à faire correspondre le document en question avec un autre échantillon du même scripteur dans l'ensemble d'apprentissage ce qui va ramènera notre problème à un problème d'identification du scripteur et plus à un problème de reconnaissance du sexe. Ainsi, dans les expérimentations où plus d'un échantillon par scripteur sont considérés, tous les échantillons d'un scripteur donné appartiennent soit à l'ensemble d'apprentissage (d'entraînement) ou à celui du test.

Le tableau 3.2. présente les taux Globale de classification sur la base de données QUWI, des échantillons de 201 scripteurs sont utilisés pour l'apprentissage (804 échantillons), tandis que ceux de 192 scripteurs (768 échantillons) sont utilisés pour les tests.

Des taux global de bonne classification de l'ordre de **71.88%** et **68.49%** sont atteints sur les bases de données QUWI-Anglais et QUWI-Arabe. Ces résultats sont comparables à ceux de l'état de l'art discutés est cependant intéressant de noter que le système proposé est évalué sur des bases de données beaucoup plus grandes par rapport aux méthodes existantes. En comparant les performances de classifieur (SVM), Une autre observation aussi intéressante

est que la combinaison des différents taille de fenêtre de découpage entraîne des améliorations marginales dans les taux globaux de classification.

Pour les expérimentations ultérieures, nous allons donc discuter les résultats des différentes catégories de tailles. En plus des évaluations sur les ensembles de données complets, nous analysons également les performances des caractéristiques proposées dans un certain nombre de scénarios spécifiques.

Base de données	Taille de codebook	Taille de fenêtre de découpage	Taux de classification SVM
QUWI-Anglais	100	13*13	68.75%
		15*15	67.45%
		17*17	66.67%
		19*19	71.87%
	200	13*13	68.23%
		15*15	69.53%
		17*17	69.01%
		19*19	71.87%
	300	13*13	69.79%
		15*15	69.27%
		17*17	68.49%
		19*19	70.31%
	400	13*13	69.27%
		15*15	**71.88%**
		17*17	69.27%
		19*19	71.61%
QUWI-Arabe	100	13*13	63.54%
		15*15	65.89%
		17*17	**68.49%**
		19*19	65.89%
	200	13*13	66.15%
		15*15	66.41%
		17*17	64.33%
		19*19	66.93%
	300	13*13	67.19%
		15*15	63.54%
		17*17	64.06%
		19*19	66.93%
	400	13*13	65.36%
		15*15	66.15%
		17*17	64.84%
		19*19	67.45%

Tableau 3.2. Taux de classification SVM Globale des évaluations sur les bases de données QUWI .

Le tableau 3.3 présente les taux de classification de sexe male, des échantillons de 201 scripteurs sont utilisés pour l'apprentissage 804 échantillons, tandis que ceux de 192 scripteurs utilisé 768 échantillons pour les tests. Des taux global de bonne classification SVM male de sexe masculin de l'ordre de **70.37%** et **70.99%** sont atteints sur les bases de données QUWI-Anglais et QUWI-Arabe.

Base de données	Taille de codebook	Taille de fenêtre de découpage	Taux de classification SVM (Male)
QUWI-Anglais	100	13*13	69.75%
		15*15	67.90%
		17*17	64.81%
		19*19	69.75%
	200	13*13	59.88%
		15*15	61.73%
		17*17	66.67%
		19*19	**70.37%**
	300	13*13	62.96%
		15*15	62.96%
		17*17	62.35%
		19*19	69.75%
	400	13*13	64.81%
		15*15	**70.37%**
		17*17	61.73%
		19*19	67.90%
QUWI-Arabe	100	13*13	59.88%
		15*15	70.37%
		17*17	**70.99%**
		19*19	67.28%
	200	13*13	66.67%
		15*15	69.75%
		17*17	67.90%
		19*19	70.37%
	300	13*13	66.67%
		15*15	60.49%
		17*17	64.20%
		19*19	67.28%
	400	13*13	64.20%
		15*15	69.14%
		17*17	66.67%
		19*19	67.90%

Tableau 3.3. Taux de classification SVM Male des évaluations sur bases de données QUWI .

Le tableau 3.4 présente les taux de classification de sexe Female Des taux global de bonne classification SVM Female de sexe féminine de l'ordre de **75.22%** et **67.57%** sont atteints sur les bases de données QUWI-Anglais et QUWI-Arabe, le taille de fenêtre de découpage 15*15 avec k= 200 taille de codebook donne meilleur résultat.

Base de données	Taille de codebook	Taille de fenêtre de découpage	Taux de classification SVM (Female)
QUWI-Anglais	100	13*13	68.01%
		15*15	67.12%
		17*17	68.02%
		19*19	73.42%
	200	13*13	74.32%
		15*15	**75.22%**
		17*17	70.72%
		19*19	72.97%
	300	13*13	74.77%
		15*15	73.87%
		17*17	72.97%
		19*19	70.72%
	400	13*13	72.52%
		15*15	72.97%
		17*17	74.77%
		19*19	74.32%
QUWI-Arabe	100	13*13	66.21%
		15*15	62.61%
		17*17	66.67%
		19*19	64.86%
	200	13*13	65.77%
		15*15	63.96%
		17*17	61.71%
		19*19	64.41%
	300	13*13	**67.57%**
		15*15	65.77%
		17*17	63.96%
		19*19	66.67%
	400	13*13	66.22%
		15*15	63.96%
		17*17	63.51%
		19*19	67.12%

Tableau 3.4 Taux de classification SVM Female des évaluations sur bases de données QUWI.

La Figure 3.8 représente des colonnes graphiques examinent des résultats de classification Femelle et Male des évaluations sur les bases de données QUWI, obtenus dans le cas codebook taille **100** sur documents manuscrites en **anglais**, où l'on note que le sexe femelle du scripteur déterminé en taux supérieur à **73%** dans le découpage **19*19** par rapport aux autres découpages (13*13. 15* 15. 17*17). D'autre part le sexe Male du scripteur déterminé en taux supérieur à **69 %** dans les deux découpages **13*13** et **19*19** par rapport aux autres découpages (15*15. 17*17).

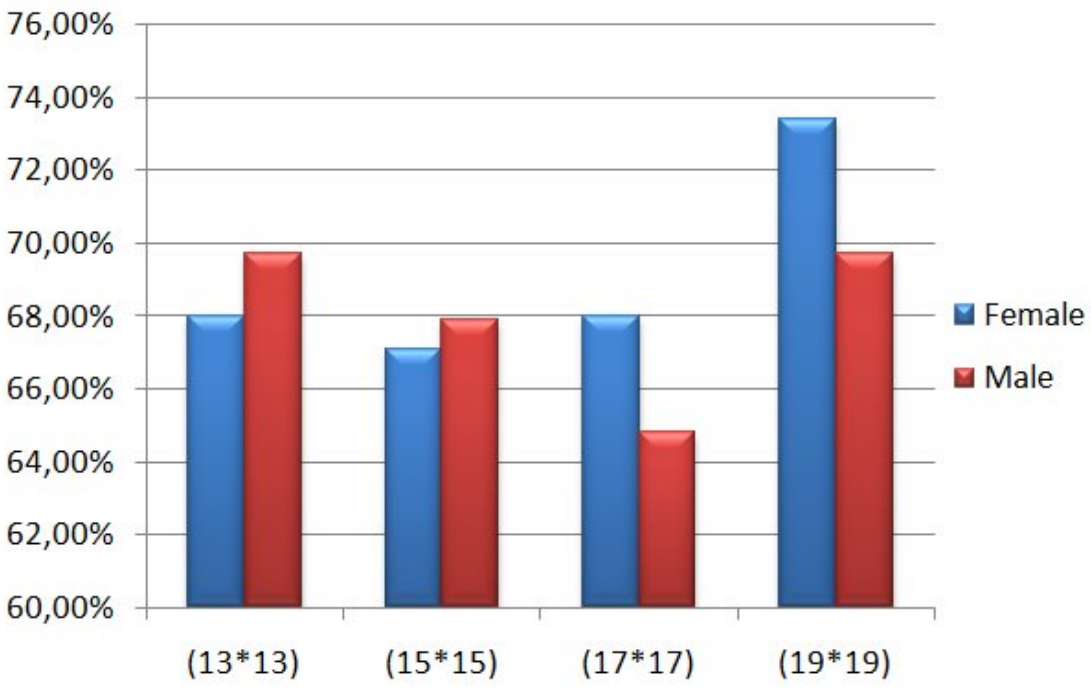


Figure 3.8 : Colonnes graphique des résultats obtenus dans le cas codebook size **100** (Anglais).

La Figure 3.9 explique une étude comparative par des colonnes graphiques examinant des résultats de classification Femelle et Male des évaluations sur les bases de données QUWI, en **arabe**, où l'on note que le sexe femelle du scripteur déterminé par **66.67 %** dans le découpage **17*17** par rapport aux autres découpages (13*13.15*15.19*19). D'autre part le sexe Male du scripteur déterminé par **70.99 %** dans le découpage **17*17** par rapport aux autres découpages (15*15. 17*17.19*19).

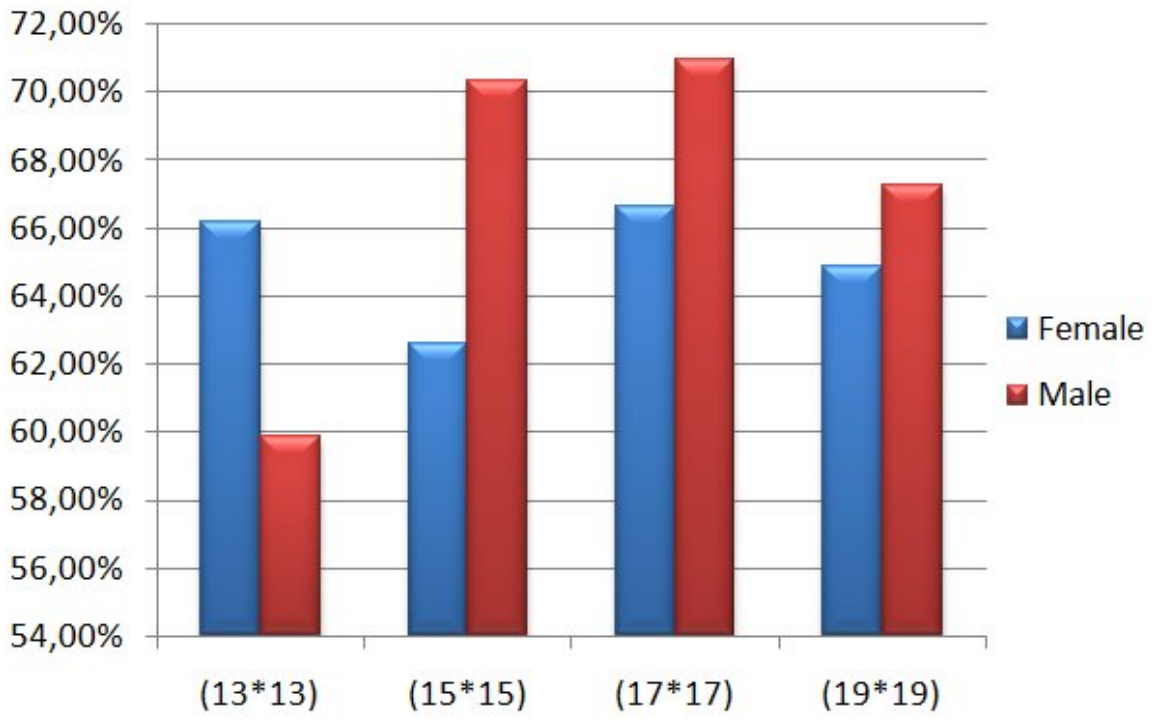


Figure 3.9 : Colonnes graphique des résultats obtenus dans le cas codebook size **100** (Arabe).

La Figure 3.10 en basant sur les colonnes graphiques qui examinent des résultats de classification Femelle et Male des évaluations sur les bases de données QUWI, signale dans le cas codebook taille **200** sur documents manuscrites en **anglais**, où l'on note que le sexe femelle du scripteur déterminé en taux supérieur à **75%** dans le découpage **15*15** par rapport aux autres découpages (13*13.17 *17.19*19). D'autre part le sexe Male du scripteur déterminé par **70.37 %** dans le découpage **19*19** par rapport aux autres découpages (13*13. 15*15.17*17).

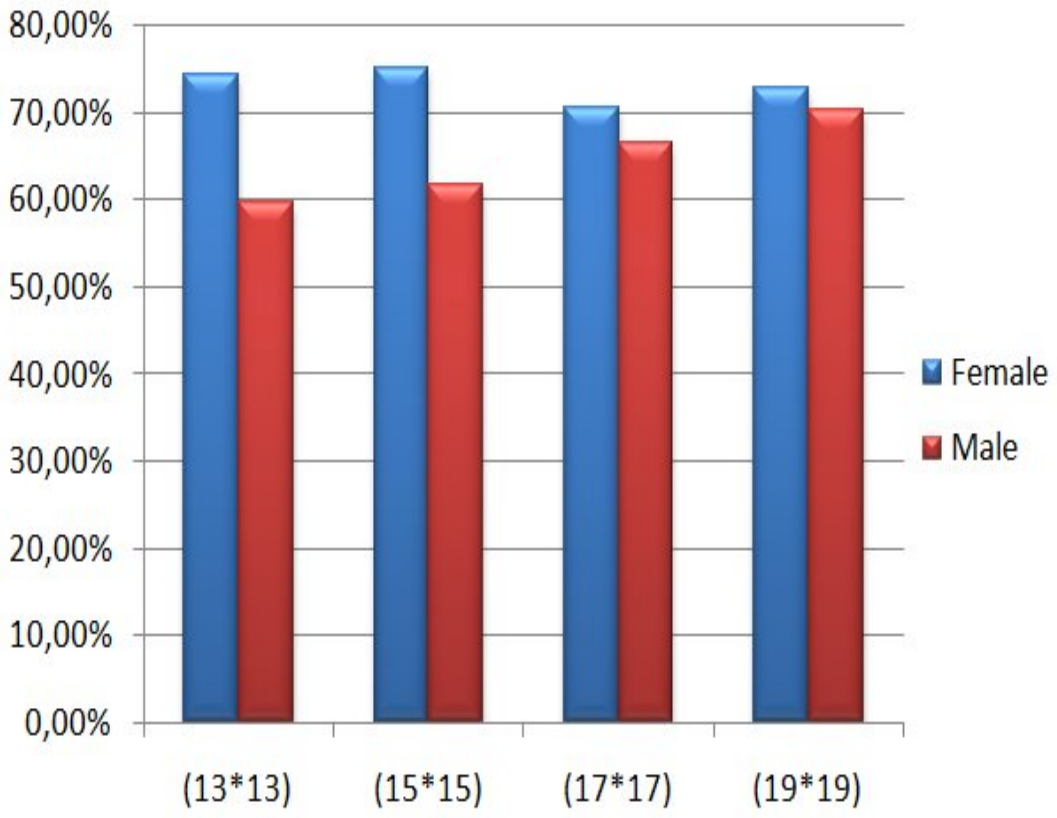


Figure 3.10 : Colonnes graphique des résultats obtenus dans le cas codebook size **200** (Anglais).

La Figure 3.11 même notion et explication par des colonnes graphiques qui examinent des résultats de classification Femelle et Male des évaluations sur les bases de données QUWI, obtenus dans le cas codebook taille **200** sur documents manuscrites en **arabe**, où l'on note que le sexe femelle du scripteur déterminé par **65.77 %** dans le découpage **13*13** par rapport aux autres découpages (15*15. 17*17. 19*19). D'autre part le sexe Male du scripteur déterminé par **70.37 %** dans le découpage **19*19** par rapport aux autres découpages (13*13. 15*15. 17*17).

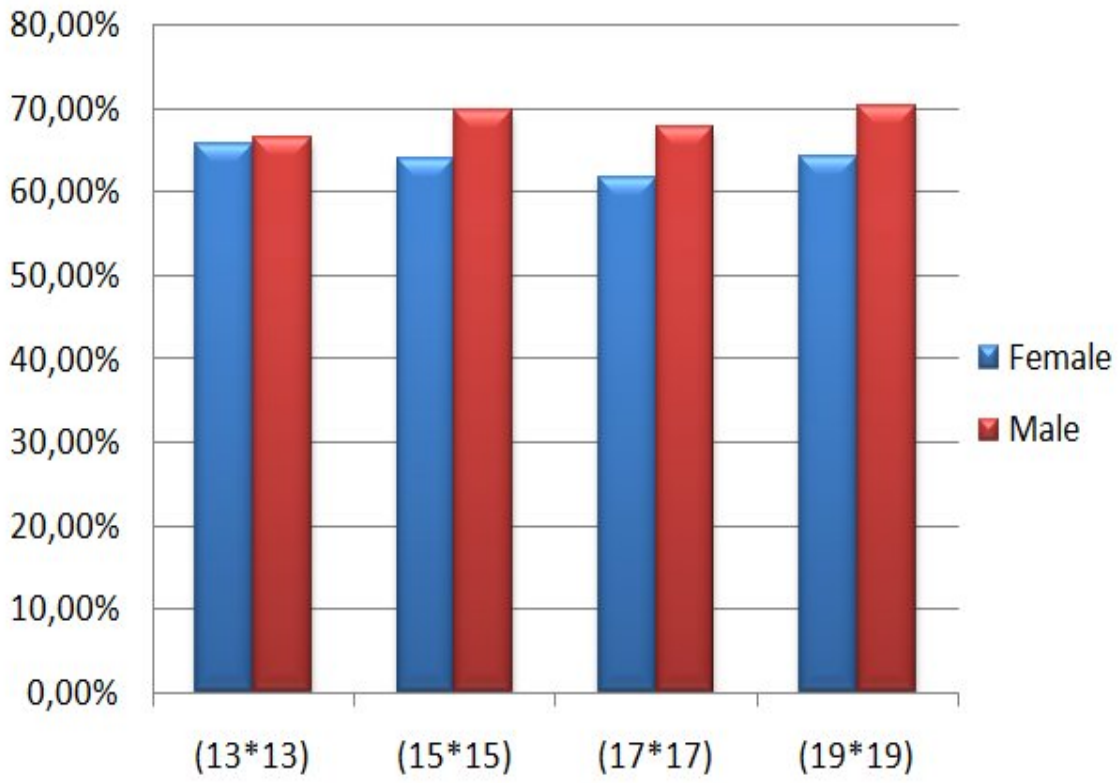


Figure 3.11 : Colonnes graphique des résultats obtenus dans le cas codebook size **200** (Arabe).

La Figure 3.12 démontre des colonnes graphiques qui examinent des résultats de classification Femelle et Male des évaluations sur les bases de données QUWI, obtenus dans le cas codebook taille **300** sur documents manuscrites en **anglais**, où l'on note que le sexe femelle du scripteur déterminé par **74.77 %** dans le découpage **13*13** par rapport aux autres découpages (15*15. 17*17. 19*19). D'autre part le sexe Male du scripteur déterminé par **69.75 %** dans le découpage **19*19** par rapport aux autres découpages (13*13. 15*15. 17*17).

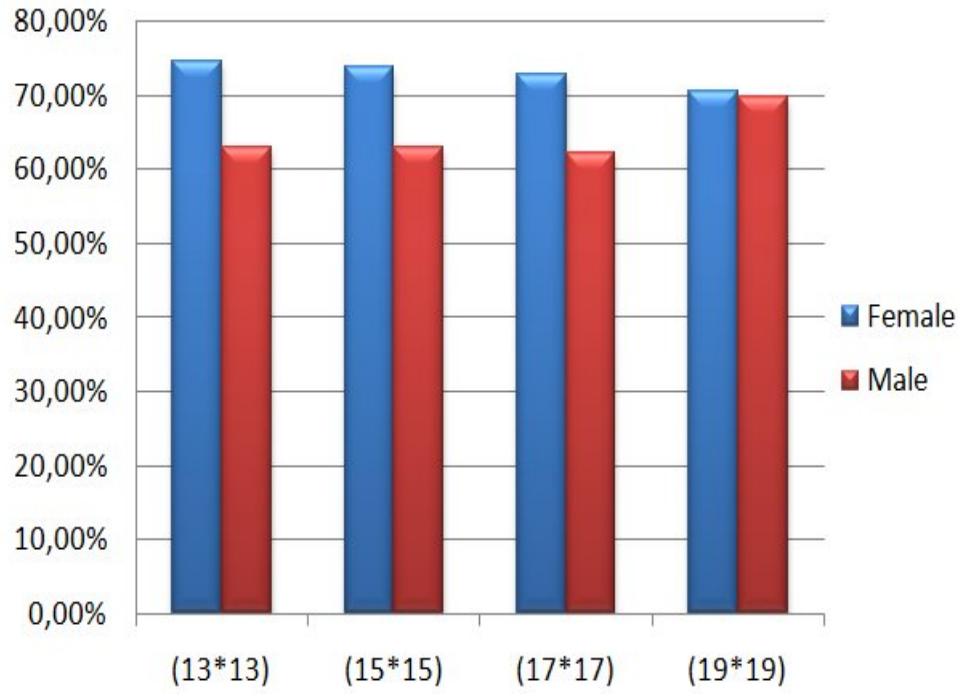


Figure 3.12 : Colonnes graphique des résultats obtenus dans le cas codebook size 3**00** (Anglais).

La Figure 3.13 représente des colonnes graphiques qui examinent des résultats de classification Femelle et Male des évaluations sur les bases de données QUWI, obtenus dans le cas codebook taille **300** sur documents manuscrites en **arabe**, où l'on note que le sexe femelle du scripteur déterminé par **67.57 %** dans le découpage **13*13** par rapport aux autres découpages (15*15. 17*17. 19*19). D'autre part le sexe Male du scripteur déterminé par **67.28 %** dans le découpage **19*19** par rapport aux autres découpages (13*13. 15*15.17*17).

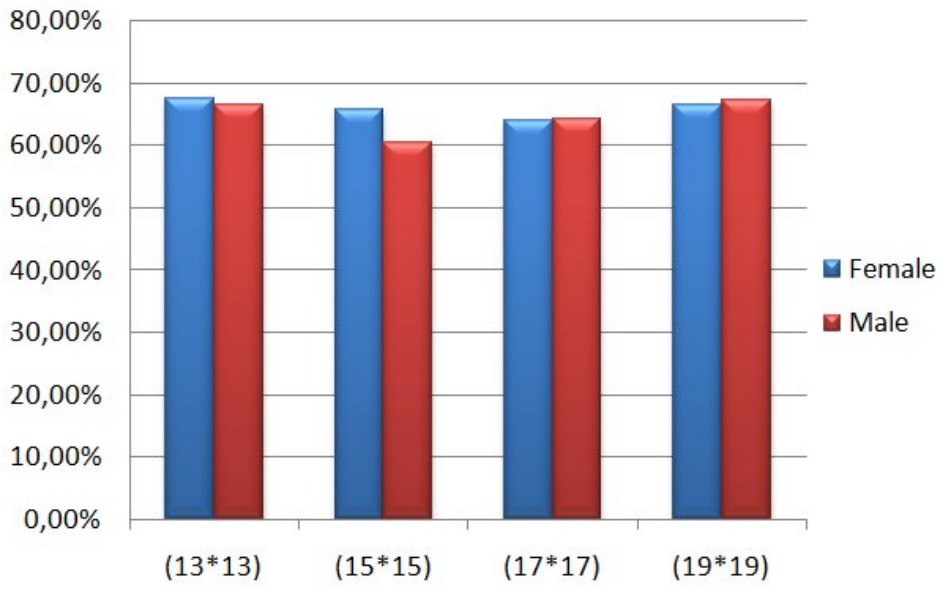


Figure 3.13 : Colonnes graphique des résultats obtenus dans le cas codebook size **300** (Arabe).

La Figure 3.14 en basant sur des colonnes graphiques qui examinent des résultats de classification Femelle et Male des évaluations sur les bases de données QUWI, obtenus dans le cas codebook taille **400** sur documents manuscrites en **anglais**, où l'on note que le sexe femelle du scripteur déterminé par **70.37 %** dans le découpage **15*15** par rapport aux autres découpages (13* 13. 17*17. 19*19). D'autre part le sexe Male du scripteur déterminé par **74.77 %** dans le découpage **17*17** par rapport aux autres découpages (13*13. 15*15.19*19).

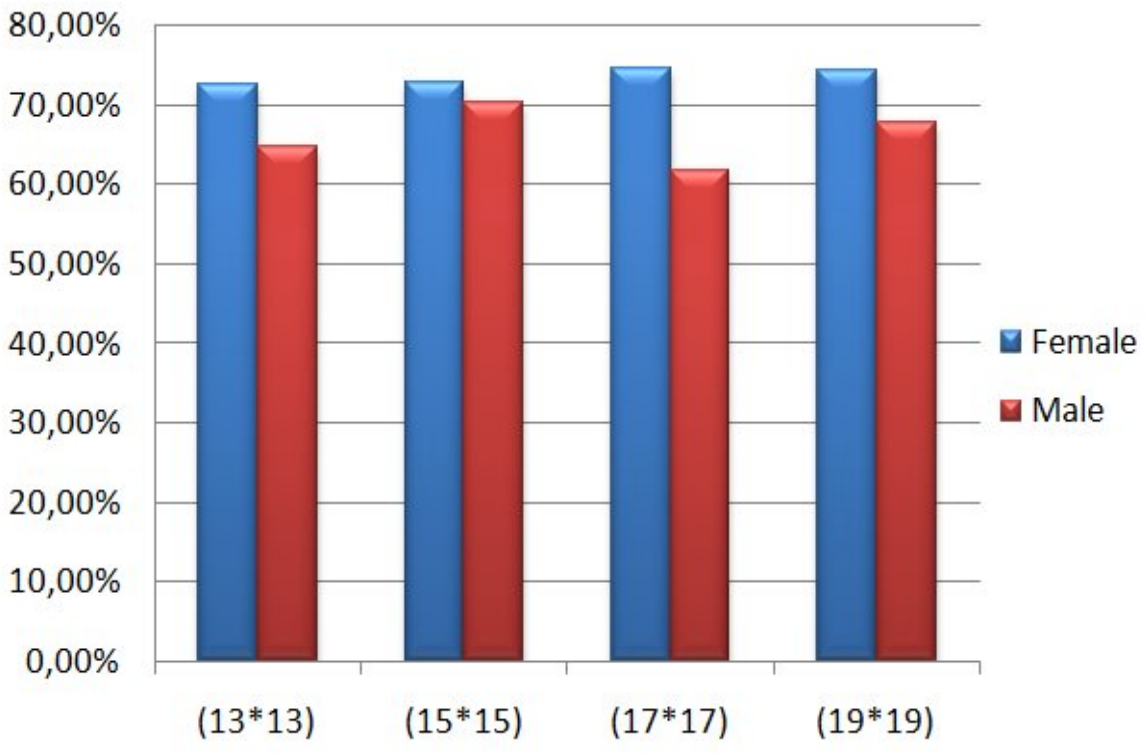


Figure 3.14 : Colonnes graphique des résultats obtenus dans le cas codebook size **400** (Anglais).

La Figure 3.15 représente des colonnes graphiques examinent des résultats de classification Femelle et Male des évaluations sur les bases de données QUWI, obtenus dans le cas codebook taille **400** sur documents manuscrites en **arabe**, où l'on note que le sexe femelle du scripteur déterminé par **69.14 %** dans le découpage **15*15** par rapport aux autres découpages (13*13. 17*17. 19*19). D'autre part le sexe Male du scripteur déterminé par **67.12 %** dans le découpage **19*19** par rapport aux autres découpages (13*13. 15*15. 17*17).

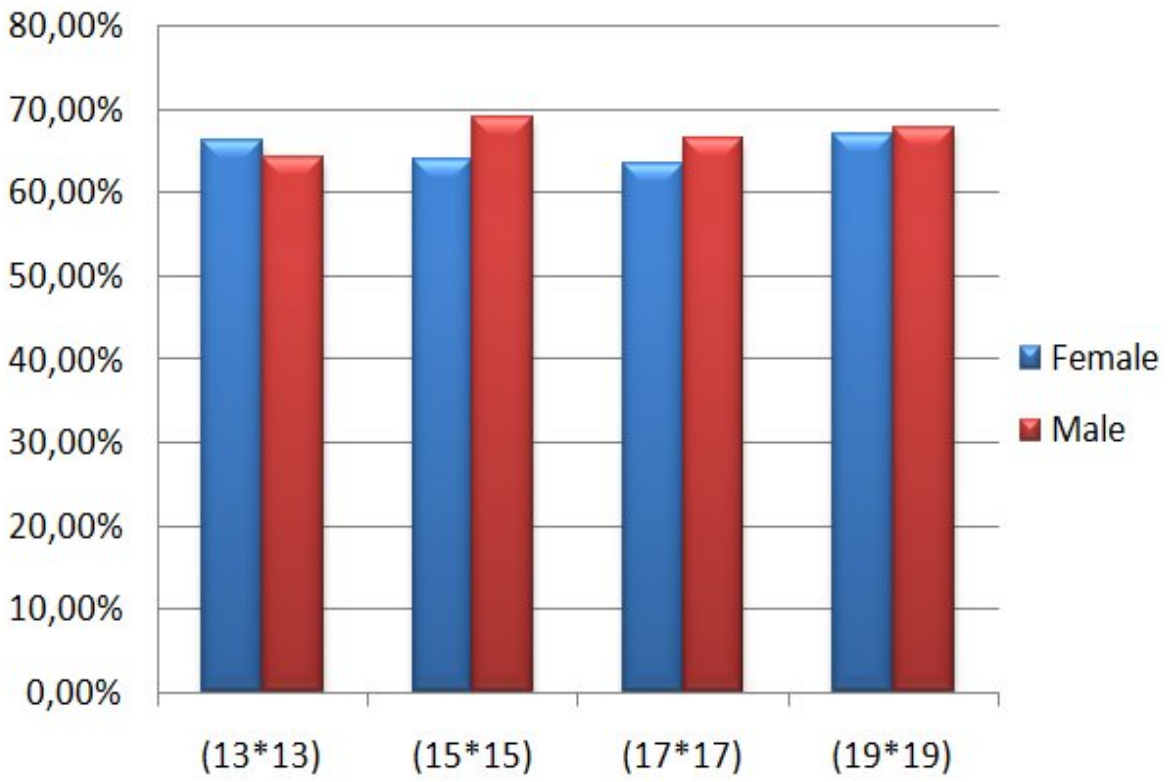


Figure 3.15 : Colonnes graphique des résultats obtenus dans le cas codebook size **400** (Arabe).

5. Conclusion

Nous avons présenté une méthode basés sur les codebooks pour la détermination de sexe d'un scripteur, qui est applicable pour les deux documents en arabe et en anglais. Cette étude est la d'abord que rapporté des résultats de classification sur l'ensemble de données QIWI, cette méthode pour la reconnaissance de sexe de scripteurs à partir de leur écriture manuscrite. la technique s'appuie sur l'extraction des formes pour un auteur spécifique, l'auteur emploie fréquemment des formes lorsqu'il dessine les caractères. les taux d'identification réalisés sont très encourageantes, Le changement de la taille de fenêtre pendant la phase de découpage de l'écriture et la combinaison de plusieurs tailles de fenêtres a amélioré les résultats significativement.

nous avons proposé un système de détermination de sexe d'un scripteur basé sur hors ligne texte manuscrit. cette méthode a été utilisé pour améliorer la détermination de sexe d'un scripteur . Des expériences ont été effectuées sur les échantillons extraits à partir de la base de données QUWI. Deux ensembles de données qui contiennent 1900 échantillons ont été sélectionnés. La comparaison avec les résultats obtenus en utilisant le même protocole expérimental, indiquent que les caractéristiques offrent une amélioration significative qui atteint 10%. En outre, les taux globale de classification qui est d'environ 71% et 68% pour les deux ensembles de données QUWI-Anglais et QUWI-Arabe, ainsi de taux de classification de sexe Male qui est 70.37% et 70.99% pour les deux ensemble de données QUWI-Anglais et QUWI-Arabe, et de taux de classification de sexe Female qui est 75.22% et 67.57% pour les deux ensemble de données QUWI-Anglais et QUWI-Arabe, respectivement. Enfin, nous en déduisons que caractéristiques de données locales constituent une approche prometteuse pour résoudre la tâche de détermination de sexe d'un scripteur.

Conclusion

CONCLUSION ET PERSPECTIVES

Ce travail présente une méthode pour la détermination de sexe d'un scripteur à partir de l'analyse automatique de son écriture manuscrite. La méthode proposée est basée sur l'extraction de petits fragments d'écriture, ces fragments sont extraits par un découpage adaptatif de l'écriture en imagettes de taille $n \times n$ pixels. Ces imagettes sont ensuite regroupées en utilisant l'algorithme de regroupement k-means pour avoir un codebook qui sera utilisé pour la caractérisation des écritures féminines et masculines. Les caractéristiques proposées ont la particularité de pouvoir décrire un échantillon d'écriture qu'il soit de grande ou de petite taille.

Les séparateurs à vaste marge (SVMs) ont été utilisés pour l'évaluation de la capacité des caractéristiques locales proposées à prédire le sexe d'un individu à partir de son écriture, et les résultats enregistrés sur un sous ensemble de la base de données QUWI constitué de 475 scripteurs semblent intéressants et prometteurs.

Comme perspectives d'avenir nous comptons développer l'approche proposée afin qu'elle soit appliquée à d'autres attributs de scripteurs tels que la latéralité manuelle (gaucher ou droitier), l'âge ou l'ethnie peuvent être envisagés. Il serait également intéressant d'introduire des caractéristiques supplémentaires et ensuite appliquer un mécanisme de sélection de caractéristiques pour savoir quelles sont les caractéristiques les plus discriminantes pour ce problème et pour des problèmes similaires.

Il est nécessaire de rappeler que la performance du système proposé ne dépend pas seulement de la technique de classification utilisée, mais aussi des caractéristiques choisies. Dans ce cadre, il serait très intéressant d'exploiter la combinaison de caractéristiques proposées cette étude avec celles de l'état de l'art afin d'améliorer les performances du système proposé. Nous pensons aussi que des études plus approfondies sur les stratégies de sélection de caractéristiques devraient être menées, afin de réduire la dimension de l'ensemble des caractéristiques proposées et ce pour déterminer quel sous-ensemble de caractéristiques est le plus discriminant dans la caractérisation des deux catégories de scripteurs.

Pour la méthode proposée et concernant les technique de classification utilisée, nous pensons qu'il serait intéressant d'envisager l'utilisation d'autres techniques de classification que celle que nous avons adoptée dans le présent mémoire. Il serait intéressant aussi d'envisager et d'expérimenter des possibilités de combinaison de techniques de classification.

Bibliographies

Bibliographie

[ADR 87] Adrian. F., Barrie. G., "Graphology and personality: Another failure to validate graphological analysis", In Personality and Individual Differences, Vol. 8, N°. 3, pp. 433 – 435, 1987.

[ALM 12] Al-Ma'adeed. S., Ayouby. W., Hassaine. A,, and Aljaam. J.M., "QUWI: *An Arabic and English handwriting data set for offline writer identification*". In: Proceedings of 13Th International Conference on Frontiers in Handwriting Recognition (ICFHR 2012), Bari, Italy, pp. 746 - 751, 2012.

[ARA 77] B. Arazi, "*Handwriting identification by means of run-length measurements*", IEEE Transactions on Systems, Man and Cybernetics, vol.SMC-7, no.12, pp. 878-881, Dec. 1977.

[ATA 11] Atanasiu.V, Likforman-Sulem.L, Vincent.N, Writer Retrieval - Exploration of a Novel Biometric Scenario Using Perceptual Feat ures Derived from Script Orientation , ICDAR, 2011

[BAR 09] I. Bar-Yosef, N. Hagbi, K. Kedem, and I. Dinstein. Line Segmentation for Degraded Hand-written Historical Documents. In 2009 10th *International Conference on Document Analysis and Recognition* (ICDAR), pages 1161–1165, 2009

[BAY 06] H. Bay, T. Tuytelaars, and L. Van Gool. SURF: Speeded Up Robust Features. In 20 9th European Conference on Computer Vision (ECCV), volume 3951 of Lecture Notes in Computer Science, pages 404–417. Springer Berlin Heidelberg, 2006.

[BEE 05] Beech. J and Mackintosh. I., "Do differences in sex hormones affect handwriting style? evidence from digit ratio and sex role identity as determinants of the sex of handwriting". In: Personality and Individual Differences, Vol. 39, N°. 2, pp. 459 - 468, 2005.

[BEN 02] A. Bensefia, A. Nosary, T. Paquet, L. Heutte, "*Writer identification by writer's invariants*" , Eight International Workshop on Frontiers in Handwriting Recognition (IWFHR'02), Niagara-on-the-Lake, Canada, August 6-8; pp. 274-279, 2002.

[BEN 05] Bensefia, A., Paquet, T., & Heutte , L. (2005b). A writer identification verification system. Pattern Recognition Letters, 26(13), 2080–2092.

[BOU 95] Boulétreau, V., Vincent, N., & Emptoz, H. (1995). A writing qualification invariant towards line thickness and resolution changings. In ACCV'95:In Proceedings of th Asian Conference on Computer Vision (pp. 325–329).

[BOU 97] Boulétreau, V. (1997). Vers un classement de l'écrit par des méthodes fractales. PhD thesis.

[BOU 98] Boulétreau, V., Vincent, N., Sabourin, R., & Emptoz, H. (1998). Handwriting and signature: One or two personality identifiers? In ICPR '98: Proceedings of the 14th International Conference on Pattern Recognition-Volume 2 (pp. 1758–1760). Washington, DC, USA: IEEE Computer Society.

[BOU 08] Boulehmi.H, Seddik.B, Kricha.A, Ben Amara.N, Prétraitement de documents anciens, CIFED, 2008.

[BUL 05] Bulacu M., Schomaker. L., "A Comparison of Clustering Methods for Writer Identification and Verification", In: Proceedings of the 8-Th International Conference on Document Analysis and Recognition (ICDAR 2005), Seoul, South Korea, pp. 1275 – 1279, 2005.

[BUL 07] Bulacu. M., Schomaker. L., "Text-independent writer identification and verification using textural and allographic features", In: IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 29, N°. 4, pp. 701 - 717.

[BUR 02] Burr. V., "Judging gender from samples of adult handwriting: accuracy and use of cues", In: Journal of Social Psychology, Vol. 142, N°. 6, pp. 691 - 700, 2002.

[CAM 06] W.M. Campbell, D.E. Sturim, and D.A. Reynolds. Support vector machines using GMM supervectors for speaker verification. IEEE Signal Processing Letters, 13(5):308–311, May 2006.

[CAN 05] Canu. S., Grandvalet. Y., Guigue. V., Rakotomamonjy. A., "SVM and Kernel Methods Matlab Toolbox", Perception Systèmes et Information, INSA de Rouen, Rouen, France, 2005.

[CHR 14] V. Christlein, D. Bernecker, F. Honig, and E. Angelopoulou. Writer identification and verification using GMM supervectors. In 2014 IEEE Winter Conference on Applications of Computer Vision (WACV), pages 998–1005, March 2014.

[CLA 01] E. Clavier, *"Etude des stratégies de tri : application à un système de tri de formulaires"*, thèse de l'Université de Caen, Soutenue le 21 décembre 2001.

[CRE 95] J.-P. Crettez, *"A set of handwriting families : style recognition"*, International Conference on Document Analysis and Recognition (ICDAR'95), Montréal, Canada, pp. 489-494, 1995.

[DHN 14] B.V. Dhandra and M.B. Vijayalaxmi. Text and script independent writer identification. In 2014 International Conference on Contemporary Computing and Informatics (IC3I), pages 586–590, Nov 2014.

[DJE 10] Djeddi. C., Souici-Meslati. L., "A texture based approach for Arabic writer identification and verification", In: Proceedings of the International Conference on Machine and Web Intelligence (ICMWI 2010), Algiers, Algeria, pp. 115 – 120, 2010.

[DJE 12] C. Djeddi, L. Souici-Meslati, and A. Ennaji. Writer Recognition on Arabic Handwritten Documents. In Image and Signal Processing, volume 7340 of Lecture Notes in Computer Science, pages 493–501. Springer Berlin Heidelberg, 2012.

[EFR 89] Efrat. N., Gershon. B.S., "The predictive validity of graphological inferences: A metaanalytic approach", Personality and Individual Differences, Vol. 10, N°. 7, pp. 737 - 745, 1989.

[FRI 99] Friedman, M. & Kandel, A. (1999). Introduction to Pattern Recognition : Statistical, Structural, Neural and Fuzzy Logic Approaches. World Scientific Publishing Company.

[GIL 94] M. Gilloux, "Writer *adaptation for handwritten word recognition using hidden Markov models*", 12th International Conference on Pattern Recognition (IAPR), Los Alamitos, USA, pp. 135-139 vol.2, 1994.

[GOO 45] Goodenough. F.L., "Sex differences in judging the sex of handwriting", In: Journal of Social Psychology, Vol. 22, pp. 61 - 68, 1945.

[HAS 13] Hassaïne. A., Al Maadeed. S., Aljaam. J., Jaoua. A., "*ICDAR2013 - Competition on Gender Prediction from Handwriting*", In: Proceedings of the 12th International Conference on Document Analysis and Recognition (ICDAR 2013), Washington, USA, pp. 1449 – 1453, 2013.

[HAM 96] Hamid. S., Loewenthal. K.M., ''Inferring gender from handwriting in urdu and english". In: Journal of Social Psychology, Vol. 136, N°. 6, pp. 778 - 782, 1996.

[HEU 00] L. Heutte, T. Paquet, A. Nosary, C. Hernoux, *"Handwritten text recognition using a multiple-agent architecture to adapt the recognition task"*, 7th International Workshop on Frontiers in Handwriting Recognition (IWFHR VII), Amsterdam, pp. 413-422, 2000.

[HIR 10] P.S. Hiremath, S. Shivashankar, J.D. Pujari, and R.K. Kartik. Writer identification in a handwritten document image using texture features. In International Conference on Signal and Image Processing (ICSIP), pages 139 –142, Dec 2010.

[JAI 13] R. Jain and D. Doermann. Writer Identification Using an Alphabet of Contour Gradient Descriptors. In 2013 12th International Conference on Document Analysis and Recognition (ICDAR), pages 550–554, Aug 2013.

[JAI 14] R. Jain and D. Doermann. Combining Local Features for Offline Writer Identification. In 2014 14th International Conference on Frontiers in Handwriting Recognition (ICFHR), pages 583–588, Sept 2014.

[JAM 91] James. H., "Sex differences in handwriting: A comment on spear", In: British Educational Research Journal, Vol. 17, N°. 2, pp. 141 - 145, 1991

[KAR 05] Karthik. R.B., Srihari. S.N., "Writer demographic classification using bagging and boosting", In: Proceedings of the 12th International Graphonomics Society Conference (IGS 2005), Salerno, Italy, pp. 133 - 137, 2005.

[KEB 98] S. Kebairi, B. Taconet, A.Zahour, S. Ramdane, *"A Statistical Method For an Automatic Detection of Form types"*, Proceedings of the (DAS'98), Nagano, Japan, November 4-6, pp.109-118, 1998.

[KHA96] K. Khan, I.K. Sethi, *"Handwritten signature retrieval and identification"*, Pattern Recognition Letters, vol.17, pp. 83-90, 1996.

[KLE 80] V. Klement, R.-D. Naske, K. Steinke, "The application of image processing and pattern recognition techniques to the forensic analysis of handwriting", Proceedings of the Third International Conference Security Through Science and Engineering, Lexington, Kentucky, USA, pp. 5-11, 1980.

[KLE 81] V. Klement, "*Forensic writer recognition*" , Digital Image Processing. Proceedings of the NATO Advanced Study Institute, Dordrecht, Netherlands Reidel, pp. 519-524, 1981.

[KLE 83] V. Klement, "*An Application System for the Computer-Assisted Identification of Handwritings*", Proceedings of the International Carnahan Conference on Security Technology, Lexington, Kentucky, USA, pp. 75-79, 1983.

[KLE 13] F. Kleber, S. Fiel, M. Diem, and R. Sablatnig. CVL-DataBase: An Off-Line Database for Writer Retrieval, Writer Identification and Word Spotting. In 2013 12th International Conference on Document Analysis and Recognition (ICDAR), pages 560–564, 2013

[KUC 79] W. Kuckuck, B. Rieger, K. Steinke, "*Automatico writer recognition*", Proceedings of the 1979 Carnahan Conference on Crime Countermeasures, Lexington, Kentucky, USA, pp. 57-64, 1979.

[KUC 80] W. Kuckuck, "*Writer recognition by spectral analysis*" , Proceedings of the Third International Conference Security Through Science and Engineering, Lexington, Kentucky, USA, pp. 1-3, 1980.

[LIW 06] Liwicki. M., Schlapbach. A., B. Horst., B. Samy., M. Johnny, R. Jonas., "Writer Identification for Smart Meeting Room Systems", In: Proceedings of 7th IAPR Workshop on Document Analysis Systems (DAS 2006), Nelson, New Zealand, pp. 186-195, 2006.

[LIW 11] Liwicki. M., Schlapbach. A., Bunke. H., "Automatic gender detection using on-line and off-line information", In: Pattern Analysis and Applications Journal, vol. 14, N°. 1, pp. 87 – 92, 2011.

[LOU 11] G. Louloudis, N. Stamatopoulos, and B. Gatos. ICDAR 2011 Writer Identification Contest. In 2011 11th International Conference on Document Analysis and Recognition (ICDAR), pages 1475–1479, Sept 2011

[LOU 12] G. Louloudis, B. Gatos, and N. Stamatopoulos. ICFHR2012 Competition on Writer Identification, Challenge 1: Latin/Greek Documents. In 2012 International Conference on Frontiers in Handwriting Recognition, pages 825–830, 2012.

[MAN 75] Mandelbrot, B. (1975). Les objets fractals. Flammarion.

[MAR 01] U.-V. Marti, R. Messerli, H. Bunke, "*Writer identification using text line based features*", Proceedings. 6th International Conference on Document Analysis and Recognition (ICDAR), Seattle, USA, pp. 101-105, 2001.

[MAR 02] U.-V. Marti and H. Bunke. The IAM-database: an English sentence database for offline handwriting recognition. International Journal on Document Analysis and Recognition, 5(1):39–46, 2002

[MIH 77] F. Mihelic, N. Pavesic, L. Gyergyek, "*Recognition of writers of handwritten texts*", International Conference On Crime Countermeasures, pp. 237-240,1977.

[MIK 13] H. Miklas. Zur editorischen Vorbereitung des sog. Missale Sinaiticum (Sin. slav. 5/N). InH. Miklas, V. Sadovski, and S. Richter, editors, Glagolitica - Zum Ursprung der slavis-chen Schriftkultur, volume XV-XVI, pages 117–129. (OAW, Phil.-hist. Kl., Schriften derBalkan-Kommission, Philologische Abt. 41), 2000

[NAS 80] R.-D. Naske, "*Application of a weighted least squares algorithm to writer and speaker recognition*" , Proceedings of the 5th International Conference on Pattern Recognition (ICPR), New York, USA, pp. 27-30, 1980.

[NAS 82] R.-D. Naske, "*Writer recognition by prototype related deformatio n of handprinted characters*" , Proceedings of the 6th International Conference on Pattern Recognition (ICPR), New York, USA, vol.2, pp. 819-22, 1982.

[NOS 99] A. Nosary, L. Heutte, T. Paquet, Y. Lecourtier, *"Defining writer's invariants to adapt the recognition task"*, International Conference on Document Analysis and Recognition (ICDAR'99), Bangalore (India), pp. 765-768, 1999.

[NOS 02] Nosary, A., Paquet, T., Heutte, L., & Bensefia, A. (2002). Handwritten text recognition through writer adaptation. In IWFHR '02: Proceedings of the Eighth International Workshop on Frontiers in Handwriting Recognition.

[PAR 06] Pareti, R. & Vincent, N. (2006). Global method based on pattern occurrences for writer identification. In IWFHR'06: Proceedings of the 10th International Workshop on Frontiers in Handwriting Recognition.

[PLA 89] Plamondon. R., Lorette. G., "Automatic signature verification and writer identification– the state of the art", In: Pattern Recognition Journal, vol. 22, pp. 107 – 131, 1989.

[PEA 97] Peake, G. S. & Tan, T. N. (1997). Script and language identification from document images. In BMVC'97:In Proceedings of British Machine Vision Conference, volume 2 (pp. 610–619).

[RIS 83] Richard J.K., Anat. R., "Inferring personal qualities through handwriting analysis", In: Journal of Occupational Psychology, Vol. 56, N°. 3, pp. 191 - 202, 1983.

[ROB 97] Robert. P.T., Cynthia. A.P., "The validity of handwriting elements in relation to self report personality trait measures", In: Personality and Individual Differences, Vol. 22, N°. 1, pp. 11 - 18, 1997.

[ROY 00] Roy. N.K., Derek. J.K., "Illusory correlations in graphological inference", In: Journal of Experimental Psychology Applied, Vol. 6, N° 4, pp. 336 - 348, 2000.

[ROY 99] Roy. A.H., Headrick. A.M., "Handwriting Identification: Facts and Fundamentals", CRC Press, 1999.

[SOK 12] Sokic. E., Salihbegovic. A., Ahic-Djokic. M., "Analysis of off-line handwritten text samples of different gender using shape descriptors". In: Proceedings of the 9th International Symposium on Telecommunications (BIHTEL), pp. 1 - 6, 2012.

[SAI 00] Said, H. E. S., Tan, T. N., & Baker, K. D. (2000). Personal identification based on handwriting. Pattern Recognition, 33, 149–160.

[SER02] A. Seropian, N. Vincent, "Writers Authentication and fractal Compression" , 8th International Workshop on Frontiers in Handwriting Recognition (IWFHR), Niagara-on-the-Lake (Canada), pp. 434-439, 2002

[SHA 94] Shackleton. V., Newel. S., "European management selection methods: A comparison of five countries". In: International Journal of Selection and Assessment, Vol. 2, N°. 2, pp. 91 - 102, 1994.

[SID 07] Siddiqi, I. & Vincent, N. (2007). Writer identification in handwritten documents. In ICDAR '07: Proceedings of the Ninth International Conference on Document Analysis and Recognition (ICDAR 2007), volume 1 (pp. 108–112).: IEEE Computer Society

[SID 08] Siddiqi. I., Vincent. N., ''Combining Global and Local Features for Writer Identification'', In: Proceedings of the 11th International Conference on Frontiers in Handwriting Recognition (ICFHR 2008), Montréal, Québec, pp. 48 – 53, 2008.

[SID 10] Siddiqi. I., Vincent. N., "Text independent writer recognition using redundant writing patterns with contour-based orientation and curvature features", In: Pattern Recognition Journal, vol. 43, N°. 11, pp. 3853 – 3865, 2010.

[SRI 02] Srihari. S., Arora. H., Cha. S.H., Lee. S., "Individuality of handwriting", In: Journal of Forensic Sciences, Vol. 47, N°. 4, pp. 1 - 17, 2002.

[SUN 01] Sung-Hyuk. C., Srihari. S., "A priori algorithm for sub category classification analysis of handwriting". In: Proceedings of the 6th International Conference on Document Analysis and Recognition (ICDAR 2001), Seattle, WA, USA, pp. 1022 - 1025, 2001.

[STE 09] Stefan Fiel. Novel Methods for Writer Identification and Retrieval pages 14–16, nov 2015.

[VIN 95] Vincent, N. & Emptoz, H. (1995). A classification of writings based on fractals. Fractal Reviews in the Natural and Applied Sciences, (pp. 320–331).

[WIL 96] William N.H., "Identifying sex from handwriting". In: Perceptual and Motor Skills, Vol. 83, pp. 91 - 800, 1996.

[VIN98] N. Vincent, S. Barbezieux, *"Compression of handwriting images : a way to define a writing style"*, Vision Interface, pp. 347-354, 1998.

[ZIP 49] Zipf, G. (1949). Human Behavior and the Principle of Least Effort. Addison-Wesley

[ZOI 00] Zois, E. N. & Anastassopoulos, V. (2000). Morphological waveform coding for writer identification. Pattern Recognition, 33(3), 385–398.

Résumé Le travail présenté dans ce manuscrit se situe dans le domaine de l'analyse et la reconnaissance de documents, et plus précisément, la reconnaissance hors-ligne du genre de scripteur à partir de leur écriture manuscrite. La méthode proposée est basée sur l'extraction d'un ensemble de caractéristiques de l'écriture à partir des échantillons de scripteurs de sexe masculin et de féminin, et l'entraînement d'un classifieur afin qu'il puisse distinguer entre les deux catégories. La classification est faite en utilisant les SVMs et cela en utilisant une base de données contenant des documents écrits pour 475 scripteurs avec 4 échantillons de chaque personne qu'elle a fait la base à notre humble avis quelques résultats intéressants.

Mot clés : Identification de scripteurs, Détermination du sexe, classifieur SVM.

Abstract The work presented in this manuscript can be placed within the field of document analysis and recognition, and more precisely, the off-line recognition of individuals and their gender from their handwriting. a study to predict gender of individuals from their handwriting. The proposed method is based on extracting a set of features from writing samples of male and female writers and training classifiers for learning to discriminate between the two categories. The classification is made using SVMs and this using a database containing 475 documents written for writers with 4 samples from each person that made the basis in our humble opinion some interesting results.

Keywords: Writer recognition, gender determination, SVM classifier.

الملخص

إن العمل المقدم في هذه المخطوطة يندرج في إطار مجال التحليل والتعرف على الوثائق، وعلى نحو أدق، تحديد جنس الكاتب حسب الوثائق المكتوبة بخط اليد. و تعتمد الطريقة المقترحة على استخراج مجموعة من الميزات من عينات لوثائق أشخاص ذكور و إناث و يتم ذلك بتدريب المصنفات لتكون قادرة على التمييز بين الفئتين. تمت عملية التصنيف باستخدام أجهزة المتجهات الاعتمادية. و هذا باستخدام قاعدة بيانات تحتوي على وثائق مكتوبة لـ 475 كاتب مع 4 عينات لكل شخص ، حيث حققت في عملنا المتواضع نتائج مثيرة للاهتمام.

الكلمات المفتاحية : تحديد هوية صاحب الخط، تحديد جنس صاحب الخط، أجهزة المتجهات الاعتمادية.