

وزارة التعليم العالي و البحث العلمي

BADJI MOKHTAR-ANNABA UNIVERSITY
UNIVERSITE BADJI MOKHTAR-ANNABA



جامعة باجي مختار- عنابة

Faculté des Sciences de L'Ingéniorat Année 2013-2014

Département d'Informatique

THESE

Présenté en vue de l'obtention
du diplôme de Doctorat en Sciences

*Contribution à l'analyse et la caractérisation
de l'écriture manuscrite*

Option
Intelligence Artificielle

par
Mr DJEDDI Chawki

Directeur de Thèse

Mme Labiba SOUICI-MESLATI

Professeur Université Badji Mokhtar-Annaba

DEVANT LE JURY

Président :

Mr Nadir FARAH

Professeur Université Badji Mokhtar-Annaba

Examineurs :

Mme Halima BAHY-ABIDET

Professeur Université Badji Mokhtar-Annaba

Mr Mohamed BENMOHAMMED

Professeur Université de Constantine

Mr Salah CHENIKHER

Professeur Université de Tébessa

Mr Hamid SERIDI

Professeur Université 8 Mai 1945- Guelma

DEDICACES

A ceux que j'aime ...et ceux qui m'aiment

REMERCIEMENTS

Ce fût vraiment un très grand plaisir d'avoir été encadré pour la troisième fois, par Mme Labiba Souci-Labiba, Professeur à l'Université Badji-Mokhtar d'Annaba. Qu'elle trouve ici l'expression de ma plus profonde reconnaissance et gratitude pour son aide précieuse. Je la remercie pour tout le temps qu'elle m'a consacré, ses encouragements et ses conseils constructifs. Elle a contribué à ce travail de recherche de façon incommensurable. Sa rigueur intellectuelle a permis au projet de suivre son chemin. Je la remercie également pour la confiance qu'elle a su m'accorder jusqu'à la dernière minute.

Je remercie vivement Mr Nadir Farah, Professeur à l'université Badji-Mokhtar d'Annaba pour m'avoir fait l'honneur de présider mon jury de thèse. Je remercie également Mr Mohamed Benmohammed, Professeur à l'Université de Constantine, Mme Halima Bahi, Professeur à l'Université Badji-Mokhtar d'Annaba, Mr Hamid Seridi, Professeur à l'Université de Guelma et Mr Salah Chenikher, Professeur à l'Université de Tébessa, pour l'intérêt qu'ils ont bien voulu porter à mon travail en acceptant la lourde tâche d'examineur.

J'adresse toute ma reconnaissance à Mr Abdellatif Ennaji, Maître de Conférences au Laboratoire LITIS de l'université de Rouen, pour avoir été à l'origine des stages enrichissants effectués au laboratoire LITIS de l'université de Rouen, pour son aide, sa gentillesse et sa disponibilité.

Un remerciement spécial est aussi adressé aux membres du laboratoire LITIS de l'université de Rouen, en particulier Pr Thierry Paquet pour avoir ponctué mon travail de discussions fructueuses.

Je remercie très sincèrement Mr Imran Siddiqi, Docteur à l'université Bharia au Pakistan, pour son extrême gentillesse, son aide et ses encouragements durant toute la période de préparation de cette thèse.

Un grand merci pour ma femme qui a dû supporter mon désordre, toujours présente dans les moments difficiles et aussi a mon petit garçon Haider ainsi qu'à mes deux filles Yara et Lina qui m'ont assisté par leurs sourires.

Je réserve mes derniers remerciements à toute ma famille, particulièrement à mes parents qui longtemps attendu ce moment afin de cueillir le fruit pour lequel ils ont consenti, pendant des années, tant de dévouement et de sacrifices.

Chawki

TABLE DES MATIERES

Introduction générale	1
1. Contexte de la recherche	1
2. Objectifs de la recherche	4
3. Organisation du mémoire de thèse	6
Chapitre 1. Analyse de l'écriture manuscrite pour la classification de scripteurs: Concepts et outils	8
1.1. L'écriture manuscrite: un moyen de caractérisation des individus	8
1.1.1. Facteurs causant la variabilité des écritures manuscrites	9
1.1.2. Différents types de variations des écritures manuscrites	11
1.2. Catégorisation des systèmes de classification de scripteurs	11
1.2.1. Catégorisation par Tâche	12
1.2.2. Catégorisation par dépendance du texte	14
1.2.3. Catégorisation selon le mode d'acquisition des données	14
1.3. Evaluation de performances des systèmes de classification des scripteurs	16
1.3.1. Evaluation de l'identification	16
1.3.2. Evaluation de la vérification	16
1.4. Les applications possibles de la classification des scripteurs	18
1.4.1. Systèmes adaptatifs de reconnaissance de l'écriture manuscrite	18
1.4.2. Examen forensique de documents	18
1.4.3. Reconnaissance biométrique	19
1.4.4. Bibliothèque numériques	19
1.4.5. Salles de réunions intelligentes	19
1.4.6. L'intelligence ambiante	20
1.5. Systèmes existants de la classification des scripteurs	20
1.5.1. Le système WANDA	20
1.5.2. Le système CEDAR-FOX	21
1.5.3. Le système FLASH ID	22
1.5.4. Le système SCRIPT	23
1.6. Principales bases de données	24
1.6.1. CEDAR Database	24
1.6.2. IAM Database	24
1.6.3. RIMES Database	24
1.6.4. BFL Database	25
1.6.5. CVL Database	25
1.6.6. KHATT Database	25
1.6.7. IFN/ENIT Database	26
1.6.8. QUWI Database	26
1.6.9. CASIA Database	26
1.6.10. CVCMUSCIMA Database	26
1.6.11. FIREMAKER Database	27
1.6.12. CEDARABIC Database	27
1.7. Conclusion	27
Chapitre 2. Approches de classification de scripteurs: Un état de l'art	29
2.1. Introduction	29
2.2. Analyse locale pour la classification de scripteurs	30
2.2.1. Travaux d'Audrey Seropian (2003)	30
2.2.2. Travaux de Bensefia et al (2004)	31
2.2.3. Travaux de Pervouchine & al (2007)	32
2.2.4. Travaux de Tan & al (2009)	34
2.2.5. Travaux de Woodard & al (2010)	35
2.2.6. Travaux de Jain & al (2011)	36
2.3. Analyse globale pour la classification de scripteurs	37
2.3.1. Travaux de Zois & al (2000)	37
2.3.2. Travaux de Said & al (2000)	39
2.3.3. Travaux de Marti et al (2001)	40

2.3.4. Travaux de Srihari & al (2002)	41
2.3.5. Travaux de Gazzah & al (2005, 2006)	43
2.3.6. Travaux de Feddaoui & al (2006)	44
2.3.7. Travaux de Schlapbach & al (2007)	47
2.3.8. Travaux de Bertolini & al (2013)	48
2.3.9. Travaux de Chabouni & al (2014)	49
2.3.10. Travaux de Wang et al (2003 & 2004)	50
2.3.11. Travaux de Garain et al (2009)	50
2.3.12. Autres travaux	51
2.4. Analyse combinée pour la classification de scripteurs	52
2.4.1. Travaux de Bulacu & al (2007)	52
2.4.2. Travaux de Siddiqi et al (2010)	54
2.5. Compétitions sur la classification de scripteurs	55
2.5.1. Compétitions sur l'identification de scripteurs	56
2.5.1.1. Compétitions sur l'identification de scripteurs à partir de documents arabes	56
2.5.1.2. Compétitions sur l'identification de scripteurs à partir de documents multilingues (latins et grecs)	57
2.5.1.3. Compétition sur l'identification de scripteurs à partir de partitions musicales	58
2.5.1.4. Compétition sur l'identification de scripteurs à partir des écritures de styles différents issues du même scripteur	58
2.5.2. Compétition sur la prédiction du genre	59
2.6. Synthèse de travaux de classification de scripteurs	60
Chapitre 3. Reconnaissance du scripteur en mode indépendant du texte en utilisant des documents multi-scripts	64
3.1. Introduction	64
3.2. Base de données utilisée	66
3.2.1. Description de la base de données	66
3.2.2. Similitudes et différences entre les scripts latins et grecs	66
3.3. Extraction de caractéristiques	67
3.3.1. Distribution de longueurs de segments (Run-length)	67
3.3.2. Caractéristiques de l'état de l'art	72
3.4. Indépendance des caractéristiques proposées du script considéré	72
3.4.1 Expérimentations sur la base IFN-ENIT	73
3.4.2 Expérimentations sur la base KHATT	73
3.4.3 Expérimentations sur la base CVL	73
3.4.4 Expérimentations sur les bases des compétitions ICDAR et ICFHR contenant des documents latins et grecs	74
3.5. Reconnaissance du scripteur	74
3.5.1 Identification du scripteur	74
3.5.1.1. k Plus Proche Voisins (k-NN)	75
3.5.1.2. Séparateurs à vaste marge multi-classes (SVM)	75
3.5.2. Vérification du scripteur	75
3.6. Résultats expérimentaux et discussion	76
3.6.1. Expérimentations dépendantes de la langue	76
3.6.2. Expérimentations Multi-scripts (Grec vs Anglais et inversement)	78
3.6.3. Stabilité des caractéristiques proposées	79
3.6.3.1. Performances en fonction de la quantité du texte	79
3.6.3.2. Influence du bruit sur les performances des caractéristiques proposées	81
3.7. Conclusion	83
Chapitre 4. Analyse automatique de l'écriture manuscrite pour la détermination du sexe d'un individu	84
4.1. Introduction	84
4.2. Différences de sexe dans l'écriture manuscrite	85
4.3. Bases de données	86
4.4. Extraction de caractéristiques	88
4.4.1. Orientation et courbure	88
4.4.1.1. Caractéristiques basées sur les chaînes de codes de Freeman	88
4.4.1.2. Caractéristiques basées sur les polygones	90
4.4.2. Caractéristiques fractales	91
4.4.3. Caractéristiques texturales	92
4.4.3.1. Motifs binaires locaux (LBP)	92
4.4.3.2. Coefficients autorégressifs	99
4.5. Classification	94

4.6. Résultats expérimentaux	94
4.6.1. Evaluations en modes dépendant et indépendant du texte	95
4.6.2. Evaluations en modes dépendant et indépendant du script	97
4.6.3. Evaluations inter-bases de données	98
4.7. Conclusion	99
Conclusion et Perspectives	100
Bibliographie	102
A.1. The ICDAR 2011 Arabic Writer Identification Contest	115
A.1.1. Description de la base de données utilisée	116
A.1.2. Evaluation et résultats	116
A.2. The ICDAR 2011 Music Scores Competition: Staff Removal and Writer Identification	116
A.2.1. Description de la base de données utilisée	117
A.2.2. Evaluation et résultats	117
A.3. ICDAR 2011 Writer Identification Contest	117
A.3.1. Description de la base de données utilisée	118
A.3.2. Evaluation et résultats	118
A.4 ICFHR 2012 Competition on Writer Identification Challenge 1: Latin/Greek Documents	122
A.4.1. Description de la base de données utilisée	123
A.4.2. Evaluation et résultats	123
A.5. ICFHR 2012 Competition on Writer Identification Challenge 2: Arabic Scripts	124
A.5.1. Description de la base de données utilisée	124
A.5.2. Evaluation et résultats	125
A.6. ICDAR 2013 Competition on Writer Identification	125
A.6.1. Description de la base de données utilisée	125
A.6.2. Evaluation et résultats	126
A.7. ICDAR 2013 Competitions on Signature Verification and Writer Identification for On- and Offline Skilled Forgeries	127
A.7.1. Description de la base de données utilisée	127
A.7.2. Evaluation et résultats	128
A propos de l'auteur	129
1. Biographie de l'auteur	129
2. Contributions scientifiques	130
3. Participations aux compétitions spécialisées	132

LISTE DES FIGURES

Figure 1	Images du mot 'beheaded' collectées à partir de 18 cahiers d'écriture différents [SUN 06]	1
Figure 2	Un manuscrit arabe ancien extrait de la bibliothèque de Tombouctou au Mali [SIT 01]	2
Figure 3	Une écriture obtenue à partir d'un patient atteint de la maladie de parkinson [CHU 12]	2
Figure 4	Une partie d'une lettre de menace de mort envoyée à un député à l'assemblée nationale constituante tunisienne [SIT 02]	3
Figure 5	Des exemples typiques des écritures féminines (à gauche) et masculines (à droite) de mot "anketa" (qui signifie "poll" en anglais) [SOK 12]	4
Figure 1.1	Illustration des transformations affines [SCH 98]	9
Figure 1.2	Illustration de la variabilité neuro-biomécanique [SCH 98]	9
Figure 1.3	Illustration de la variabilité de séquençement [SCH 98]	10
Figure 1.4	Illustration de la variabilité allographique [SCH 98]	10
Figure 1.5	Illustration de la variabilité intra-scripteur par la superposition d'un mot écrit 6 fois par la même personne [GAZ 06]	11
Figure 1.6	Illustration de la variabilité inter-scripteurs par la superposition d'un mot écrit par 6 personnes différentes [GAZ 06]	11
Figure 1.7	Une représentation schématique d'un système d'identification du scripteur [SID 10]	12
Figure 1.8	Une représentation schématique d'un système de vérification du scripteur [SID 10]	12
Figure 1.9	Une représentation schématique d'un système d'identification du genre d'un scripteur	13
Figure 1.10	Les deux modes d'acquisition de l'écriture, (a) Tracé du mot "sage", (b) Image du mot "dix" [BEL 01]	15
Figure 1.11	Distributions des taux de vraisemblance des utilisateurs légitimes et des imposteurs d'un système biométrique	17
Figure 1.12	Courbe ROC pour un système de vérification de scripteurs	18
Figure 1.13	Fenêtre principale du module de mesures interactives (WAM) du système WANDA [FRA 04]	21
Figure 1.14	La fenêtre principale de l'outil de comparaison de deux documents manuscrits du système CEDAR-FOX [SRI 07]	22
Figure 1.15	Interface de la station de travail du système FLASH ID au laboratoire du FBI [FIS 12]	23
Figure 2.1	Exemple de formes invariantes associées à un scripteur	30
Figure 2.2	Illustration des contours supérieurs et inférieurs de la composante connexe "manuscrit", (a) Contour supérieur, (b) Composante connexe originale, (c) Contour inférieur	31
Figure 2.3	Illustration de la segmentation de la composante connexe "manuscrit" en, (a) graphèmes, (b) bigrammes et (c) trigrammes	31
Figure 2.4	Exemples de groupes d'invariants obtenus sur une page d'écriture	32
Figure 2.5	Certaines des caractéristiques extraites à partir des images des caractères "d", "y", "f" et du graphème "th" [PER 07]	33
Figure 2.6	Différents allographes de la lettre « f » issus de différents scripteurs [TAN 09]	34
Figure 2.7	Schéma du système proposé par Tan et al [TAN 09]	34
Figure 2.8	Exemples de régions d'intérêt détectées par différents types de détecteurs (a) image originale, (b) détecteur DOG, détecteur MSER, et, (d) détecteur Harris-Hesse-Laplace	36
Figure 2.9	Exemple d'un regroupement en utilisant la quantification vectorielle (QV), (a) Premier mot visuel, (b) deuxième mot visuel, et (c) un troisième mot visuel	36
Figure 2.10	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	37
Figure 2.11	Codebook de 3-AS	37
Figure 2.12	Prétraitements et projections dans [ZOI 00], (a) Échantillon scanné d'un mot, (b) Mot binarisée et prétraité, (c) Projection avec les espaces, (d) Projection sans les espaces	38
Figure 2.13	Illustration de l'application de l'ouverture morphologique sur la fonction de projection (a) Fonction de projection initiale, (b) : ouverture avec un élément structurant de taille 3, (c) ouverture avec un élément structurant de taille 7	39
Figure 2.14	Exemple de lignes de référence [MAR 01]	41
Figure 2.15	Micro-caractéristiques extraites à partir de l'image du chiffre «6»	42
Figure 2.16	Variations structurelles de l'écriture, (a) : hauteur d'une ligne de texte, (b) : espaces entre pseudo mots, (c) : inclinaisons de l'écriture, (d) : différentes formes de points diacritiques	43
Figure 2.17	Prétraitements dans [FED 06], (a) Image de document binarisée; (b) Correction des angles de pente, (c) script horizontal, (d) projection horizontale, (e) normalisation des espaces inter-lignes, (f) projection verticale, (g) normalisation des espaces inter-mots et (h) remplissage de texte	45

Figure 2.18	Résultats de l'application des 16 filtres de Gabor considérées avec différentes fréquences et orientations [FED 06]	46
Figure 2.19	Caractéristiques locales extraites de chaque colonne de pixels dans la fenêtre glissante [SCL 07]	47
Figure 2.20	Le processus de génération de texture (a) le remplissage d'une ligne et (b) la texture résultante	48
Figure 2.21	Un échantillon de la base IAM et ses blocs de texture	49
Figure 2.22	Architecture du système proposé par Chaabouni et al [CHB 14]	50
Figure 2.23	Exemple illustrant la méthode d'extraction des directions des contours à partir de la lettre arabe « 9 » [BUL 07a]	53
Figure 2.24	Un Codebook contenant 400 graphèmes arabes [BUL 07a]	53
Figure 2.25	Illustration du découpage adaptatif de l'image du mot «headlines» en imogettes [SID 09]	54
Figure 2.26	Un codebook universel de taille 100 obtenu à partir d'échantillons de la base RIMES [SID 10]	55
Figure 2.27	Deux écritures en styles différents issues du même scripteur extraites de la base de la compétition ICDAR 2013 Competitions on Signature Verification and Writer Identification for On- and Offline Skilled Forgeries [MAL 13], (a) style 1, (b) style 2	59
Figure 3.1	Images de deux paragraphes du même scripteur	66
Figure 3.2	Textes grecs et anglais: (a) le même paragraphe écrit en grec et en anglais (b) les lettres minuscules et majuscules des alphabets grec et anglais	67
Figure 3.3	Calcul de la matrice de longueur de segments (a) Une image 8×6 avec deux valeurs de couleurs (0 et 1) (b) matrice de longueur de segments pour la direction 45° (c) matrice de longueur de segments pour la direction 90° (d) matrice de longueur de segments pour la direction 135° (e) matrice de longueur de segments pour la direction 180°	68
Figure 3.4	Distributions des longueurs des segments dans les directions (0°, 45°, 90° et 135°) sur (a) Pixels blancs. (b) Pixels noirs	69
Figure 3.5	Distributions des longueurs des segments de (a) Deux textes manuscrits issus du même scripteur, (b) Textes manuscrits de deux scripteurs différents	71
Figure 3.6	Performances en utilisant différentes quantités de texte: (a) Taux d'identification avec le classifieur SVM, (b) Taux d'identification avec la classifieur KNN, (c) Taux d'erreur égale	81
Figure 3.7	(a) Image Originale, (b-f) Images bruitées avec densités de 1%, 2%,3%,4% et 5% respectivement	81
Figure 3.8	Performances du système en fonction de la densité du bruit: (a) Taux d'identification avec le classifieur SVM, (b) Taux d'identification avec le classifieur KNN, (c) Taux d'erreur égale	83
Figure 4.1	Échantillons d'écriture, a) d'une femelle b) d'un male	86
Figure 4.2	Échantillons extraits (a) de la base MSHD, (b) de la base QUWI	87
Figure 4.3	Une image d'un caractère avec ses contours et la représentation en chaîne de codes de Freeman	89
Figure 4.4	Des échantillons d'écriture et leurs distributions de codes de Freeman respectives	89
Figure 4.5	La paire de chaîne de codes de Freeman (7, 6) représentant un angle de 135° au pixel de position p_i	90
Figure 4.6	Polygonisation, a) image originale b) Contours polygonisés	90
Figure 4.7	Angle entre les segments voisins des contours polygonisés	91
Figure 4.8	Construction d'un motif binaire et calcul du code LBP pour un pixel central	93
Figure A.1	Un échantillon de la base CVC	117
Figure A.2	Exemple d'un texte produit en quatre langues : (a) Grec, (b) Anglais, (c) Français, (d) Allemand, de la base de données de la compétition "The ICDAR 2011 Writer Identification contest"	118

LISTE DES TABLEAUX

Tableau 1.1	Synthèse des principales bases de données d'images de documents manuscrits	27
Tableau 2.1	Aperçu des compétitions organisées dans le cadre des conférences ICDAR 2011, ICFHR 2012 et ICDAR 2013	56
Tableau 2.2	Aperçu des méthodes proposées dans le domaine de la classification de scripteurs	61
Tableau 3.1	Une vue d'ensemble des caractéristiques implémentées et leurs dimensionnalités	72
Tableau 3.2	Résultats de l'identification de scripteurs en utilisant la base IFN-ENIT	73
Tableau 3.3	Résultats de l'identification de scripteurs en utilisant la base KHATT	73
Tableau 3.4	Résultats de l'identification de scripteurs en utilisant la base CVL	74
Tableau 3.5	Résultats de l'identification de scripteurs en utilisant les bases des compétitions ICDAR et ICFHR	74
Tableau 3.6	Résultats de l'identification et la vérification du scripteur en utilisant les textes grecs	76
Tableau 3.7	Résultats de l'identification et la vérification du scripteur en utilisant les textes anglais	77
Tableau 3.8	Les résultats de l'identification et de la vérification du scripteur en utilisant les documents grec en apprentissage et les documents anglais en test et vice-versa	78
Tableau 4.1	Répartition de la base de données QUWI	87
Tableau 4.2	Récapitulatif des caractéristiques	94
Tableau 4.3	Taux de classification sur les bases QUWI et MSHD	95
Tableau 4.4	Taux de classification des évaluations dépendantes et indépendantes du texte sur les bases de données QUWI et MSDB	96
Tableau 4.5	Taux de classification des évaluations dépendantes du script sur les bases de données QUWI et MSHD	97
Tableau 4.6	Taux de classification des évaluations dépendantes du script sur les bases de données QUWI et MSHD	98
Tableau 4.7	Taux de classification des évaluations inter-bases de données	99
Tableau A.1	Performances des systèmes participants à la compétition "The ICDAR2011 Arabic Writer Identification Contest" en %.	116
Tableau A.2	Performances des méthodes participantes à la compétition "The ICDAR 2011 Music Scores Competition: Staff Removal and Writer Identification" en %.	117
Tableau A.3	Systèmes et pays des participants à la compétition "The ICDAR 2011 Writer Identification contest".	119
Tableau A.4	Résultats d'identification avec le critère "soft TOP-N" en utilisant toute la base de données en %.	119
Tableau A.5	Résultats d'identification avec le critère "hard TOP-N" en utilisant toute la base de données en %.	120
Tableau A.6	Résultats d'identification avec le critère "hard TOP-N" en utilisant les documents grecs en %.	120
Tableau A.7	Résultats d'identification avec le critère "hard TOP-N" en utilisant les documents anglais en %.	120
Tableau A.8	Résultats d'identification avec le critère "hard TOP-N" en utilisant les documents français en %.	120
Tableau A.9	Résultats d'identification avec le critère "hard TOP-N" en utilisant les documents allemands en %.	121
Tableau A.10	Résultats d'identification avec le critère "soft TOP-N" en utilisant toute la base des images recadrées en %.	121
Tableau A.11	Résultats d'identification avec le critère "hard TOP-N" en utilisant toute la base des images recadrées en %.	121
Tableau A.12	Résultats d'identification avec le critère "hard TOP-N" en utilisant les documents grecs recadrés en %.	121
Tableau A.13	Résultats d'identification avec le critère "hard TOP-N" en utilisant les documents anglais recadrés en %.	122
Tableau A.14	Résultats d'identification avec le critère "hard TOP-N" en utilisant les documents français recadrés en %.	122
Tableau A.15	Résultats d'identification avec le critère "hard TOP-N" en utilisant les documents allemands recadrés en %.	122

Tableau A.16	Résultats d'identification avec le critère "soft TOP-N" en utilisant toute la base de données en %.	123
Tableau A.17	Résultats d'identification avec le critère "hard TOP-N" en utilisant toute la base de données en %.	124
Tableau A.18	Résultats d'identification avec le critère "soft TOP-N" en utilisant les documents anglais en %.	124
Tableau A.19	Résultats d'identification avec le critère "soft TOP-N" en utilisant les documents grecs en %.	124
Tableau A.20	Résultats d'identification de certains systèmes parmi ceux qui ont participé à la compétition "ICFHR2012 Competition on Writer Identification - Challenge 2: Arabic Scripts " en %.	125
Tableau A.21	Résultats d'identification avec le critère "soft TOP-N" en utilisant toute la base de données en %.	126
Tableau A.22	Résultats d'identification avec le critère "hard TOP-N" en utilisant toute la base de données en %.	126
Tableau A.23	Résultats d'identification avec le critère "soft TOP-N" en utilisant les documents grecs en %.	127
Tableau A.24	Résultats d'identification avec le critère "soft TOP-N" en utilisant les documents anglais en %.	127
Tableau A.25	Performances des systèmes participants à la tâche "Writer Identification" de la compétition "ICDAR2013 Competitions on Signature Verification and Writer Identification for On- and Offline Skilled Forgeries".	128



INTRODUCTION GENERALE

1. Contexte de la recherche

L'écriture est l'un des plus anciens modes de communication dans notre civilisation, ce mode s'est beaucoup développé et a évolué significativement à travers les siècles. Un individu apprend à écrire en copiant des formes à partir d'un cahier d'écriture standard qui, lui aussi, diffère selon la localisation géographique ainsi que les circonstances temporelles, sociales et culturelles. Cependant, le style appris à partir du cahier d'écriture (voir figure 1), s'atténue avec le temps et la personne développe ses préférences personnelles d'écriture. Dans ce contexte, contrairement aux textes électroniques ou imprimés, un texte manuscrit apporte des informations supplémentaires sur l'individu qui l'a produit. Ceci rend l'analyse de l'écriture un domaine de recherche attractif pour les psychologues, les examinateurs médico-légaux de documents, les paléographes, les graphologues et les analystes légistes.

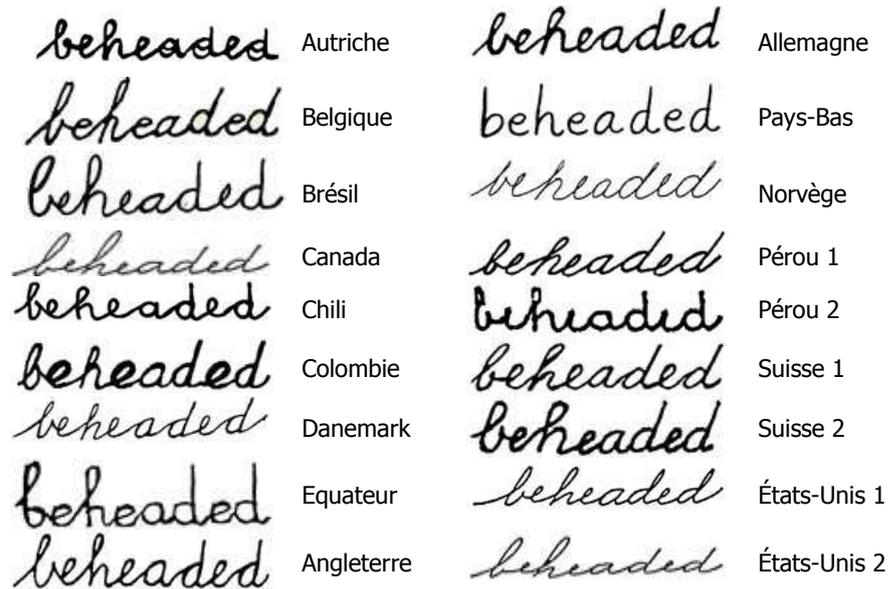


Figure 1. Images du mot 'beheaded' collectées à partir de 18 cahiers d'écriture différents [SUN 06].

Un texte manuscrit peut présenter des intérêts variés. Les manuscrits anciens (voir figure 2), par exemple, pourraient servir à étudier l'évolution de la forme et du style d'écriture d'une société au fil du temps, ce qui, à son tour reflète les changements historiques et culturels de cette société. Des connaissances sur les différentes lettres, les ligatures, les signes de ponctuation, les abréviations et la façon avec laquelle ils ont évolué, permet aux paléographes

et historiens d'identifier les périodes au cours desquelles un manuscrit a été écrit. La quantité de manuscrits anciens stockés dans des archives, des bibliothèques et des collections privées est considérable et il serait très utile de développer des systèmes informatiques qui pourraient aider les paléographes à dater, classer et authentifier ces manuscrits.



Figure 2. Un manuscrit arabe ancien extrait de la bibliothèque de Tombouctou au Mali [SIT 01].

L'écriture manuscrite a aussi une relation intéressante avec les neurosciences et plusieurs troubles neurologiques (qui pourraient affecter la motricité de l'écriture) sont reflétés par l'écriture du patient. L'écriture manuscrite pourrait donc avoir une valeur diagnostique pour les neurologues, particulièrement s'ils ont des échantillons antérieurs de l'écriture du patient. A titre d'exemple, les changements dans l'écriture d'une personne peuvent être analysés pour le diagnostic de la maladie de Parkinson dont les effets se manifestent, même dans les premiers stades, dans l'écriture du patient. Cette dernière a tendance à devenir de plus en plus petite à mesure que la main progresse vers la fin d'une ligne et en passant d'une ligne à l'autre (voir figure 3). Les changements des formes de l'écriture pourraient ensuite être utilisés pour suivre la progression de la maladie de Parkinson.

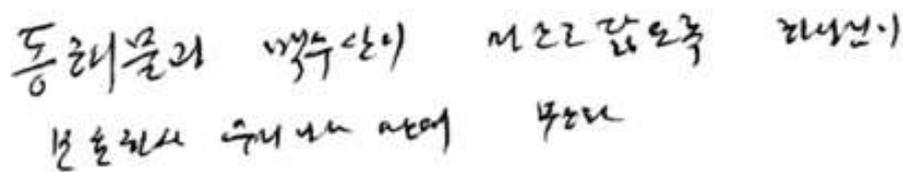


Figure 3. Ecriture d'un patient atteint de la maladie de Parkinson [CHU 12].

Du point de vue de la graphologie, l'écriture est un moyen perspicace de profilage de la personnalité, qui met en évidence les traits de caractère ainsi que le suivi des sentiments et des émotions d'une personne. Effectivement, les graphologues affirment que l'écriture pourrait révéler plus de deux cents traits de personnalité tels que l'humeur, le tempérament, le mode de pensée, la peur, la maturité et les interactions sociales. C'est peut-être pour cette raison que l'écriture manuscrite 'handwriting' est aussi appelé 'brainwriting'. Non seulement elle révèle les traits de la personnalité de celui qui écrit, mais un changement dans notre manière d'écrire peut aussi entraîner des changements de la personnalité. La graphothérapie, qui consiste à

améliorer une écriture altérée, illisible, trop lente ou fatigante, va permettre le développement de traits de personnalité positifs et l'atténuation, ou la disparition, des traits négatifs indésirables.

En ce qui concerne l'analyse légale et judiciaire, l'écriture manuscrite est devenue un élément ayant une importance significative dans la biométrie comportementale. Les examinateurs légaux de documents pourraient avoir besoin de vérifier l'authenticité et/ou la falsification d'écrits, de documents ou de signatures ou encore d'en déterminer l'auteur comme c'est le cas pour une demande de rançon, une lettre frauduleuse, une lettre de suicide suspecte ou une lettre de menaces (voir figure 4). Cette analyse est une procédure fastidieuse pour les examinateurs légaux de documents, par conséquent, des systèmes informatisés pour l'analyse des écritures manuscrites pourraient servir d'outils précieux dans l'analyse des documents judiciaires. Ces systèmes prennent comme entrées des images scannées de documents manuscrits et emploient des techniques de reconnaissance de formes et de traitement d'images afin de résoudre un problème donné. Bien entendu, les résultats de ces systèmes peuvent ne pas être acceptés comme preuves dans une affaire judiciaire et l'intervention d'experts humains est inévitable. Néanmoins, ils se sont toujours avérés très utiles pour accélérer considérablement le processus d'expertise.

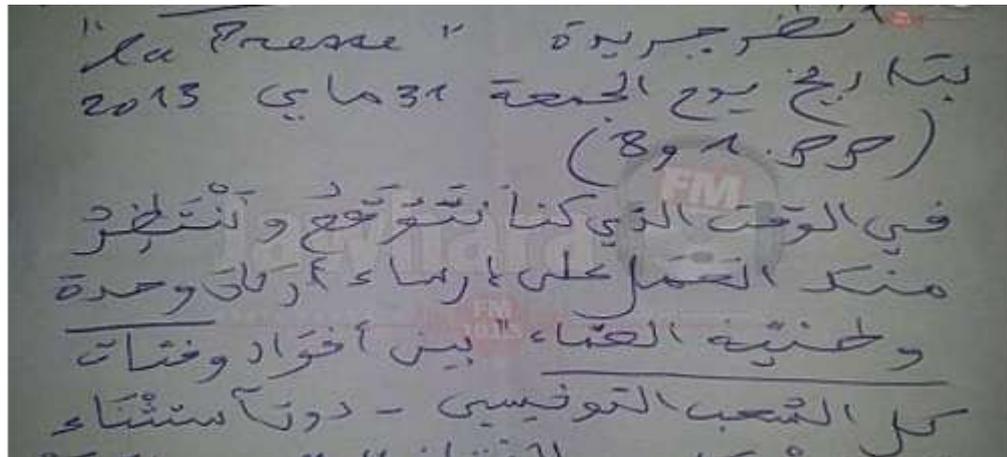


Figure 4. Une partie de lettre de menace de mort envoyée à un député à l'assemblée nationale constituante tunisienne [SIT 02].

L'écriture manuscrite peut également être utilisée pour déterminer le sexe d'un individu (voir figure 5), la main dont il se sert pour écrire, son âge ainsi que son origine ethnique. Un intérêt tout particulier a été porté à la détermination du sexe d'un individu à partir de son écriture manuscrite. Ceci est probablement dû à ses nombreuses applications telles que l'expertise légale où cela peut aider les enquêteurs à se concentrer davantage sur une certaine catégorie de suspects. L'identification du sexe d'un scripteur peut également être utilisée pour améliorer les performances d'un système de reconnaissance de l'écriture manuscrite. La variabilité qui existe entre les deux types de scripteurs (masculin et féminin) est plus petite que dans une population de scripteurs, ce qui nous permet d'entraîner des reconnaisseurs spécialisés. Les études démographiques représentent une application assez intéressante. Un exemple concret serait d'analyser les écritures manuscrites disponibles sur le 'World Wide Web' et de savoir combien de personnes de chaque catégorie (masculine/féminine) ont contribué aux données.



Figure 5. Des exemples typiques des écritures féminines (à gauche) et masculines (à droite) du mot "anketa" (qui signifie "poll" en anglais) [SOK 12].

Pour résumer, nous pouvons dire qu'une écriture manuscrite comporte des informations considérables sur son auteur, son style d'écriture, son sexe ainsi que sa personnalité. Forts de ce constat, nous avons entrepris une étude sur le développement d'algorithmes qui permettent une analyse automatique de l'écriture manuscrite. Étant donné que l'analyse de l'écriture manuscrite à différents aspects, nous avons focalisé notre recherche sur la classification des écritures manuscrites, en particulier, la reconnaissance du scripteur d'un document manuscrit ainsi que la détermination de son sexe.

La tâche de reconnaissance du scripteur est étroitement liée à celle de la reconnaissance de l'écriture manuscrite. En effet, la reconnaissance du scripteur doit tirer profit de la variabilité des écritures manuscrites afin de les discriminer, alors que la reconnaissance de l'écriture doit, au contraire, parvenir à s'affranchir de la variabilité entre les scripteurs pour identifier le message textuel quel qu'en soit le scripteur. En dépit des différences entre les deux tâches, la reconnaissance de scripteurs peut être utile pour la reconnaissance de l'écriture manuscrite, en exploitant le principe de l'adaptation du système au type du scripteur [NOS 04].

La tâche de détermination du sexe d'un scripteur à partir de documents manuscrits est aussi étroitement liée à celle de la reconnaissance de scripteurs, la seule différence réside dans le fait que l'identification de scripteurs est un problème de prise de décision entre N classes (scripteurs) tandis que la détermination du sexe est un problème de prise de décision entre deux classes : masculine et féminine. Malgré cette différence entre les deux tâches, la détermination du sexe d'un scripteur peut être utilisée afin de réduire la complexité de la tâche de reconnaissance de scripteurs.

2. Objectifs de la recherche

Les caractéristiques utilisées dans le domaine de la reconnaissance de scripteurs se divisent en deux grandes catégories: texturales et allographiques. Parmi les caractéristiques texturales, nous pouvons citer les directions des contours, les charnières des contours, l'autocorrélation, les cooccurrences des angles des contours, les filtres de Gabor, les matrices de cooccurrences et

les chaînes de codes de Freeman. Les caractéristiques allographiques sont extraites par la segmentation du texte en lignes, mots, caractères, graphèmes, et parfois en allotraits etc. Une technique importante est de regrouper les graphèmes et générer un codebook qui sera utilisé afin de caractériser les scripteurs.

Des bases de données expérimentales contenant des données issues d'une vingtaine jusqu'à un millier de scripteurs ont été considérées dans les travaux de l'état de l'art. On peut constater que la tendance qui prévaut dans ces études est de traiter des documents monolingues, c'est à dire que l'entraînement et le test des systèmes de reconnaissance de scripteurs sont effectués en utilisant des documents écrits dans la même langue. Comment va se comporter une méthode particulière quand on change de script ou de langue?. Ce scénario n'a pas été considéré jusqu'à présent... La capacité d'une méthode à traiter des échantillons de textes issus de scripteurs qui écrivent en utilisant différentes langues et différents scripts n'a pas été étudiée aussi...

L'utilisation des caractéristiques allographiques nécessite des connaissances sur le script traité pour déterminer comment segmenter les lignes en mots et les mots en caractères ou en graphèmes et, par conséquent, l'extension des méthodes basées sur des caractéristiques allographiques n'est pas simple, surtout dans un environnement multi-scripts où les scripteurs ne sont pas obligés d'utiliser un seul script. L'utilisation des caractéristiques extraites à partir du contour exige également une connaissance parfaite sur les contours des caractères.

D'un autre côté, les caractéristiques texturales comme l'autocorrélation, les cooccurrences des angles des contours, les longueurs de segments, etc. peuvent représenter un choix approprié pour être utilisées dans un environnement multi-scripts et ceci peut être justifié par le fait que l'extraction de ces caractéristiques n'a pas besoin de connaissances approfondies sur la nature du script. Malheureusement, les performances de ces caractéristiques sont tellement faibles qu'il est difficile d'élaborer un système de reconnaissance du scripteur basé sur ces dernières.

Notre première contribution dans le cadre de cette thèse consiste à étendre la recherche dans le domaine de la reconnaissance de scripteurs à un environnement multi-scripts. Les matrices de longueurs de segments sont introduites pour la caractérisation des scripteurs. La motivation derrière l'utilisation de ces caractéristiques vient de son application réussie dans divers domaines, y compris l'analyse de texture [GAL 75]. Aucune connaissance spécifique sur le script n'est nécessaire pour utiliser ces caractéristiques, ce qui offre un cadre viable pour traiter des documents issus d'individus utilisant plus d'un script.

Plusieurs études psychologiques affirment que l'un des facteurs influençant certains aspects de la production de l'écriture manuscrite d'un scripteur est son sexe. La vitesse et la pression, en moyenne moins élevées chez les femmes que chez les hommes, permettent de discriminer les écritures masculines et féminines [MEU 89]. D'autres études ont montré que certains paramètres spatiaux tels que la rondeur, l'inclinaison et le style d'écriture (script versus cursif) permettent également d'établir cette différence [MAA 86].

La délicatesse et l'aspect décoratif de l'écriture manuscrite ont été considérés afin de déterminer les écritures féminines [HAM 96]. Aussi, l'embellissement peut être utilisé comme un élément pour décrire les écritures féminines [DIL 98]. Dans certaines études [BUR 02, JAM 91], lorsque des enfants ont été invités à imiter l'écriture du sexe opposé, les garçons ont essayé de rendre leur écriture plus petite et plus propre, tandis que les filles ont essayé de rendre leur écriture plus grande.

L'objectif de notre deuxième contribution est de déterminer si on peut prédire 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 ou document manuscrit) peut être attribuée spécifiquement à l'une ou l'autre des deux catégories du genre (masculin/féminin)? Si oui, quelles sont les caractéristiques les mieux adaptées pour ce type de problème et avec quel degré de précision? Est-ce que le contenu textuel des échantillons a une influence sur les performances ?, Est-ce que l'utilisation en test d'un script différent de celui utilisé en apprentissage (entraînement) peut affecter la précision du système?, la localisation géographique des scripteurs affecte-elle aussi les performances du système ?

Pour répondre à toutes ces questions, nous avons implémenté un ensemble d'attributs discriminants parmi ceux définis par les psychologues afin qu'ils soient utilisés pour la discrimination entre les écritures féminines et masculines. L'inclinaison et l'orientation, la rondeur et la courbure, la propreté et la lisibilité ainsi que la texture des écritures manuscrites ont été considérées. La classification est effectuée à l'aide des réseaux de neurones (ANN) ainsi que les séparateurs à vaste marge (SVM). Nous avons évalué notre système sur deux bases de données multi-scripts issues de deux régions complètement différentes (la première de l'Algérie du Maghreb Arabe et l'autre de Qatar du Moyen Orient). Plusieurs scénarios d'évaluation ont été considérés, ces scénarios incluent la détermination du sexe d'un scripteur à partir de son écriture, l'étude de l'influence du contenu textuel ainsi que celui du script utilisé sur la précision du système proposé et enfin l'étude de l'influence de la localisation géographique des scripteurs sur les performances de classification.

3. Organisation du mémoire de thèse

Le présent document est structuré en quatre chapitres répartis en deux parties. La première partie, contenant les chapitres 1 et 2 est consacrée à la présentation des principaux concepts, outils et travaux relatifs à l'étude entreprise. Dans la deuxième partie du mémoire, représentée par les chapitres 3 et 4, nous abordons de manière détaillée nos choix conceptuels, la mise en œuvre ainsi que les résultats obtenus par les systèmes proposés pour la reconnaissance des scripteurs ainsi que la détermination de leur sexe.

Chapitre 1. Analyse de l'écriture manuscrite pour la classification de scripteurs: Concepts et outils.

Ce chapitre est consacré à la présentation des principaux concepts et outils du domaine de la classification de scripteurs. Il présente d'abord les mécanismes de production de l'écriture manuscrite ainsi que les principales sources de variabilités pour comprendre comment reconnaître un individu ou le classer dans une catégorie spécifique (sexe, âge, latéralité ...etc) en utilisant son écriture manuscrite. Nous adoptons une catégorisation des différents systèmes de classification de scripteurs en fonction de la tâche envisagée (identification ou vérification), de la dépendance au texte (dépendant ou indépendant du texte) ainsi que du mode d'acquisition de l'écriture (en-ligne ou hors ligne). Les méthodes utilisées pour l'évaluation des performances de ces systèmes sont exposées ainsi que les applications possibles de la classification de scripteurs. Nous passons en revue les systèmes existants et nous terminons le chapitre par un survol des principales bases de données utilisées dans le domaine.

Chapitre 2. *Approches de classification de scribes: 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 scribes. Nous nous concentrons dans sa première partie sur la présentation des principaux travaux de recherche dans le domaine de la reconnaissance de scribes. Nous distinguons entre les approches basées sur des caractéristiques locales, globales ou sur la combinaison des deux. Ensuite, nous discutons les diverses compétitions dans le domaine de la classification de scribes et enfin nous terminons le chapitre par une comparaison des différents travaux du domaine en nous basant, 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 3. *Reconnaissance du scribe en mode indépendant du texte en utilisant des documents multi-scripts*

Ce chapitre se détache des aspects théoriques abordés dans les deux premiers chapitres et s'oriente vers la présentation de notre première contribution principale qui consiste en une approche d'identification de scribes utilisant des documents multi-scripts (latins et grecs), basée sur la caractérisation des différents scribes par les distributions de longueurs de segments (Run-length). Nous décrivons la base de données utilisée ainsi que les similitudes et différences entre les scripts latins et grecs avant de nous focaliser sur la présentation détaillée de la méthode d'extraction de caractéristiques proposée, des caractéristiques de l'état de l'art implémentées pour des fins de comparaison, des classificateurs employés pour la reconnaissance de scribes (identification et vérification). Les expérimentations effectuées concernent aussi bien un environnement dépendant de la langue qu'un environnement Multi-scripts (Grec versus Anglais et vice versa). A la fin de ce chapitre, les résultats sont exposés et discutés avant de présenter une étude détaillée sur la stabilité des caractéristiques proposées en fonction de la quantité du texte disponible et de la présence de bruit dans les images considérées.

Chapitre 4. *Analyse automatique de l'écriture manuscrite pour la reconnaissance du genre d'un individu*

Ce quatrième 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 scribes. 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 qui concernent les modes dépendant et indépendant du texte ainsi que les évaluations croisées entre les bases de données (cross-databases evaluations) sont exposées et discutées la fin de ce chapitre et valident l'hypothèse qu'une corrélation existe entre une écriture manuscrite et le sexe de son scribe.

A la fin de ce mémoire, nous émettons nos conclusions sur les recherches que nous avons entreprises dans le domaine de la classification de scribes. Comme aucune machine ne peut égaler les performances humaines dans le domaine de l'analyse et la reconnaissance de l'écriture manuscrite, plusieurs perspectives restent envisageables pour faire évoluer les propositions que nous avons présentées dans ce document.

ANALYSE DE L'ÉCRITURE MANUSCRITE POUR LA CLASSIFICATION DE SCRIPTEURS: CONCEPTS ET OUTILS

Ce chapitre est consacré à la présentation des principaux concepts et outils du domaine de la classification de scripteurs. Il présente les mécanismes de production de l'écriture manuscrite ainsi que les principales sources de variabilités pour comprendre comment reconnaître un individu ou le classer dans une catégorie démographique spécifique (genre, âge, latéralité) en se basant sur son écriture manuscrite. Nous adoptons une catégorisation des différents systèmes de classification de scripteurs en fonction de la tâche envisagée (identification ou vérification), de la dépendance au texte (dépendant ou indépendant du texte) ainsi que du mode d'acquisition de l'écriture (en-ligne ou hors ligne). Les méthodes utilisées pour l'évaluation des performances de ces systèmes sont exposées ainsi que les applications possibles. Nous passons en revue les systèmes existants et nous terminons le chapitre par un survol des principales bases de données utilisées dans le domaine.

1.1. L'écriture manuscrite: un moyen de caractérisation des individus

Le terme "écriture" peut se référer aux mouvements complexes effectués par la main au cours de la production d'un texte [SES 12] ou aux résultats de ce processus. En tant que processus, l'écriture est une tâche perceptivo-motrice complexe, une compétence qui s'acquiert généralement à l'école. La main est un mécanisme extrêmement complexe et délicat qui contient 27 os contrôlés par plus de 40 muscles différents [HUB 99], la plupart de ces muscles sont situés dans l'avant bras et sont connectés aux doigts par un ensemble complexe de tendons. Elle est innervée par 3 nerfs qui exercent des fonctions sensorielles et motrices différentes.

L'écriture a reçu beaucoup d'attention à partir du moment où les êtres humains ont commencé à écrire [GIL 11]. Son individualité, ainsi que sa facilité l'ont rendue un moyen de communication majeur. Avec l'avènement de supports électroniques de communication et de stockage rapides et faciles à utiliser, l'écriture est devenue un centre d'intérêt pour les chercheurs du domaine de l'analyse et la reconnaissance automatique de l'écriture manuscrite. L'écriture manuscrite peut contenir divers types d'informations, on peut citer, par exemple, les informations sur ce qui a été écrit (le texte), les informations sur celui qui a écrit, telles que son identité, son genre, son âge ainsi que sur sa personnalité. L'extraction du premier type d'information se fait par le biais de techniques de reconnaissance de l'écriture manuscrite alors que le deuxième type d'information peut être exploité par des systèmes de classification des scripteurs.

Bien que la tâche de classification des scripteurs soit liée étroitement à celle de la reconnaissance de l'écriture manuscrite [SID 10, BUL 07, BEN 04], elle ne semble pas poser le

même type de difficultés. Le but dans la reconnaissance de l'écriture est d'obtenir l'invariance et la généralisation, alors que, dans la classification de scribes, on essaye d'obtenir tout à fait l'opposé dans le but d'exploiter au maximum les variabilités des écritures manuscrites pour la discrimination et la classification de leurs scribes. Il est important, cependant, de mentionner que la classification de scribes pourrait réduire certaines ambiguïtés dans le processus de reconnaissance de l'écriture si les informations sur les habitudes et les particularités générales de l'écriture d'un scribe, ou d'une catégorie de scribes, sont disponibles au système de reconnaissance de l'écriture [CRE 95, NOS 99, BUL 07a, SID 10].

1.1.1. Facteurs causant la variabilité des écritures manuscrites

Quatre facteurs influençant la production de l'écriture ont été identifiés par Schomaker [SCH 98]: Le premier facteur de variation des écritures concerne les transformations affines que le scribe impose à l'écriture (voir figure 1.1) et qui sont sous son contrôle volontaire. Parmi ces transformations, on peut citer, les transformations de taille, de translation, de rotation, et de cisaillement. Elles représentent certainement une nuisance pour la classification des écritures manuscrites, mais pas un obstacle fondamental. En particulier, l'inclinaison de l'écriture constitue un paramètre habituel qui peut être exploité par les systèmes de classification de scribes.

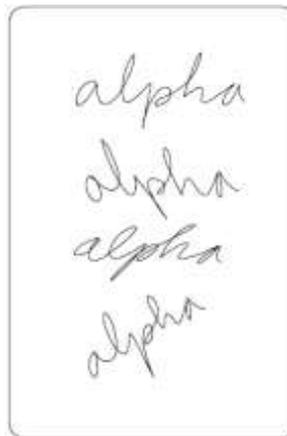


Figure 1.1. Illustration des transformations affines [SCH 98].

La variabilité neuro-biomécanique (voir figure 1.2) représente le facteur le plus fréquemment rencontré causant de nombreuses variations des écritures manuscrites. Cette variabilité est assimilée à un manque de soin dans la production des formes des différents caractères à cause d'un effort irrégulier produit par le scribe, ce qui influe remarquablement sur la lisibilité de son écriture. Ce facteur est davantage lié à l'état du scribe qu'à son identité.

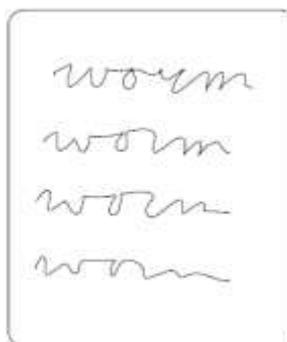


Figure 1.2. Illustration de la variabilité neuro-biomécanique [SCH 98].

La variabilité de séquençement représente également un facteur causant des variations à l'écriture. Ce facteur est hautement dépendant de l'état instantané du scripteur durant le processus d'écriture (voir figure 1.3). L'ordre des traits peut varier stochastiquement comme dans la production d'un E, un E à quatre traits peut être produit à $4! \times 2^4 = 384$ permutations [SCH 98]. Dans la production de certains scripts asiatiques, tel que HANZI, la permutation de l'ordre de traits des lettres rend la signification complètement différente (malgré le fait que l'apprentissage dans les écoles de l'ordre des traits dans l'écriture soit assez strict) ce qui représente un problème bien connu en reconnaissance de l'écriture.

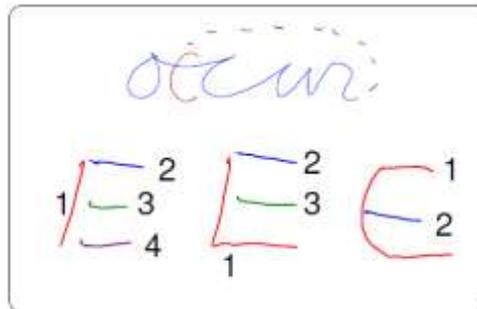


Figure 1.3. Illustration de la variabilité de séquençement [SCH 98].

Un autre facteur qui peut survenir est celui des variations allographiques (voir figure 1.4) ou la quantité de formes de caractères utilisées dans une population de scripteurs pour une lettre donnée de l'alphabet. Il y a de grandes différences de formes entre les caractères produits par différents scripteurs, surtout quand ils sont de nationalités ou de générations différentes. Ce facteur produit la plupart des problèmes en reconnaissance de l'écriture manuscrite, mais, parallèlement, fournit des informations essentielles pour la reconnaissance du scripteur.

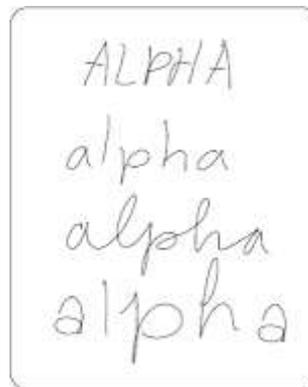


Figure 1.4. Illustration de la variabilité allographique [SCH 98].

Il existe d'autres facteurs qui peuvent influencer considérablement l'écriture d'un scripteur [HUB 99, CRE 98], on peut citer, par exemple, le sexe, l'âge, l'origine ethnique, la main dont se sert un scripteur pour écrire (gaucher ou droitier), l'état de santé physique et mentale du scripteur, le système d'écriture appris, le sujet (contenu), le protocole d'écriture (de mémoire, dictée ou copie), la nature du support utilisé (papier, plastique, etc.), le type de stylo (plume, feutre, bille, etc.), la vitesse d'écriture, la tâche à effectuer (lettre personnelle, lettre officielle, brouillon, notes personnelles) ainsi que les changements dans l'écriture d'un individu au fil du temps, etc.

1.1.2. Différents types de variations des écritures manuscrites

Bien que les individus d'un pays donné reçoivent une éducation commune, et subissent, notamment, un apprentissage commun de l'écriture, les graphismes qu'ils produisent sont extrêmement variables [CRE 98]. Il existe trois types de variabilités des écritures manuscrites : la variabilité inter-scripteurs qui dénote la variation du style d'écriture entre différentes personnes, la variabilité intra-scripteur qui représente les variations de l'écriture d'une même personne au cours du temps et qui dépend de son état physique et psychologique (variations intrinsèques à chaque scripteur) ainsi que la variabilité inter-classes qui dénote la variation du style d'écriture entre les deux catégories de scripteurs (masculine et féminine).

La reconnaissance de scripteurs est seulement possible dans la mesure où la variabilité inter-scripteurs excède la variabilité intra-scripteur [SID 10]. Autrement dit, la capacité des systèmes de reconnaissance de scripteurs à reconnaître une personne repose essentiellement sur la capacité à discriminer les personnes grâce aux variabilités des écritures. Nous considérons la superposition d'un mot écrit six fois par le même scripteur (voir figure 1.5) et celle produite par six scripteurs différents (voir figure 1.6). Ces figures illustrent les variations intra et inter-scripteurs et montrent une stabilité relative de l'écriture pour un même scripteur.



Figure 1.5. Illustration de la variabilité intra-scripteur par la superposition d'un mot écrit 6 fois par la même personne [GAZ 06].



Figure 1.6. Illustration de la variabilité inter-scripteurs par la superposition d'un mot écrit par 6 personnes différentes [GAZ 06].

D'un autre côté, la classification de scripteurs en fonction de leur catégorie (le sexe, l'âge, l'origine ethnique, la latéralité) doit s'affranchir de la variabilité des écritures des différents scripteurs et se focaliser sur ce qui ressortent comme caractéristiques communes de la catégorie à laquelle ils appartiennent. Ainsi, deux écritures semblables appartiendront à une même catégorie même si elles sont produites par deux scripteurs différents.

1.2. Catégorisation des systèmes de classification de scripteurs

Sur le plan méthodologique, la complexité des systèmes de classification de scripteurs peut être évaluée en tenant compte d'un ensemble de critères dont les plus connus sont : la tâche de reconnaissance (identification ou vérification), la dépendance du texte (dépendant ou indépendant du texte) et enfin le mode d'acquisition de l'écriture (en-ligne ou hors-ligne).

1.2.1. Catégorisation par tâche

La classification de scripteurs est un terme générique qui répond à plusieurs définitions selon le scénario applicatif envisagé. Les scénarios applicatifs qui seront abordés dans ce mémoire peuvent être regroupés en trois catégories principales :

- L'identification de scripteurs.
- La vérification de scripteurs.
- La classification de scripteurs en fonction de leur sexe.

Pour la tâche de l'identification du scripteur [SID 10], étant donné un échantillon de document manuscrit S inconnu et une base de données contenant des échantillons de documents manuscrits appartenant à N scripteurs connus, l'objectif est de trouver le scripteur (ou la liste probable de scripteurs) du document manuscrit S dans la base de données. La figure 1.7 illustre une représentation schématique du principe de base de la tâche d'identification du scripteur.

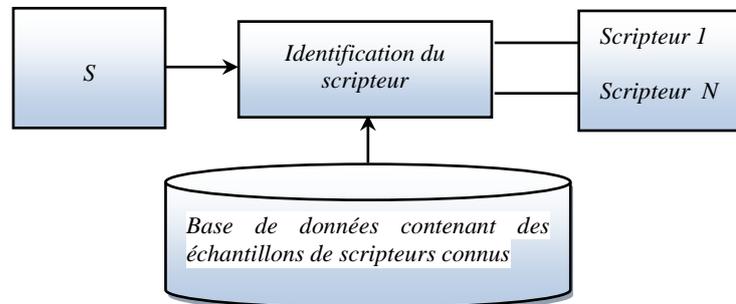


Figure 1.7. Une représentation schématique d'un système d'identification du scripteur [SID 10].

Deux modes sont possibles en identification de scripteurs: l'identification en ensemble fermé dans laquelle on suppose que le document manuscrit est effectivement écrit par un scripteur connu du système et l'identification en ensemble ouvert dans laquelle le scripteur peut ne pas être connu. Il est important de noter que tous les travaux présentés jusqu'à présent dans le domaine de l'identification du scripteur procèdent à une identification en ensemble fermé à l'exception de ceux présentés par Schlapbach et al. [SCL 04a, SCL 04b, SCL 05, SCL 06, SCL 07b, SCL 07a].

Pour la tâche de vérification du scripteur [SID 10], étant donné deux échantillons de documents manuscrits S_1 et S_2 l'objectif est de déterminer si les deux ont été écrits par la même personne ou par deux personnes différentes. La tâche de vérification du scripteur est illustrée dans la figure 1.8.

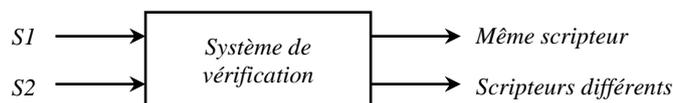


Figure 1.8. Une représentation schématique d'un système de vérification du scripteur [SID 10].

En ce qui concerne la tâche de classification de scripteurs en fonction de leur sexe, étant donné un échantillon de document manuscrit S inconnu et une base de données contenant des échantillons de documents manuscrits appartenant à deux catégories différentes de scripteurs (masculine et féminine), l'objectif est de déterminer si le document manuscrit S a été écrit par un scripteur de sexe masculin ou féminin. Une illustration schématique de la tâche d'identification du sexe d'un scripteur est présentée dans la figure 1.9.

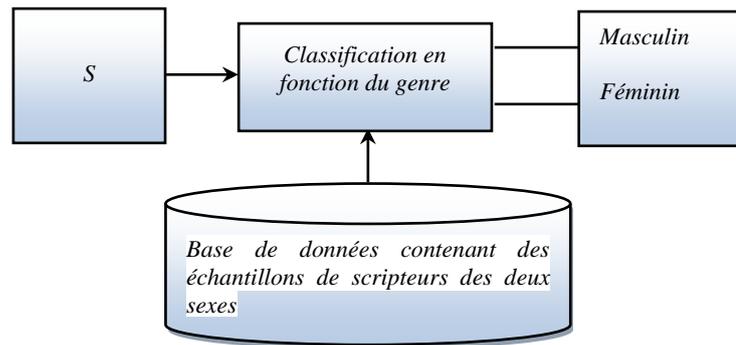


Figure 1.9. Une représentation schématique d'un système d'identification du genre d'un scripteur.

Lors du développement d'un système de classification de scripteurs, quatre étapes principales peuvent être distinguées :

Prétraitements : Cette étape implique la collecte et la préparation des données dans une forme appropriée [GIL 11] afin qu'elles soient facilement utilisables. Cette étape est évidemment très spécifique au type de données à classifier (en-ligne ou hors-ligne). Pour la classification de scripteurs en utilisant des images de documents, l'étape de prétraitements peut inclure la numérisation des images de documents, l'élimination du bruit, la normalisation, la correction de l'inclinaison, la sélection de régions d'intérêt, la binarisation, ainsi que toute autre opération permettant de soutenir les étapes ultérieures.

Extraction de caractéristiques : Généralement, les échantillons de textes présents en entrée d'un système de classification de scripteurs contiennent beaucoup d'informations qui ne peuvent pas être utilisées pour leur classification, et une grande partie de ces informations n'est pas assez pertinente pour déterminer la bonne classe. L'objectif essentiel de cette étape repose sur l'extraction, à partir de l'échantillon de texte, d'un ensemble d'attributs qui sont des indicateurs utiles sur la classe à laquelle il appartient [GIL 11, SCL 07a]. Ces attributs sont connus comme des "caractéristiques", et leur choix représente une étape importante dans le domaine de la classification de scripteurs. En outre, il y a quelques étapes facultatives qui peuvent être effectuées en dérivant des caractéristiques améliorées de celles initialement extraites. Les caractéristiques peuvent être combinées (combinaison de caractéristiques) de différentes manières pour générer des caractéristiques plus performantes, ou elles peuvent être soumises à une phase de test préliminaire afin de choisir les attributs les plus précis et les plus discriminatoires (sélection de caractéristiques).

Apprentissage (ou entraînement): Dans cette étape, et pour chaque classe, un modèle est construit et entraîné en utilisant des échantillons de textes issus de cette classe [SCL 07a]. Le modèle représente la distribution des vecteurs de caractéristiques extraites à partir échantillons de textes de l'ensemble d'apprentissage. À la suite de l'étape d'apprentissage ou d'entraînement, un modèle représentant chaque classe (scripteur ou catégorie de scripteurs) est obtenu. Ce modèle sera utilisé lors de l'étape suivante (de classification) pour la classification d'échantillons inconnus.

Classification : Une fois que l'échantillon de texte en entrée du système de classification de scripteurs est représenté par un ensemble de caractéristiques, il peut être comparé avec les modèles de référence générés par l'étape précédente (étape d'apprentissage) afin de déterminer à quelle classe il appartient. Dans la reconnaissance de scripteurs, ces classes correspondent à des individus (scripteurs): chaque personne est modélisée comme un

générateur d'un ensemble de formes particulières d'écriture décrivant son style personnel d'écriture. Dans la classification de scripteurs en fonction de leur sexe, les classes correspondent à deux catégories différentes (masculine et féminine): chaque catégorie est modélisée comme un générateur d'un ensemble de formes particulières d'écriture décrivant cette catégorie particulière. Le processus d'attribution d'une classe à un échantillon d'écriture inconnu représente l'étape de "classification".

1.2.2. Catégorisation par dépendance du texte

Une deuxième classification à l'intérieur des catégories mentionnées dans la section 1.2.1 repose sur le niveau de dépendance au texte. Les systèmes de classification de scripteurs peuvent être catégorisés en deux grandes familles: les systèmes indépendants du texte et les systèmes dépendants du texte [PLA 89, BUL 07].

Les systèmes dépendants du texte sont plus contraints et exigent qu'un scripteur écrive un texte prédéfini particulier qui sera utilisé pour identifier (vérifier) son identité ou pour reconnaître son sexe. Les systèmes dépendants du texte sont très semblables aux systèmes de vérification de signatures manuscrites et emploient la comparaison entre différents caractères ou mots de contenu sémantique connu. Ces méthodes exigent donc la localisation et la segmentation antérieures d'informations appropriées. Ceci est habituellement exécuté interactivement par un utilisateur humain ou automatiquement par un algorithme de segmentation.

Si n'importe quel texte peut être utilisé afin d'établir (de vérifier) l'identité ou le sexe du scripteur, la tâche est dite indépendante du texte car aucune restriction relative au sujet du texte écrit ne peut être faite, la tâche est plus générale et, ainsi, plus difficile à accomplir. D'un point de vue utilisateur, il est important de mentionner qu'une quantité minimale d'écriture (par exemple, un paragraphe contenant quelques lignes de texte) est nécessaire afin de dériver des caractéristiques stables, peu sensibles au contenu du texte, à partir de l'échantillon d'écriture.

Chacun des deux types de systèmes (dépendant ou indépendant du texte) a ses propres avantages et inconvénients. Les systèmes dépendants du texte atteignent des performances très élevées en utilisant une quantité réduite de données, ce qui n'est pas possible pour les systèmes indépendants du texte. Cependant, ils sont plus enclins à la contrefaçon, car le texte de vérification est connu à l'avance. Pour le cas des systèmes indépendants du texte, la contrefaçon n'est pas un problème important car ces systèmes extraient, à partir d'un document manuscrit, les propriétés les moins fréquentes qui sont difficile à contrefaire [GUP 08]. L'utilisation d'un système dépendant ou indépendant du texte dépend entièrement du domaine d'application et de la disponibilité des données. Par exemple, dans le cas des applications légales, nous ne pouvons pas nous assurer que les données qui sont disponibles pour la vérification correspondent à l'écriture de référence, donc la seule alternative est l'utilisation d'un système indépendant du texte [GUP 08].

1.2.3. Catégorisation selon le mode d'acquisition des données

Une différence fondamentale entre les systèmes de classification de scripteurs est liée au moyen utilisé pour l'acquisition des données. Une distinction est faite entre les systèmes en-ligne (en temps réel, ou encore dynamique), où le tracé de l'écriture est récupéré pendant sa génération et les systèmes hors-ligne (en différé, ou encore statique) où la numérisation du tracé est effectuée après sa génération [BEN 04, SCL 07a, BEL 01]. Si seulement une image de l'écriture est disponible, la tâche est effectuée en différé (hors-ligne), par contre, si les données

temporelles et spatiales de l'écriture sont disponibles, la tâche est effectuée en temps réel (en ligne).

Les données temporelles et spatiales de l'écriture contiennent plus d'informations sur le style d'écriture d'un individu, parmi ces informations on peut citer : l'ordre des traits, la position des poser et lever de crayon, la pression, la vitesse,.....etc [CRE 98]. Ces données ne sont pas disponibles en mode hors-ligne, ainsi la classification de scripteurs en temps réel (en-ligne) est moins difficile que celle en différé (hors-ligne) et elle est censée fournir des résultats plus précis.

Dans le cas des systèmes hors ligne, le signal est une image en noir et blanc, en niveaux de gris ou en couleurs. Il est bidimensionnel, c'est une matrice représentant la luminosité des pixels, selon les axes x et y. Par rapport aux données en-ligne, les informations dynamiques, telles que l'ordre et le sens des traits, sont absentes. La segmentation est plus difficile, surtout pour les écritures liées. En effet, certaines segmentations détectables dans des tracés en-ligne ne le sont pas sur des tracés hors-ligne équivalents. De plus, la classification de scripteurs sur des documents hors-ligne nécessite davantage de prétraitements: binarisation, élimination des bruits, segmentation des blocs de texte en lignes de texte, puis en mots. L'épaisseur du tracé, inexistante pour les signaux en-ligne, dépend ici de l'instrument d'écriture et de la résolution de l'image.

Il est important de mentionner que des tentatives pour transformer l'un des deux types de données (en-ligne ou hors-ligne) en l'autre ont été rapportées dans la littérature. Convertir l'écriture en-ligne en une écriture hors-ligne est assez simple mais l'inverse est plus compliqué parce qu'il exige de reconstruire l'ordre temporel des différents traits individuels de l'image d'un texte manuscrit donné. Néanmoins, un certain nombre de solutions à ce problème ont été rapportées [SCL 07a]. Il est intéressant aussi de noter que certains chercheurs ont opté pour la combinaison des caractéristiques extraites à partir des deux types de signaux (en-ligne et hors-ligne) et ce, pour tirer profit des avantages de chacun d'eux, dans le but d'améliorer les performances de leurs systèmes d'identification du scripteur [CHB 11] et d'identification du sexe du scripteur [LIW 11]. La figure 1.10 montre des exemples de mots relevant des deux modes d'acquisition de l'écriture manuscrite. La figure 1.10 (a) montre une traduction de l'analyse du mouvement du tracé du mot "sage" par repérage des points importants alors que la figure 1.10 (b) présente l'image du mot cursif "dix" en représentant ses pixels par des carrés noirs de même taille.

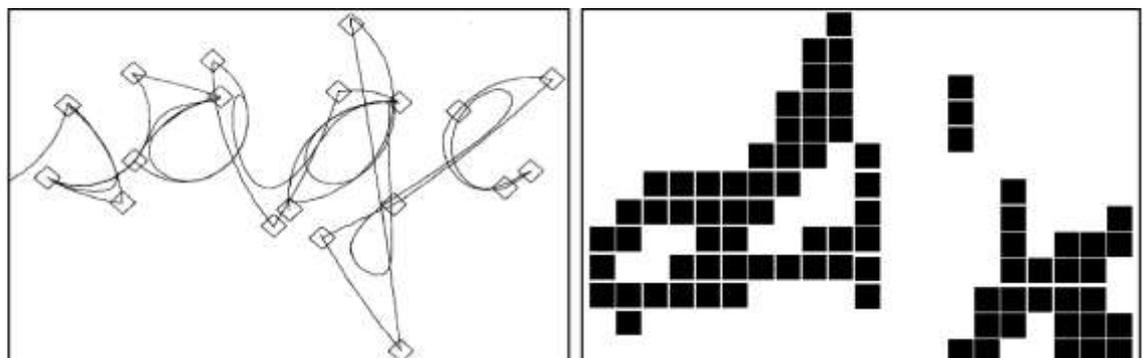


Figure 1.10. Les deux modes d'acquisition de l'écriture, (a) Tracé du mot "sage".
(b) Image du mot "dix" [BEL 01].

1.3. Evaluation de performances des systèmes de classification des scripteurs

Les performances d'un système de classification des scripteurs peuvent se mesurer principalement à l'aide de trois critères [SCH 08] : sa précision, son efficacité ou sa vitesse d'exécution ainsi que le volume de données à traiter. Nous nous concentrerons, dans cette section, sur le premier aspect. Comme nous l'avons déjà mentionné dans la section 1.2.1, l'identification et la vérification ont des modes opératoires différents, elles nécessitent donc des mesures de précision différentes que nous étudierons dans les sections suivantes.

1.3.1. Evaluation de l'identification

Pour l'identification d'un scripteur ou l'identification de son sexe, un échantillon de son écriture est comparé à toutes les écritures de référence connues du système (d'identification de scripteurs ou de reconnaissance de sexe du scripteur), résultant en un ensemble de mesures de similarité en entrée du processus de décision. Aussi, ce processus a pour tâche de rechercher la mesure de similarité maximale (ou minimale dans le cas de mesures de distance) et de désigner l'identité ou le sexe du scripteur considéré.

Dans ce contexte, la mesure des performances d'un système d'identification de scripteurs ou de reconnaissance de sexe est généralement donnée en termes de taux d'identification correcte [SCH 08]. Ce taux s'obtient par la formulation suivante :

$$\text{Taux d'identification correcte} = \frac{\text{Nombre de tests ayant mené à une identification correcte}}{\text{Nombre total de tests}}$$

Pour les systèmes d'identification de scripteurs, il est très utile aussi de savoir si le bon scripteur se trouve dans les N premiers. On calcule alors le score cumulé (cumulative Top-N correct-writer identification proportions) qui représente la probabilité que le bon scripteur se trouve parmi les N premiers [SCH 08]. Top-1 indique que le bon scripteur se trouve en première position dans la liste de scripteurs possibles retournée par le système d'identification de scripteurs, Top-10 correspond à la possibilité de trouver le bon scripteur parmi les 10 premières positions dans cette liste.

Dans le cas où il existe plusieurs instances, pour chaque scripteur, dans la base de données, les mesures classiques des systèmes de recherche d'information telles que la précision et le rappel peuvent être utilisées. Ces mesures s'obtiennent par les formulations suivantes :

$$\text{Précision} = \frac{\text{Nombre d'instances correctement retrouvées}}{\text{nombre total d'instances retrouvées}}$$

$$\text{Rappel} = \frac{\text{Nombre d'instances correctement retrouvées}}{\text{Nombre total d'instances qui auraient dû être retrouvées}}$$

1.3.2. Evaluation de la vérification

Tout système de vérification du scripteur peut faire deux types d'erreurs [SCH 08]. Il peut rejeter un scripteur légitime et dans ce premier cas on parle de faux rejet (false rejection). Il peut aussi accepter un imposteur et on parle dans ce second cas de fausse acceptation (false acceptance). La performance d'un système de vérification de scripteurs se mesure donc à son

taux de faux rejet (False Rejection Rate ou FRR) et à son taux de fausse acceptation (False Acceptance Rate ou FAR).

Pour une mesure de similitude X , les distributions $P_{\text{genuine}}(x)$ pour des distances intra-scripteur et $P_{\text{impostor}}(x)$ pour des distances inter-scripteurs peuvent être calculées en évaluant un grand nombre de paires de documents manuscrits. Par la suite, en changeant un seuil θ pour accepter ou rejeter des identités réclamées, les distributions de probabilités cumulées peuvent être obtenues, pour le cas d'une mesure de similitude X .

$$FAR(\theta) = \int_0^{\theta} P_{\text{impostor}}(x) dx$$

et

$$FRR(\theta) = \int_0^{\theta} P_{\text{genuine}}(x) dx$$

Ainsi, pour chaque valeur de θ , les probabilités d'accepter faussement un imposteur ou celle de rejeter faussement un utilisateur légitime peuvent être déterminées. Nous avons représenté, sur la figure 1.11, la distribution hypothétique des taux de vraisemblance qu'obtiendraient les utilisateurs légitimes et les imposteurs d'un système de vérification de scripteurs. Les FAR et FRR sont représentés en grisé. Idéalement, le système devrait avoir des FAR et FRR égaux à zéro. Comme ce n'est jamais le cas en pratique, il faut choisir un compromis entre FAR et FRR. Plus le seuil de décision θ est bas, plus le système acceptera d'utilisateurs légitimes, mais plus il acceptera aussi d'imposteurs. Inversement, plus le seuil de décision θ est élevé, plus le système rejettera d'imposteurs mais plus il rejettera aussi d'utilisateurs légitimes. Il est donc impossible, en faisant varier le seuil de décision, de faire diminuer les deux types d'erreurs en même temps. La courbe dite ROC (Receiver Operating Characteristic), illustrée par la figure 1.12 permet de représenter graphiquement la performance d'un système de vérification de scripteur pour les différentes valeurs de θ [SCH 08].

Le taux d'égale erreur (Equal Error Rate ou EER) correspond à un point particulier de la courbe ROC indiquant un taux identique d'erreurs de fausse acceptation et de faux rejet ($FAR = FRR$), c'est-à-dire graphiquement à l'intersection de la courbe ROC avec la première bissectrice. Il est fréquemment utilisé pour donner un aperçu de la performance d'un système de vérification de scripteurs. Cependant, il est important de souligner que le taux d'égale erreur (ERR) ne résume, en aucun cas, toutes les caractéristiques d'un système de vérification de scripteurs. Le seuil θ doit donc être ajusté en fonction de l'application ciblée : haute sécurité, basse sécurité ou compromis entre les deux.

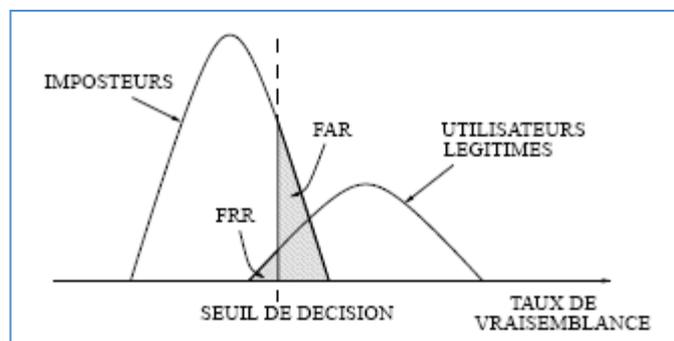


Figure 1.11. Distributions des taux de vraisemblance des utilisateurs légitimes et des imposteurs d'un système biométrique.

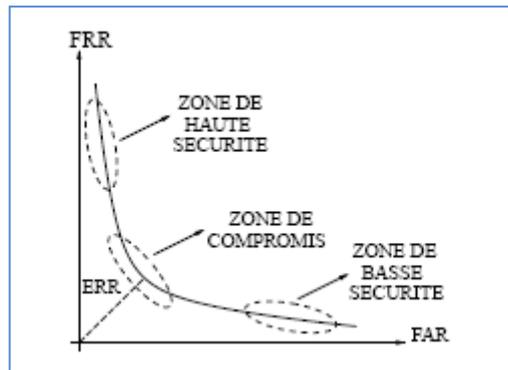


Figure 1.12. Courbe ROC pour un système de vérification de scripteurs.

1.4. Les applications possibles de la classification des scripteurs

La classification des scripteurs a des applications dans divers domaines, particulièrement celui de la reconnaissance biométrique qui a reçu, ces dernières années, une attention accrue due à la nécessité d'identifier certains individus. Les sections suivantes décrivent brièvement certaines de ces applications.

1.4.1. Systèmes adaptatifs de reconnaissance de l'écriture manuscrite

Un système de reconnaissance de l'écriture manuscrite ne peut pas atteindre les performances souhaitées puisqu'il est difficile d'apprendre toutes les variations dues aux différents styles d'écriture [GUP 08]. Cependant, un système adaptatif (de reconnaissance de l'écriture manuscrite) pour un ensemble d'individus doit apprendre les traits personnels de l'écriture des scripteurs considérés. Puisque le style d'écriture d'un individu présente certaines particularités, le taux de reconnaissance sera probablement meilleur. L'utilisation d'un processus identifiant le scripteur concerné [GUP 08, SER 03], ou son genre comme étape de prétraitement à la reconnaissance de l'écriture manuscrite, permettrait d'effectuer, de manière globale, une approche multi-scripteurs tout en s'adaptant automatiquement à chaque individu ou aux caractéristiques du style représentant son genre (masculin ou féminin).

1.4.2. Examen légal de documents

L'examen légal de documents (Forensic Document Examination) a pour objectif principal de déterminer l'identité ou le sexe du scripteur d'un document manuscrit [BUL 07, SCL 07a]. Ses applications incluent la vérification de l'authenticité et/ou la falsification d'écrits, de documents ou de signatures ou la détermination de l'auteur (par exemple, une demande de rançon, une lettre frauduleuse, une note de suicide suspecte ou une lettre de menaces).

L'examen légal de documents est, en grande partie, basé sur l'analyse manuelle par des experts humains. Cette analyse est une procédure fastidieuse pour les examinateurs légaux de documents, par conséquent, des systèmes informatisés pour l'analyse des écritures manuscrites pourraient servir d'outils précieux dans l'examen légal de documents. Ces systèmes prennent comme entrées des images scannées de documents manuscrits et emploient des méthodes de traitement d'images et de reconnaissance de formes. Bien entendu, les résultats de ces

systemes peuvent ne pas être acceptés comme preuves dans une affaire judiciaire et l'intervention d'experts humains est inévitable. Néanmoins, ils se sont toujours avérés très utiles pour accélérer considérablement le processus. Certains de ces systèmes sont présentés dans la section 1.5.

1.4.3. Reconnaissance biométrique

La reconnaissance biométrique dénote la reconnaissance automatique des individus en se basant sur leurs caractéristiques physiologiques ou comportementales [BEN 04]. Les caractéristiques physiologiques sont basées sur une propriété physique du corps humain (telle que le visage, les empreintes digitales, l'iris, l'ADN,etc.). Les caractéristiques comportementales utilisent les traits du comportement d'une personne pour l'identification (telles que la voix, la démarche, la dynamique de frappe au clavier, la signature, l'écriture,..... etc.). Puisqu'on suppose que ces caractéristiques sont uniques à chaque personne, elles sont plus fiables et plus distinctives que les techniques basées sur la possession de connaissances (mot de passe ou code PIN) ou celles basées sur la possession d'un jeton physique (carte, badge, document ou clef) pour différencier entre une personne autorisée et un imposteur [SCL 07a]. Les caractéristiques physiologiques sont des modalités fortes qui donnent des performances d'authentification plus élevées que les caractéristiques comportementales, cependant, elles sont difficiles à acquérir et souvent socialement mal admises.

1.4.4. Bibliothèque numériques

Les bibliothèques numériques représentent un ensemble de ressources électroniques liées avec des techniques automatiques permettant la création, la recherche et l'utilisation de l'information [SCL 07a]. Dans ce sens, elles sont une extension et un perfectionnement des systèmes de stockage et de récupération d'informations qui utilisent des données numériques dans n'importe quel milieu. Les bibliothèques numériques comprennent des données et des métadonnées qui décrivent différents aspects des données stockées. Un aspect intéressant des bibliothèques numériques est de préserver les documents anciens manuscrits. Le patrimoine culturel est préservé par la numérisation et ensuite mis à la disposition des utilisateurs sous une forme numérique [SCL 07a]. Une application possible de l'identification et de la vérification de scripteurs, dans ce contexte, est la récupération des documents historiques qui n'ont pas encore été attribués à un scripteur ou la validation de l'identité de l'auteur d'un de ces documents.

1.4.5. Salles de réunions intelligentes

L'objectif des salles de réunion intelligentes est d'automatiser des tâches standards, généralement effectuées par l'homme, dans une réunion, et à développer des techniques qui aident à obtenir les informations nécessaires d'une réunion [SCL 07a, LIW 06]. Pour enregistrer une réunion, une salle de réunion intelligente est équipée avec des appareils d'enregistrement audio-visuels synchronisés. Les recherches sur les salles de réunion intelligentes visent à élaborer des méthodes automatiques pour capter, stocker, structurer les données acquises afin d'y naviguer et d'y effectuer des recherches. Une tâche importante dans une salle de réunion intelligente est de capturer l'écriture à partir d'un tableau blanc (whiteboard) au cours d'une réunion. Cette tâche consiste également à identifier l'auteur d'un texte écrit sur un tableau blanc. Résoudre ce problème permet à ses utilisateurs d'étiqueter l'écriture avec l'identité du

scripteur. En outre, elle permet de valider les résultats des systèmes d'identification de personnes utilisant des données vidéo ou audio.

1.4.6. L'intelligence ambiante

L'intelligence ambiante (*Ambient Intelligence*) vise à créer des environnements qui sont sensibles et réceptifs à la présence de personnes [SCL 07a]. L'idée est d'améliorer la qualité de vie en créant l'atmosphère désirée et en fournissant les fonctionnalités souhaitées par des systèmes intelligents et personnalisés et des services de la vie quotidienne. Deux concepts clés de l'intelligence ambiante sont la sensibilité au contexte et la personnalisation. Dans ce contexte, un dispositif électronique permettant de capter l'écriture manuscrite identifierait automatiquement son utilisateur et utiliserait alors un système optimisé de reconnaissance de l'écriture manuscrite conçu en fonction de l'écriture de l'utilisateur.

1.5. Systèmes existants de la classification des scribes

L'analyse des écritures manuscrites pour la classification des scribes a fait l'objet d'une forte collaboration entre des organismes étatiques et privés tels que la police judiciaire, les ministères de justice, les cabinets d'experts en écritures ainsi que ceux des avocats avec les laboratoires de recherche scientifique. Cette collaboration a donné naissance à des systèmes capables de traiter et d'analyser de grandes bases de documents manuscrits. Dans les sections suivantes, nous présentons un aperçu des principaux systèmes existants.

1.5.1. Le système WANDA

WANDA est un outil logiciel utilisé pour la comparaison des écritures manuscrites en se basant sur des caractéristiques bien définies et sur les connaissances et l'expérience d'un expert en écriture [VAN 03]. Ce système permet la numérisation des échantillons d'écriture grâce à un dispositif optique, tel qu'un scanner ou une caméra numérique ou à l'aide d'un stylo ou une tablette électronique. WANDA comporte un ensemble d'outils pour le traitement des images de documents, l'extraction de caractéristiques de manière manuelle, semi-automatique ou automatique, la gestion de la base de données, et enfin l'identification de scribeur.

WANDA dispose d'un module de mesures interactives, appelé WAM (voir figure 1.13). A travers ce module, des caractéristiques manuelle peuvent être extraites, elles incluent la hauteur, l'épaisseur, l'inclinaison, l'interligne moyen en plus de mesures caractérisant certains aspects sur la forme des boucles. Il est aussi possible d'extraire des caractéristiques de manière complètement automatique, on peut citer par exemple : les fonctions d'autocorrélation qui servent à détecter la présence de régularités dans l'écriture, les distributions des longueurs de segments horizontaux et verticaux, les distributions de la densité d'encre présente en fin des traits, les distributions des directions des contours ainsi que les distributions des charnières des contours [FRA 04].

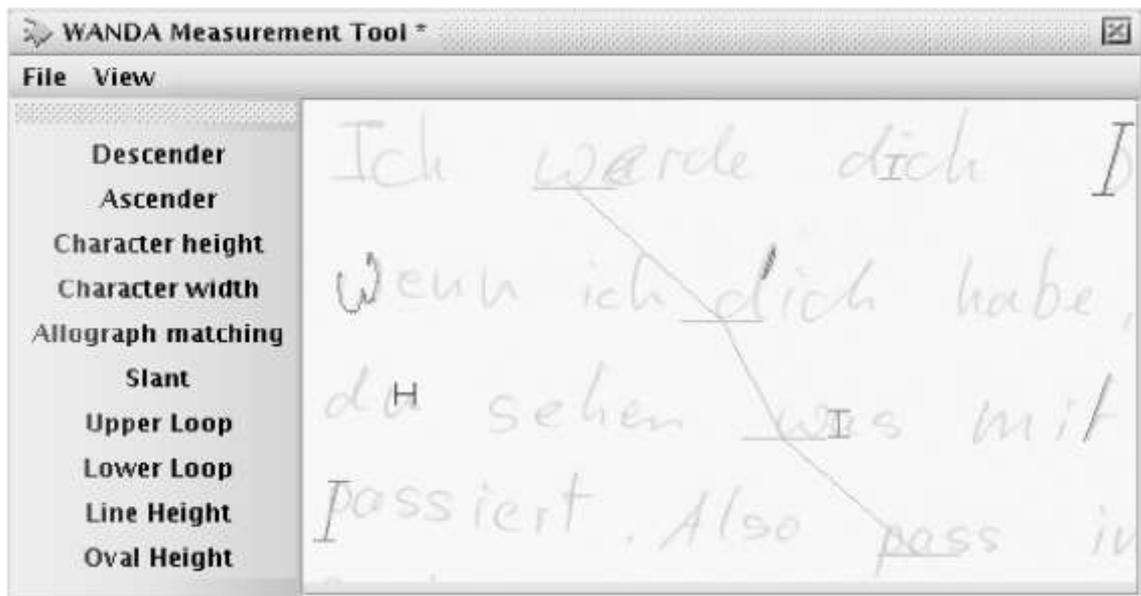


Figure 1.13. Fenêtre principale du module de mesures interactives (WAM) du système WANDA [FRA 04].

1.5.2. Le système CEDAR-FOX

CEDAR-FOX est un système informatique [SRI 07] permettant d'analyser des écritures manuscrites pour des applications légales. Ce système a été réalisé par une équipe de recherche au Center of Excellence for Document Analysis and Recognition (CEDAR) de l'Université de Buffalo (New York, USA) en collaboration avec le National Institute of Justice (NIJ). CEDAR-FOX dispose de diverses fonctions qui sont utilisées par les experts en écriture, parmi ce fonctions, on peut citer la reconnaissance de l'écriture, la reconnaissance de scripteurs, la vérification de signatures, le traitement d'images, la segmentation de l'écriture et plusieurs modalités de recherche.

En tant que système de gestion de documents pour l'analyse légale et judiciaire, CEDAR-FOX offre aux utilisateurs trois fonctionnalités majeures. Il peut être utilisé comme un système d'analyse de documents ou pour créer une bibliothèque numérique de documents manuscrits légaux, et enfin, comme un système de gestion de base de données pour la recherche de documents et la reconnaissance de scripteurs. Comme c'est un système interactif d'analyse de documents, une interface graphique est fournie, elle permet de numériser ou charger une image d'un document manuscrit. Le système va d'abord extraire automatiquement des caractéristiques en se basant sur des techniques de traitement et de reconnaissance d'images de documents manuscrits. L'utilisateur peut alors utiliser les outils fournis pour effectuer l'examen des documents et en extraire des métriques. Ces outils incluent des fonctionnalités telles que la sélection de l'image, l'amélioration de l'image et l'affichage du contour. Lors de la vérification du scripteur, lorsqu'un document connu est comparée au document considéré, CEDAR-FOX analyse ce dernier et calcule un score de similarité qui est utilisé pour décider si les deux documents sont du même scripteur ou non.

En outre, le système a la capacité d'apprendre à partir de documents connus pour la tâche de vérification du scripteur. Le système nécessite, au minimum, quatre échantillons du même scripteur afin qu'il puisse être entraîné sur l'écriture du scripteur. Après l'entraînement du système, les documents en question peuvent ensuite être comparés aux échantillons connus du scripteur pour des fins de vérification. CEDAR-FOX utilise une fonctionnalité de traitement par

lots pour l'identification du scripteur à partir d'un ensemble de documents connus, le système peut trouver les documents les plus similaires au document considéré.

CEDAR-FOX dispose de deux modes de vérification de signature. Dans le premier mode, une signature connue et une autre inconnue (questioned signature) sont comparées; le système génère alors un score indiquant si la signature inconnue considérée est authentique ou non. Le second mode de vérification de signature correspond à la capacité d'apprendre à partir des échantillons de signature connue. Le système recommande qu'un minimum de quatre échantillons connus doit être utilisé. Après l'entraînement du système sur la signature du scripteur connu, les signatures inconnues peuvent être comparées et un score est généré, il indique si les signatures inconnues considérées sont susceptibles d'être authentiques ou non.

CEDAR-FOX a été testé par l'agence des services frontaliers du Canada, les services secrets américains, le bureau d'enquête américain (FBI) et est actuellement en cours d'évaluation par les services de police de San Diego. En outre, le système a été autorisé par le *Netherlands Forensic Institute* (NFI) et par l'Agence des services frontaliers du Canada. Une version d'essai de CEDAR-FOX est disponible au téléchargement [CEDAR]. Une version arabe du CEDAR-FOX, connue sous le nom CEDARABIC, est également disponible. La figure 1.14 montre la fenêtre de l'outil de comparaison de deux documents manuscrits du système CEDAR-FOX

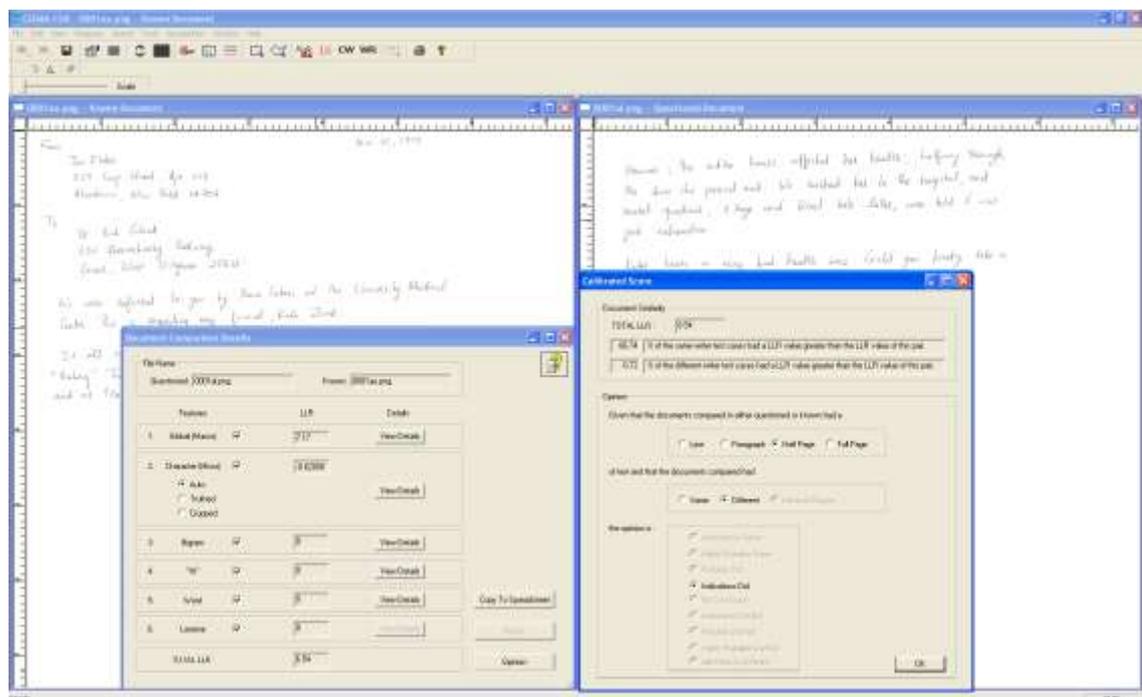


Figure 1.14. La fenêtre principale de l'outil de comparaison de deux documents manuscrits du système CEDAR-FOX [SRI 07].

1.5.3. Le système FLASH ID

Le *Forensic Language-Independent Analysis System for Handwriting Identification* (FLASH ID), est un système entièrement automatisé et fonctionnel qui utilise l'écriture manuscrite en tant qu'identifiant biométrique. FLASH ID est un logiciel utilisé par le bureau fédéral de recherche (FBI) des Etats Unis d'Amérique (USA). FLASH ID extrait, d'une manière complètement automatique, des données graphiques à partir de documents manuscrits, puis analyse ces données en utilisant des méthodes statistiques connues et classe ensuite les documents en

fonction de la ressemblance des documents manuscrits. Le système extrait les caractéristiques graphiques à partir des caractères individuellement et ces caractéristiques peuvent être dépendantes du scripteur [FIS 12]. FLASH ID est capable d'agir sur n'importe quel échantillon de document manuscrit inconnu et retourne la valeur la plus proche dans sa base de données de documents manuscrits, il fournit ainsi la correspondance la plus proche de l'échantillon du document manuscrit considéré. La technologie FLASH ID est indépendante de la langue. L'interface de la station de travail du système FLASH ID au laboratoire du bureau fédéral de recherche (FBI) est présentée sur la figure 1.15.

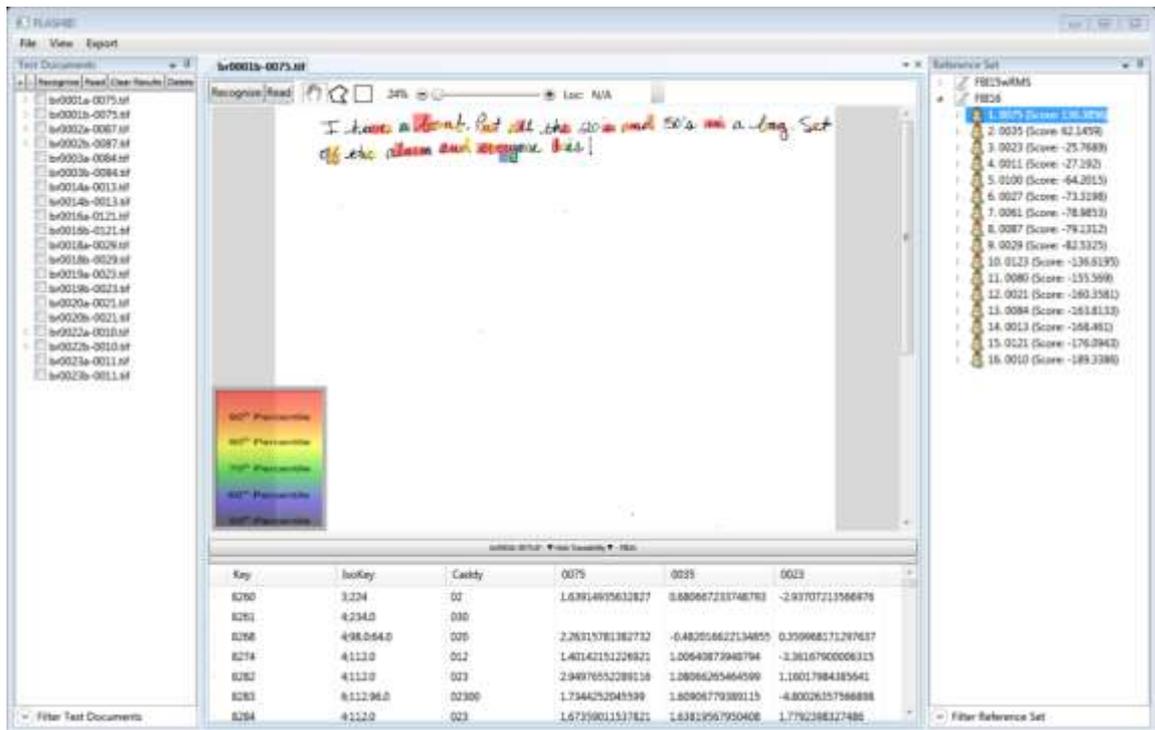


Figure 1.15. Interface de la station de travail du système FLASH ID au laboratoire du FBI [FIS 12].

1.5.4. Le système SCRIPT

Le système Script est le fruit d'une importante collaboration entre le NIFO (Netherlands Institute for Forensic Examinations and Research) et l'Université de Delft et le TNO (Dutch Organization for Applied Scientific Research). Ce système avait pour objectifs de remplacer le processus d'analyse manuelle des écritures manuscrites par une technique assistée par ordinateur, de sauvegarder les informations importantes dans un fichier et de retrouver rapidement tous les échantillons similaires à l'échantillon en entrée du système.

Le système SCRIPT utilise des attributs complètement indépendants pour la caractérisation des documents manuscrits, cette indépendance existe aussi vis-à-vis du déguisement, de l'effet de l'alcool ou de la drogue ainsi que du système d'éducation. Ces attributs sont basés principalement sur des mesures d'angles et de distances. Il est important de noter que le système SCRIPT n'a pas été développé spécialement afin qu'il soit utilisé par des experts en écritures manuscrites [MAR 07].

A la fin de cette section, il est plus que nécessaire de noter que les performances des différents systèmes qui viennent d'être présentés ne sont pas complètement publiées ; le risque d'erreur

existe, une erreur qui peut devenir cruciale dans certains cas. Les systèmes développés ne sont donc pas destinés à un utilisateur ordinaire mais à un expert en écriture, ceci afin que les résultats d'analyse de l'écriture puissent être validés finalement par l'expert. Il est aussi important de noter qu'aucune tentative n'a été faite pour le développement d'un système pour la reconnaissance du genre d'un scripteur, ceci est dû à la jeunesse de ce domaine qui reste loin des préoccupations des informaticiens pour des raisons inconnues.

1.6. Principales bases de données

Les bases de données d'images de textes manuscrits constituent le point central dans l'évaluation des systèmes d'analyse et de reconnaissance des écritures manuscrites. Elles fournissent un moyen efficace pour l'unification et la comparaison des travaux effectués au sein des différentes équipes de recherche à travers le monde. Dans la section suivante, nous présentons les principales bases de données utilisées pour la classification des scripteurs.

1.6.1. CEDAR Database

La base de données CEDAR a été développée à l'Université de Buffalo, et 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 [SRI 02]. Chaque scripteur a recopié trois exemplaires de la lettre CEDAR [SRI 02], scannée avec une résolution de 300 ppp. Cette lettre est un document qui contient 156 mots, à partir d'un lexique de 124, qui inclut tous les caractères (lettres et chiffres). Le document a été soigneusement conçu pour contenir chaque lettre de l'alphabet en majuscule en position initiale d'un mot et en minuscule en position initiale, intermédiaire et finale d'un mot.

1.6.2. IAM Database

La base de données IAM [MAR 02] 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 public, 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].

1.6.3. RIMES Database

RIMES [AUG 06] est une base de données relativement nouvelle, elle 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].

1.6.4. BFL Database

La base de données BFL (Brazilian Forensic Letter) [FRE 08] est composée de 315 scripteurs, avec trois échantillons par scripteur, soit un total de 945 images. Les échantillons ont été collectés auprès des étudiants du premier cycle en trois sessions différentes sur une période d'un mois. Les textes ont été recueillis sur une feuille blanche de format A4, sans carreaux ni lignes, puis numérisés en niveaux de gris à 300 ppp (3760×2448 pixels). Chaque scripteur a été autorisé à utiliser son propre stylo, ce qui signifie que de nombreux modèles stylos ont été utilisés. Le texte est concis (131 mots en portugais), et complet dans le sens où il contient tous les caractères (lettres et chiffres) et certaines combinaisons de caractères d'intérêt. Cela le rend approprié aussi pour la classification des écritures en mode indépendant du texte. Il est à noter que cette base de données a fait l'objet de travaux sur la reconnaissance de scripteurs [BER 13].

1.6.5. CVL Database

La base de données CVL [KLE 13] est une base de données publique créée en 2013, 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 contribué 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 œuvres littéraires. Cette base de données a été utilisée pour la reconnaissance de scripteurs par Fiel et Sablatnig [FIE 13].

1.6.6. KHATT Database

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. Elle a été rendue publique en septembre 2012. 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.

1.6.7. IFN/ENIT Database

La base de données IFN/ENIT [PEC 02] comprend des noms manuscrits arabes de villes et villages tunisiens rassemblés auprès de 411 scripteurs, les images sont binaires avec une résolution de 300 ppp. Cette base d'images a été conçue pour l'apprentissage et le test des systèmes de reconnaissance de l'écriture arabe manuscrite [MAR 05]. Les données de la base IFN/ENIT ont été employées également pour la reconnaissance de scripteurs [BUL 07a, DJE 08]. Cette base de données, par sa taille et sa disponibilité (gratuite pour des travaux académiques) s'est imposée comme la base de référence des travaux en reconnaissance de scripteurs arabes et en reconnaissance d'écriture arabe manuscrite.

1.6.8. QUWI Database

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 systèmes 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'une compétition sur la prédiction du genre à partir de documents manuscrits qui s'est déroulée dans le cadre de la conférence ICDAR 2013 [HAS 13].

1.6.9. CASIA Database

CASIA [LIU 11] est une base de données qui contient des échantillons de caractères chinois isolés ainsi que des textes chinois manuscrits. Ces échantillons ont été produits par 1020 scripteurs différents en utilisant un stylo électronique. Les échantillons de la base de données sont divisés en six ensembles, trois pour les caractères isolés et trois pour les textes manuscrits. La base de données de textes manuscrits contient environ 5090 pages et 1,35 million échantillons de caractères. La base de données CASIA peut être utilisée pour la reconnaissance de caractères et de textes manuscrits ainsi que la reconnaissance de scripteurs.

1.6.10. CVCMUSCIMA Database

La base de données CVCMUSCIMA [FOR 12, DAL 08] a été spécialement conçue pour être utilisée en détection et suppression des lignes de portée à partir des images de partitions musicales. Cette base de données contient 20 images de partitions musicales pour chacun des 50 scripteurs qui ont participé à sa création. Les partitions musicales ne présentent pas de rotations importantes ou des changements d'échelle et les scripteurs ont utilisé le même stylo. Afin qu'elle puisse être utilisée en reconnaissance de scripteurs, les lignes de portée des documents de la base de données ont été enlevées. Il est à noter que la base CVCMUSCIMA a fait l'objet de travaux sur la reconnaissance de scripteurs [GOR 13], elle a été utilisée aussi lors des compétitions qui ont été organisées dans le cadre des conférences ICDAR 2011 [FOR 11] et ICDAR 2013 [VIS 13].

1.6.11. FIREMAKER Database

La base de FIREMAKER [SCH 00] contient des documents manuscrits issus de 250 scripteurs néerlandais, sachant que la plupart de ces scripteurs sont des étudiants. Chacun des scripteurs a été invité à écrire 4 pages différentes. Sur la page 1, ils ont été invités à copier un texte donné, sur la page 2, ils devaient écrire, avec leurs propres mots, la description d'un dessin animé donné. Sur les pages 3 et 4, les scripteurs ont été invités à écrire des échantillons majuscules en style forgé. Cette base de données a fait l'objet de travaux sur la reconnaissance de scripteurs [BUL 03].

1.6.12. CEDARABIC Database

La base de données CEDARABIC [BAL 06] comprend 10 scripteurs différents. Chacun d'entre eux a écrit 10 pages complètes différentes de texte manuscrit. Chaque page comprend entre 150 et 200 mots, soit un total de 100 documents contenant environ 20000 mots. Les documents sont scannés en 300 dpi. Une annotation complète des documents est faite manuellement : segmentation en mots, mais aussi séquence des lettres arabes, ainsi que la prononciation et la traduction en anglais. Cette base de données est donc particulièrement adaptée pour des travaux sur le repérage de mots (word spotting) [SRI 05] ainsi que la reconnaissance de scripteurs [SRI 08].

Le tableau 1.1 présente une synthèse des principales bases de données d'images de documents manuscrits utilisées pour la classification des écritures manuscrites, plus particulièrement, pour la classification des scripteurs à partir de leurs écritures manuscrites. Il est intéressant de noter que la plupart de ces bases de données sont disponibles gratuitement pour effectuer des recherches académiques, parmi les 12 bases de données présentées, les bases QUWI [ALM 12] et CEDAR [SRI 02] sont les seules qui contiennent des informations sur les catégories démographiques des scripteurs (genre, âge, gaucher ou droitier) et, par conséquent, elles peuvent être utilisées pour l'évaluation des systèmes de classification des scripteurs en fonction de leurs catégories démographiques.

Base de données	Nombre d'échantillons	Nombre de scripteurs	Infos genre disponibles	Base Publique	langue
CEDAR Database[SRI 02]	4701	1567	Oui	Non	Anglais
IAM Database [MAR 02]	1539	657	Non	Oui	Anglais
RIMES Database [AUG 06]	12600	1300	Oui	Non	Français
BFL Database [FRE 08]	945	315	Non	Oui	Portugais
CVL Database [KLE 13]	2163	311	Non	Oui	Anglais et Allemand
KHATT Database [MAH 12]	4000	1000	Non	Oui	Arabe
IFN/ENIT Database [PEC 02]	2055	411	Non	Oui	Arabe
QUWI Database [ALM 12]	4068	1017	Oui	Non	Arabe et Anglais
CASIA Database [LIU 11]	5090	1020	Non	Oui	Chinois
CVC-MUSCIMA Database [FOR 12]	1000	50	Non	Oui	Partitions musicales
FIREMAKER Database [SCH 00]	1000	250	Non	Oui	Néerlandais
CEDARABIC Database [BAL 06]	100	10	Non	Non	Arabe

Tableau 1.1 - Synthèse des principales bases de données d'images de documents manuscrits.

1.7. Conclusion

Dans ce chapitre, nous avons présenté dans un premier temps, l'écriture qui représente l'entité de base sur laquelle s'articule le travail de cette thèse, nous avons ensuite décrit les facteurs

causant des variabilités à l'écriture manuscrite ainsi que les différents types de variations nécessaires afin qu'un système de classification de scripteurs puisse identifier un individu, vérifier son identité ou l'attribuer à une catégorie démographique donnée.

Après avoir opté pour une catégorisation des systèmes de classification de scripteurs selon la tâche à effectuer (identification ou vérification), la dépendance par rapport au contenu textuel des échantillons d'écriture à utiliser (dépendant ou indépendant) ainsi que le mode adopté pour l'acquisition de l'écriture, nous avons présenté les mesures nécessaires pour évaluer les performances de n'importe quel système de classification de scripteurs. Ces mesures seront utilisées dans les chapitres suivants pour étudier les performances sur certaines des bases de documents manuscrits présentées dans ce chapitre.

Dans le chapitre suivant, nous allons présenter un état de l'art sur les différentes approches proposées dans la littérature pour la classification de scripteurs.

APPROCHES DE CLASSIFICATION DE SCRIPTEURS: UN ETAT DE L'ART

Ce chapitre est dédié à la description de l'état de l'art dans le domaine de l'analyse de l'écriture manuscrite pour la classification de scripteurs. Nous nous concentrons dans ses premières sections sur la présentation des principaux travaux de recherche dans le domaine en les classant suivant les caractéristiques utilisées (locales, globales ou la combinaison des deux). Ensuite, les diverses compétitions dans le domaine de la classification de scripteurs sont présentées et nous terminons le chapitre par une comparaison des différents travaux du domaine en utilisant, comme critères de comparaison, la taille de la base de données utilisée, les caractéristiques choisies, la taille des échantillons ainsi que le script considéré.

2.1. Introduction

La classification de scripteurs est un domaine de recherche qui a attiré l'attention des chercheurs depuis quatre décennies, et reste encore un domaine de recherche attractif et très ouvert. Beaucoup de méthodes utilisées dans les domaines 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, ces dernières années, de bases de données publiques de grande taille, ont permis de mettre au point de nouvelles méthodes de plus en plus complexes et, par conséquent, les performances 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 dépendant 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 grandes et sans aucune restriction en matière de contenu textuel des échantillons. Durant la dernière décennie, les travaux dans le domaine se sont multipliés, les types de caractéristiques proposées ainsi que les scripts considérés se sont diversifiés et les bases de données employées pour l'évaluation des systèmes développés se sont élargies.

Depuis 2011, des compétitions scientifiques ayant pour objectif l'évaluation des systèmes de classification de scripteurs ont été mises en place dans le cadre de conférences très spécialisées telles que ICDAR et ICFHR. Les conclusions de ces compétitions sont très intéressantes et mettent en évidence de nouveaux problèmes qui accompagnent la recherche dans le domaine de la classification de scripteurs, à savoir que les performances de ces systèmes sont sensibles à la taille des bases de données et des échantillons utilisés pour l'évaluation et que la tâche de détermination du sexe d'un scripteur a encore besoin d'amélioration.

Dans le cadre de notre travail, et après avoir présenté les concepts de base liés à classification de scripteurs dans le chapitre 1, nous nous concentrons, dans ce chapitre, 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. La première est dédiée aux méthodes utilisant des caractéristiques locales (voir section 2.2), la deuxième présente les méthodes basées sur une analyse globale des textes représentant les scripteurs (voir section 2.3) alors que la troisième résume les travaux combinant les deux premiers types de caractéristiques afin de tirer profit des avantages de chacun d'eux dans l'objectif d'améliorer les performances des systèmes de classification de scripteurs (voir section 2.4).

2.2. Analyse locale pour la classification de scripteurs

Nous présentons, dans cette section, 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.

2.2.1. Travaux de Audrey Seropian (2003)

Audrey Seropian [SER 03] a effectué une étude sur l'extraction de caractéristiques fractales de l'écriture, elle propose un système utilisant la compression/décompression fractale. L'identification de chaque scripteur repose sur l'extraction d'un ensemble de caractéristiques qui doivent être propres à l'auteur du document. Les propriétés d'autosimilarité dans l'écriture sont exploitées. Pour cela, des formes invariantes caractérisant l'écriture d'un scripteur sont extraites (voir figure 2.1) au cours d'un traitement proche du processus de la compression fractale. Ces formes sont ensuite organisées dans une base de référence permettant d'analyser une écriture inconnue dans un processus d'appariement ou mise en correspondance (Pattern Matching). Les résultats de l'analyse sont évalués par le rapport signal/bruit qui permet, en fonction d'un ensemble de bases de référence, d'identifier le scripteur d'un texte dont on cherche l'identité de l'auteur. Sur une vingtaine de scripteurs, un taux d'identification de plus de 85% a été obtenu.

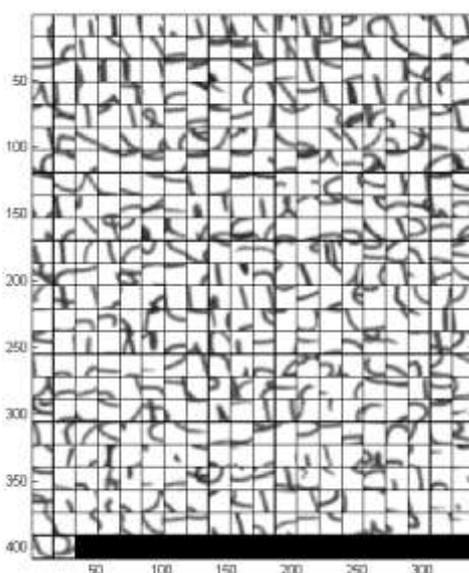
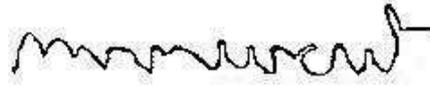


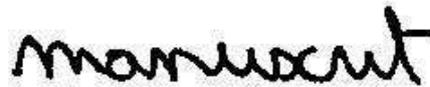
Figure 2.1. Exemple de formes invariantes associées à un scripteur.

2.2.2. Travaux de Bensefia & al (2004)

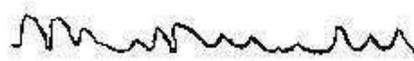
Le concept d'invariants de scripteurs introduit par Ali Nosary [NOS 99] a été développé par Bensefia et al dans [BEN 04] en proposant un système d'identification du scripteur basé sur la mise en correspondance directe des graphèmes extraits des écritures à comparer. L'idée est que la redondance des formes dans une écriture, définie au sens des invariants du scripteur, permet de compresser l'écriture tout en maintenant de bonnes performances d'identification.



(a)

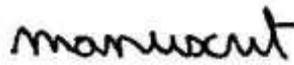


(b)

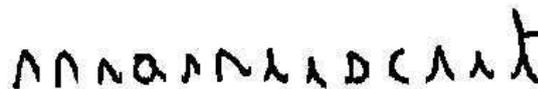


(c)

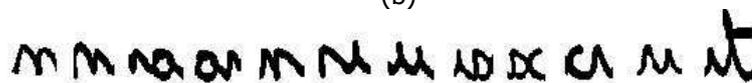
Figure 2.2. Illustration des contours supérieurs et inférieurs de la composante connexe "manuscrit", (a) Contour supérieur, (b) Composante connexe originale, (c) Contour inférieur.



(a)



(b)



(c)



(d)

Figure 2.3. Illustration de la segmentation de la composante connexe "manuscrit" en, (a) graphèmes, (b) bigrammes et (c) trigrammes.

Le système utilise trois étapes traditionnelles, prétraitements, extraction de caractéristiques et identification. Les composantes connexes du document sont d'abord extraites et analysées afin d'éliminer certains graphiques comme les ratures qui ne caractérisent pas l'écriture. Les autres composantes sont ensuite segmentées en graphèmes en se basant sur l'analyse de leurs

contours supérieurs et inférieurs (voir figure 2.2) [NOS 99]. Les auteurs introduisent également le concept de bigrammes (trigrammes) obtenu par la concaténation de deux (trois) graphèmes adjacents (voir figure 2.3).

Une fois que les graphèmes sont obtenus, ceux qui sont morphologiquement semblables sont regroupés dans le même groupe en utilisant un algorithme de regroupement séquentiel simple et rapide qui n'a pas besoin de connaître a priori le nombre de clusters.

Deux graphèmes x et y sont comparés en utilisant la corrélation comme une mesure de similarité [BEN 04]:

$$sim(x, y) = \frac{n_{11}n_{00} - n_{10}n_{01}}{\sqrt{(n_{11} + n_{10})(n_{01} + n_{00})(n_{11} + n_{01})(n_{10}n_{00})}}$$

n_{ij} étant le nombre de pixels pour lesquels les deux graphèmes x et y ont les valeurs i et j , respectivement, au niveau des positions de pixels correspondants.

Les groupes obtenus sont sensibles à l'ordre dans lequel les graphèmes sont présentés au cours de l'opération de regroupement. Pour remédier à ce problème, l'algorithme du regroupement séquentiel est exécuté plusieurs fois avec une sélection aléatoire des graphèmes.

Enfin, seuls les graphèmes qui sont toujours regroupés ensemble lors des différentes itérations de la procédure du regroupement, sont retenus pour constituer un groupe (voir figure 2.4), appelé groupe d'invariant.

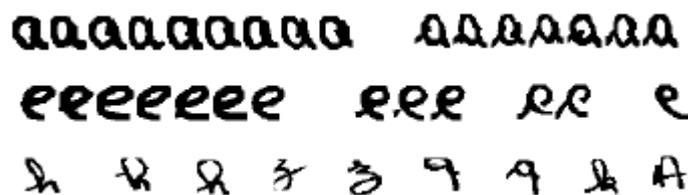


Figure 2.4. Exemples de groupes d'invariants obtenus sur une page d'écriture.

Le système proposé par les auteurs a été évalué sur deux bases de données différentes : La base PSI (Une base de données de 88 scripteurs différents auxquels il a été demandé de recopier une lettre au choix parmi deux proposées, chacune d'elles étant composée de 107 et 98 mots respectivement) ainsi qu'un sous ensemble de 150 scripteurs de la base de données IAM [MAR 02]. Les taux d'identification enregistrés sont de l'ordre de 93% et 87% pour les bases PSI et IAM respectivement. Au niveau bigramme, les performances du système en utilisant la base de données IAM se dégradent significativement, alors qu'elles s'améliorent pour atteindre les 96% d'identifications correctes sur la base de données PSI. Une dégradation très remarquable des performances du système a été enregistrée sur les deux bases de données (PSI et IAM) en utilisant des trigrammes.

2.2.3. Travaux de Pervouchine & al (2007)

Pervouchine et al. [PER 07] ont proposé une méthode basée sur un ensemble de caractéristiques structurelles correspondant à celles utilisées par les examinateurs légaux de documents, ces caractéristiques sont extraites à partir des squelettes des lettres "d", "y", "f" ainsi que du squelette du graphème "th". La plupart de ces caractéristiques sont représentées schématiquement dans la figure 2.5.

Le choix des auteurs est motivé par le fait que ces trois lettres et ce graphème sont les plus fréquents dans la langue anglaise, afin d'établir la puissance d'individualité de chacune des caractéristiques structurelles utilisées: taille, largeur, allongement, présence de la boucle, structure de la barre de la lettre t, structure de la tige de la lettre h, épaisseur de traits, nombre de traits,..., etc.

L'évaluation de l'efficacité de caractéristiques proposées a été effectuée sur un total de 21000 images de caractères extraites manuellement à partir de 600 documents issus de 200 scripteurs différents de la base CEDAR (voir section 1.6.1). Dix échantillons du caractère "d", huit échantillons du caractère "y", huit échantillons du caractère "f" et neuf échantillons du graphème "th" ont été extraits de chacun des 600 documents, ce qui fait un total de 30, 24, 24 et 27 échantillons des caractères correspondants par scripteur.

Les auteurs ont effectué une analyse du pouvoir discriminant des différentes caractéristiques. La classification a été réalisée à l'aide d'un réseau de neurones et les ensembles des caractéristiques optimales ont été déterminés en utilisant les algorithmes génétiques. Cette étude a montré que la plupart des caractéristiques structurelles proposées possèdent un pouvoir discriminant important, ce qui justifie leur utilisation dans l'analyse légale de documents. Les résultats montrent également que le graphème "th" possède un pouvoir discriminant nettement plus élevé que les trois autres caractères.

D'après les auteurs, les performances atteintes sont assez intéressantes et elles sont de l'ordre de 16% pour le caractère "d", 20% pour le caractère "y", 26% pour le caractère "f", 36% pour le graphème "th" ainsi que 56% pour la combinaison des quatre.

L'intérêt de cette méthode vient de l'utilisation de simples caractères et graphèmes, et de caractéristiques généralement utilisées par les experts légaux de documents, malgré qu'elle n'atteigne pas les mêmes performances des experts humains. Son inconvénient majeur est l'extraction manuelle des caractères étudiés, ce qui la rend semi-automatique.

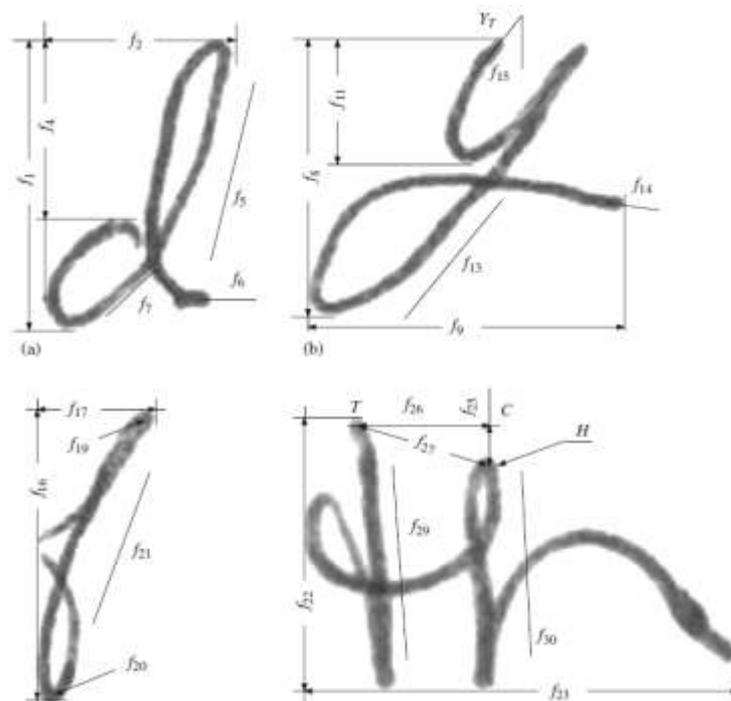


Figure 2.5. Certaines des caractéristiques extraites à partir des images des caractères "d", "y", "f" et du graphème "th" [PER 07].

2.2.4. Travaux de Tan & al (2009)

La méthode proposée par Tan et al. [TAN 09] est basée sur la création de prototypes de référence pour représenter les différents allographes d'une lettre (voir figure 2.6). Ces prototypes sont utilisés pour assigner chaque instance d'une lettre au prototype le plus similaire, la distribution de ces prototypes d'allographes est utilisée par les auteurs comme caractéristique afin de caractériser les styles d'écriture de différents scripteurs. Les auteurs justifient leur choix par le fait que deux scripteurs différents n'utilisent pas le même allographe pour la même lettre.

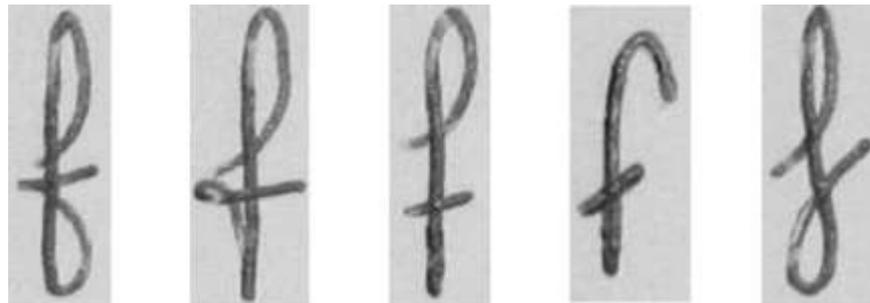


Figure 2.6. Différents allographes de la lettre « f » issus de différents scripteurs [TAN 09].

La méthode proposée s'articule autour de trois étapes (voir figure 2.7) : l'étape de construction des prototypes, l'étape de codage ou vectorisation des documents de référence et enfin, l'étape d'identification du scripteur du document de test en se basant sur les documents de référence.

Pour la construction des prototypes représentant les différents allographes de lettres, les auteurs ont segmenté 16585 mots issus la base de données IRONOFF [VIA 99] pour avoir 89760 caractères répartis sur les différentes lettres de l'alphabet. Ensuite, l'algorithme de regroupement k-means (k-moyennes) est utilisé pour déterminer les N prototypes d'allographes qui représentent les différents styles d'écritures de chaque lettre.

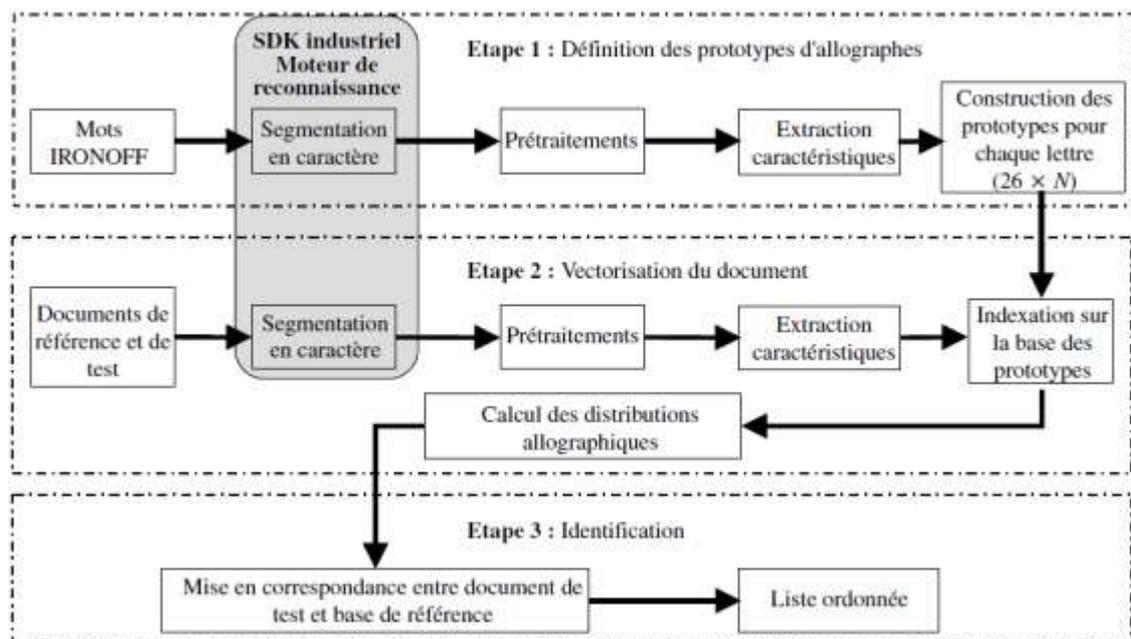


Figure 2.7. Schéma du système proposé par Tan et al [TAN 09].

Avant de procéder à la vectorisation des documents de référence et de test, ces derniers sont segmentés en caractères, et ces caractères sont reconnus par un moteur de reconnaissance

industriel afin qu'ils soient utilisés par l'étape de vectorisation. Cette dernière consiste à représenter le style d'écriture de chaque scripteur par une distribution statistique d'allographes par la mise en correspondance entre les allographes extraits à partir des documents de test ou de référence et les N prototypes construits précédemment lors de l'étape de construction prototypes d'allographes. Les auteurs ont adopté une approche fuzzy c-means (c-moyennes floues) pour l'estimation des distributions statistiques d'allographes. Pour la classification, les auteurs ont adopté un classifieur à distance minimale pour ordonner et classer les scripteurs en fonction de leur similarité avec le document de test.

Le meilleur taux d'identification réalisé par la méthode proposée est de l'ordre de 99.2%, ce résultat a été enregistré sur une base de 120 scripteurs. D'après les auteurs, la seule limitation de cette méthode est qu'une quantité minimale de texte doit être présente dans le document afin d'atteindre des taux d'identification suffisants. Cette quantité minimale de texte est d'environ 160 caractères (l'équivalent de 3 lignes de texte).

Un des inconvénients majeurs de cette méthode réside dans l'utilisation d'une étape de segmentation de mots en caractères, de reconnaissance et d'étiquetage de ces caractères. Le succès ou l'échec de la méthode proposée dépend fortement de cette étape. L'échec de la segmentation de mots en caractères et de la reconnaissance des caractères issus du processus de segmentation, entraîne une détérioration considérable des performances de la méthode d'identification de scripteurs.

2.2.5. Travaux de Woodard & al (2010)

Une autre méthode locale est proposée par Woodard et al [WOO 10]. Cette méthode statistique n'est pas fondée sur l'exploitation des relations spatiales entre les pixels de l'image, ni sur la segmentation explicite des unités linguistiques, et ne nécessite pas un apprentissage supervisé ou une étape de prétraitements.

Les auteurs proposent l'utilisation d'une représentation par Sac De Mots (SDM) ou Bag-of-Words pour la modélisation de scripteurs. Le calcul d'un descripteur SDM commence par la détection de régions d'intérêt (voir Figure 2.8) en utilisant l'un des détecteurs suivants: DoG, MSER, Hessian-Affine, Harris-Laplace, Harris-Affine, Harris-Hessian-Laplace ou Edge-Laplace [TUY 08], et, ensuite, des descripteurs locaux SIFT [LOW 04] sont calculés sur les zones localisées ou identifiées.

Une fois les descripteurs locaux SIFT obtenus, ils seront ensuite quantifiés en utilisant une méthode de classification non supervisée, la quantification vectorielle (VQ) qui est similaire à l'algorithme k-means. Ce quantificateur vectoriel est communément dénommé vocabulaire visuel, et un élément de son dictionnaire est appelé un mot visuel. La quantification consiste à assigner chaque descripteur SIFT à son plus proche voisin euclidien, ce qui revient à l'assimiler au mot visuel le plus proche. La représentation vectorielle de l'image de document manuscrit est calculée comme l'histogramme des fréquences d'apparition des mots visuels. Des exemples de mots visuels différents sont montrés dans la figure 2.9.

Enfin, la classification est effectuée en se basant sur ces mots visuels en utilisant une technique générative qui provient du domaine de la recherche d'informations, cette technique est appelée Analyse Sémantique Latente Probabiliste (pour Probabilistic latent semantic analysis : PLSA). Pour l'évaluation des performances de la méthode proposée, les auteurs ont utilisée une base comportant 153 images de textes issus de 51 scripteurs différents. Le meilleur taux

d'identification réalisé est de l'ordre de 98% en Top-1 et 99.3% en Top-10. Une erreur de vérification de l'ordre de 1.6% est enregistrée.

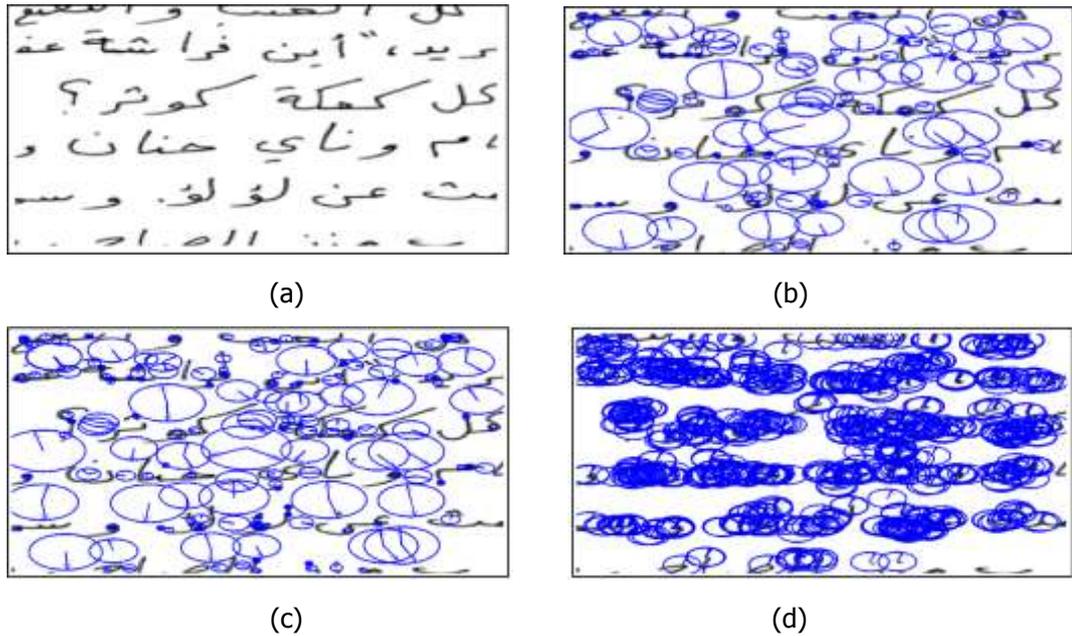


Figure 2.8. Exemples de régions d'intérêt détectées par différents types de détecteurs (a) image originale, (b) détecteur DOG, détecteur MSER, et, (d) détecteur Harris-Hesse-Laplace.

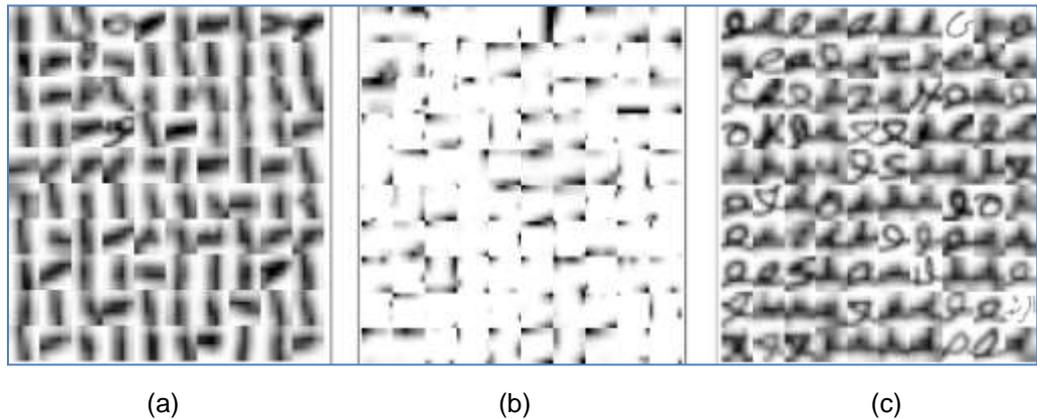


Figure 2.9. Exemple d'un regroupement en utilisant la quantification vectorielle (QV), (a) Premier mot visuel, (b) deuxième mot visuel, et (c) un troisième mot visuel.

2.2.6. Travaux de Jain & al (2011)

Dans les travaux de Jain et al. [JAI 11], les auteurs 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 2.10). Puis les segments ayant leurs extrémités proches (segments adjacents) sont reliés ensemble. Le descripteur k-AS est alors calculé à partir de chaque sous-ensemble connexe de 2, 3 et 4 points successivement.

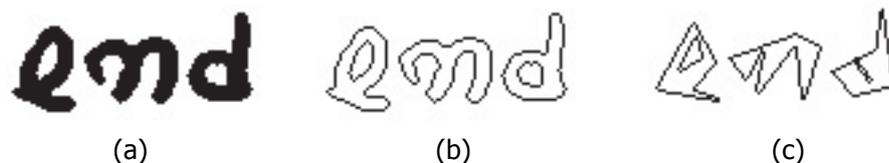


Figure 2.10. Illustration de l'extraction de segments à partir de l'image du mot "end", (a) : Image originale, (b) : Contour de l'image et (c) Segments extraits à partir des contours.

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 2.11 illustre un codebook généré à partir d'un ensemble de 3-AS.

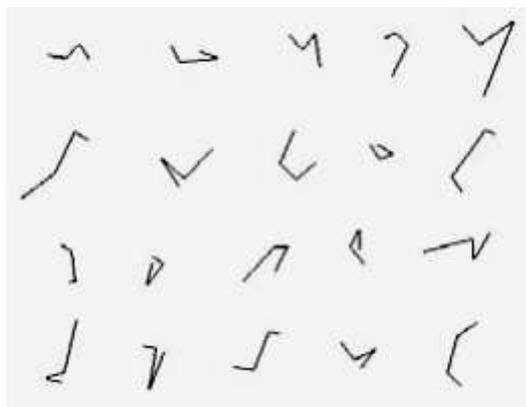


Figure 2.11. Codebook de 3-AS.

Les auteurs ont évalué leur méthode sur deux bases de données différentes : la base IAM (voir section 1.6.2) 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).

2.3. Analyse globale pour la classification de scripteurs

Dans cette section nous ne considérons que les méthodes basées sur une analyse globale de l'écriture manuscrite, cette analyse peut inclure : l'analyse de la texture, l'analyse fractale, l'analyse structurelle. Nous avons essayé de détailler les méthodes que nous avons jugées les plus intéressantes, les autres sont brièvement discutées.

2.3.1. Travaux de Zois & al (2000)

Zois et al. [ZOI 00] présentent un système d'identification de scripteurs à partir d'un seul mot, les auteurs ont introduit des caractéristiques basées sur une transformée morphologique de la fonction de projection de l'image d'un mot. L'image d'un mot est binarisée et squelettisée et ensuite projetée sur la direction horizontale pour avoir sa fonction de projection.

Afin de prendre en compte les espaces vides entre deux lettres, deux versions de la fonction de projection sont créées: une en éliminant les espaces vides (voir figure 2.12.d) et la seconde en les gardant (voir figure 2.12.c). Les fonctions sont ensuite ré-échantillonnées afin d'être indépendantes de la longueur du mot. Les projections obtenues sont morphologiquement traitées (ouverture) afin d'obtenir le vecteur de caractéristiques désiré. Les différences entre les ouvertures successives indiquent la quantité d'informations qui est enlevée par l'augmentation de la taille de l'élément structurant.

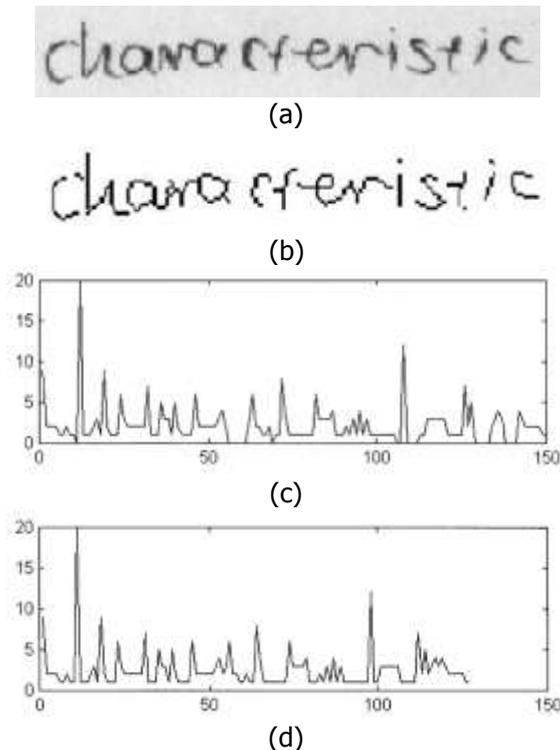


Figure 2.12. Prétraitements et projections dans [ZOI 00], (a) Échantillon scanné d'un mot, (b) Mot binarisé et prétraité, (c) Projection avec les espaces, (d) Projection sans les espaces.

La figure 2.13 montre les résultats de l'application d'ouvertures successives à la fonction de projection avec des éléments structurants de tailles 3 et 7 respectivement. Le vecteur de caractéristiques final est créé en partitionnant les projections en nombres de segments et en mesurant la surface éliminée, par l'opérateur de fermeture, dans chaque bloc par les deux éléments structurants de tailles 3 et 7.

Les caractéristiques proposées ont été évaluées sur une base qui comporte 50 scripteurs, différents, chacun d'entre eux a écrit le mot «*characteristic*» 45 fois, le même procédé a été répété pour l'équivalent grec du mot. Les deux modes de reconnaissance de scripteurs ont été considérés lors des expérimentations effectués par les auteurs. En mode identification, un taux d'identification de l'ordre de 92.5% a été obtenu en moyenne sur les mots grecs et anglais, alors qu'en mode vérification, le taux d'erreur enregistré est de l'ordre de 5%.

Il est assez important de noter que les auteurs se sont intéressés à un aspect particulier et très difficile dans le domaine de la reconnaissance de scripteurs : il s'agit de la reconnaissance d'un scripteur à partir d'un mot qui nécessite de disposer d'un nombre important d'échantillons en base de référence (45 mots pour chaque scripteur).

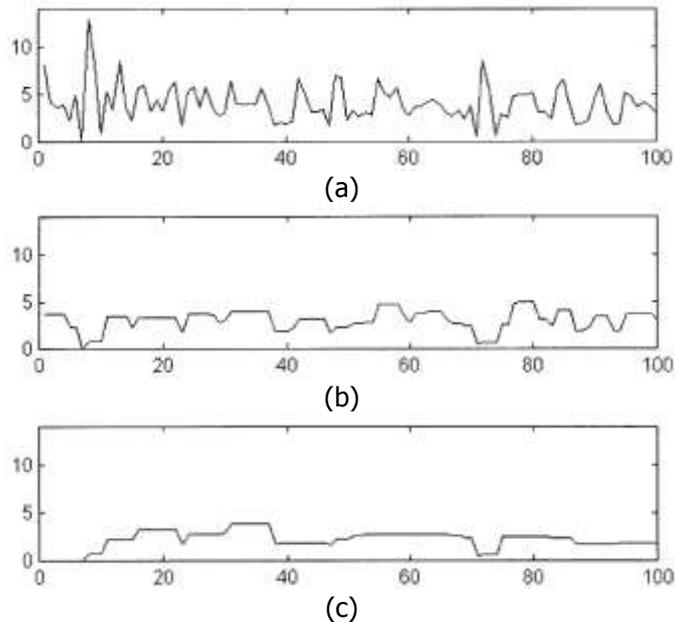


Figure 2.13. Illustration de l'application de l'ouverture morphologique sur la fonction de projection (a) Fonction de projection initiale, (b) : ouverture avec un élément structurant de taille 3, (c) ouverture avec un élément structurant de taille 7.

2.3.2. Travaux de Said & al (2000)

Said et al. [SAI 00] ont présenté une méthode globale qui est devenue un classique dans le domaine de la reconnaissance de scripteurs. L'écriture de chaque individu a été considérée comme une texture potentiellement différente de sorte que des algorithmes de reconnaissance de texture puissent être appliqués. Plus précisément, les auteurs extraient deux types de caractéristiques texturales, les premières sont obtenues à l'aide des filtres de Gabor et les secondes à partir des matrices de cooccurrences en niveaux de gris.

- Un filtre de Gabor s'exprime comme suit:

$$h(x, y) = g(x, y)e^{-2\pi j(ux+vy)}$$

Chaque voie coaxiale s'écrit à l'aide d'une paire de filtres de Gabor $h_e(x, y; f, \theta)$ et $h_o(x, y; f, \theta)$. Les deux filtres sont posés et sont donnés par :

$$\begin{aligned} h_e(x, y; f, \theta) &= g(x, y) \cos(2\pi f(x \cos \theta + y \sin \theta)) \\ h_o(x, y; f, \theta) &= g(x, y) \sin(2\pi f(x \cos \theta + y \sin \theta)) \end{aligned}$$

Le filtrage est réalisé par simple produit dans le domaine de Fourier, puis l'image résultat est obtenue par transformée de Fourier Inverse. Prenant la transformée de Fourier, les résultats des deux filtres sont combinés par :

$$q(x, y) = \sqrt{q_o^2 f(x, y) + q_e^2 f(x, y)}$$

Les deux principaux paramètres dans un filtre de Gabor sont la fréquence f et l'orientation θ . Les auteurs ont utilisé les fréquences de 4, 8, 16 et 32, et, pour chaque fréquence, ils ont utilisé les orientations $\theta = 0, 45, 90$ et 135 . Ceci donne au final 16 images en sortie. Le vecteur de caractéristiques est constitué de la moyenne et de l'écart type du niveau de gris de chaque image, soit au total 32 caractéristiques.

- Les matrices de cooccurrences en niveaux de gris sont coûteuses en calcul, en effet, pour une image en N niveaux de gris, la matrice aura une taille de $N \times N$. Pour une image binaire

N vaudra uniquement 2, ce qui rend le calcul des matrices de cooccurrences plus raisonnable. Les auteurs calculent ces matrices pour cinq distances ($d = 1, 2, 3, 4, 5$) et quatre directions ($\alpha = 0, 45, 90, 135$). Ce qui fournit, pour chaque image, 20 matrices de dimension 2×2 . Sachant que ces matrices sont symétriques, il n'existe que trois valeurs indépendantes qui sont utilisées comme caractéristiques. Finalement les auteurs construisent un vecteur de 60 caractéristiques (20 matrices \times 3 valeurs) par image.

Pour l'évaluation de leur méthode, les auteurs ont utilisé une base de données composée de 1000 textes collectés auprès de 40 individus différents, ayant, chacun, recopié 25 fois une lettre dans une page au format A4, scannée avec une résolution de 150 points par pouce. A partir de chaque page, 25 blocs de taille 128×128 pixels ont été extraits. Deux groupes sont formés pour chaque scripteur : groupe A (15 blocs d'apprentissage et 10 blocs de test) et groupe B (10 blocs d'apprentissage et 15 blocs de test).

Pour la classification, les auteurs ont envisagé l'utilisation d'un classifieur de type k plus proches voisins (k -PPV) ainsi qu'un autre utilisant la distance euclidienne pondérée [ALP 04, BEL 92, BIS 06, COR 02, MIT 97]. En utilisant les filtres de Gabor et la distance euclidienne pour la classification, les taux d'identification enregistrés sont de l'ordre de 96%. Alors qu'avec les matrices de cooccurrences, le taux d'identification obtenus sont de l'ordre de 72,2% avec le même classifieur. Avec les k -PPV, le taux de bonne identification en utilisant les filtres de Gabor est de l'ordre de 86%, tandis qu'avec les matrices de co-occurrences ce taux atteint 75%.

L'approche présentée ci-dessus est basée sur des caractéristiques extraites par des filtres de Gabor et des matrices de cooccurrences. Ces caractéristiques globales, couramment utilisées en traitement d'images, nécessitent de disposer de plusieurs échantillons d'écritures. Pour pallier au manque de données, les auteurs ont complété les lignes d'écriture par du texte extrait dans la même page, ce qui, bien évidemment, biaise considérablement les caractéristiques finales.

2.3.3. Travaux de Marti et al (2001)

Marti et al. [MAR 01], extraient un ensemble de 12 caractéristiques à partir de lignes de textes manuscrits, qui sont ensuite classées en utilisant un classifieur de type k plus proche voisin (k -PPV) ainsi qu'un classifieur neuronal (*feed forward neural network*), pour l'identification de scripteurs.

En utilisant une technique de projection, l'image du texte manuscrit est d'abord segmentée en lignes qui sont ensuite binarisées en employant la méthode Otsu [OTS 79].

Un ensemble de 12 caractéristiques est extrait à partir de chaque ligne de texte, ces caractéristiques correspondent principalement à la hauteur, la largeur, l'inclinaison et la lisibilité des caractères et à l'espace inter-mots. La hauteur de l'écriture est définie à l'aide des lignes de références (voir figure 2.14) déterminées par la projection horizontale des lignes de textes. En outre, des caractéristiques basées sur le comportement fractal de l'écriture, qui sont en corrélation avec la lisibilité de l'écriture, sont également utilisées.

Les auteurs ont évalué leur système a été évalué sur 100 pages issus de 20 scripteurs différents de la base de données IAM (voir section 1.6.2) avec un total de 912 lignes de texte divisées en cinq groupes de tailles égales (182 lignes par groupe). Quatre groupes ont servi comme base d'apprentissage tandis que le groupe restant a été utilisé comme base de test. Un taux d'identification moyen de l'ordre de 87,8% a été atteint avec l'utilisation de 7 des 12 caractéristiques proposées en employant la technique de classification des k -plus proches

voisins et un taux de l'ordre de 90,7% a été enregistré en utilisant les réseaux de neurones artificiels avec toutes les caractéristiques proposées.

Les taux d'identification de scripteurs obtenus avec ces caractéristiques sont très prometteurs, mais ils sont basés sur un protocole d'évaluation qui exige l'utilisation de 4 échantillons par scripteur pour l'entraînement du système. Cependant, dans des problèmes pratiques, on risque de ne pas avoir autant de texte disponible pour chacun des scripteurs, ce qui rend discutable l'efficacité de ces caractéristiques. Il est aussi important de noter que la taille de la base de données (nombre de scripteurs) utilisée ne peut pas vraiment refléter le pouvoir discriminant des caractéristiques proposées.

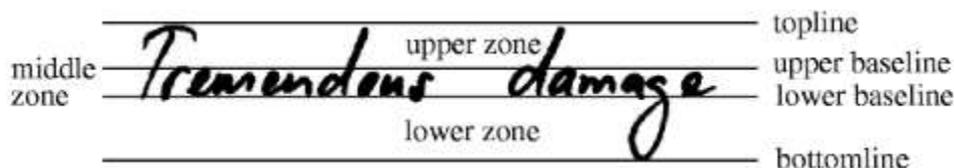


Figure 2.14. Exemple de lignes de référence [MAR 01].

2.3.4. Travaux de Srihari & al (2002)

Srihari et al. [SRI 02] ont présenté une étude dont l'objectif principal est de prouver l'individualité de l'écriture manuscrite. Des échantillons d'écriture issus de 1500 individus représentatifs de la population des Etats-Unis en fonction du sexe, de l'âge et de l'ethnie ont été collectés. Chaque individu a recopié trois exemplaires d'une lettre composée de 156 mots, scannée avec une résolution de 300 ppp. Un ensemble de macro et micro-caractéristiques a été ensuite extrait de chacun des échantillons d'écriture. Les macro-caractéristiques peuvent être extraites à partir d'un document, d'un paragraphe, d'un mot ou à partir d'un caractère tandis que les micro-caractéristiques ne peuvent être calculées qu'au niveau caractère.

Au total, onze macro-caractéristiques ont été employées, ces caractéristiques sont : l'entropie du niveau de gris (indique le niveau de variation dans l'image), seuil de niveau de gris (un indicateur de la pression du stylo), le nombre de pixels noirs (indicateur de la pression du stylo, de la taille de l'écriture et de l'épaisseur des traits), le nombre de contours intérieurs et extérieurs (détermine le mouvement de l'écriture), les inclinaisons des contours horizontales, verticales, négatives ou positives (définissent la nature des formations des tracés), l'inclinaison moyenne des lignes verticales et pseudo-verticales et enfin la hauteur moyenne.

Les micro-caractéristiques (voir figure 2.15) proposées consistent en 512 caractéristiques binaires, dont 192 structurelles, 128 de concavités et 192 du gradient. Ce vecteur de caractéristiques binaires de 1024 composantes est le seul qui intervient au niveau caractère.

Deux séries de tests ont été conduites par les auteurs pour les deux modes de reconnaissance de scripteurs (identification et vérification):

En mode d'identification, une décision du type plus proche voisin a été envisagée en recherchant dans la base de référence le vecteur de caractéristiques le plus proche du vecteur de caractéristiques extrait du document inconnu. Les tests ont été conduits en faisant varier le nombre de scripteurs dans la base de référence et en utilisant des échantillons d'écriture à identifier de taille variable (document, paragraphe, mot : un caractère ou tous les caractères du mot). Ainsi, sur une base de 100 scripteurs, les auteurs parviennent à identifier le bon scripteur à partir d'un document entier (la lettre CEDAR) dans 82% des cas, et dans 49% des cas à partir d'un paragraphe pour chuter à 28% en utilisant un mot. Ces taux baissent à 59% (document),

25% (paragraphe) et 9% (mot) sur une base de 900 scribes. Bien entendu, plus la taille de l'échantillon à identifier est faible, plus le taux d'identification diminue, de même qu'il diminue lorsque le nombre de scribes augmente.

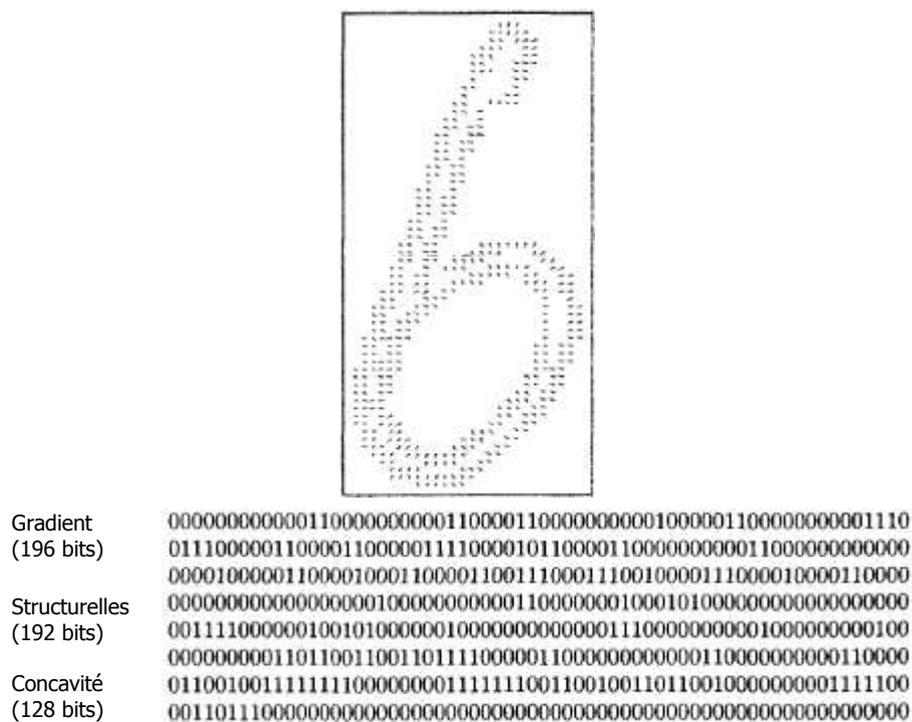


Figure 2.15. Micro-caractéristiques extraites à partir de l'image du chiffre «6».

En mode de vérification, les mêmes caractéristiques ont été conservées pour l'estimation des distances intra et inter scribes. Les distributions de ces distances ont fait apparaître deux erreurs de classification : l'erreur de premier type (probabilité d'attribuer, par erreur, les deux écritures à un même scribe) et l'erreur de second type (probabilité d'attribuer, par erreur, les deux écritures à deux scribes différents). La base a été découpée en 4 groupes, un groupe pour l'apprentissage, un groupe pour la validation et les deux groupes restants pour le test. Pour la décision, un réseau de neurones à deux classes a été envisagé. Le réseau de neurones utilisé possède trois couches (8 unités en entrée, 5 unités cachées, une unité pour la sortie : 0 ou 1). Une moyenne de 95% de bonne vérification a été obtenue en vérifiant deux documents entiers et 83% en vérifiant deux paragraphes.

L'approche proposée par les auteurs est intéressante du point de vue de la taille de la base de données ainsi que du point de vue de la variété des expérimentations réalisées. Cependant, la segmentation manuelle des mots en caractères est un réel inconvénient dans un système qui se proclame automatique. Quant aux caractéristiques utilisées, vu leur nombre et leur diversité, on pourrait s'attendre à avoir de meilleurs résultats quand la base de référence est réduite à 100 scribes, en plus du fait que le système dispose de 3 échantillons (pages) d'écriture au même contenu textuel pour chaque scribe.

2.3.5. Travaux de Gazzah & al (2005, 2006)

Gazzah & al [GAZ 05, GAZ 06] ont proposé une méthode d'identification de scripteurs qui a été évaluée sur une base de données créée au niveau de leur laboratoire de recherche. Pour construire leur base de données, les auteurs ont rédigé une lettre qui comporte un texte arabe composé de 505 caractères, 15 chiffres et 6 signes de ponctuation. Le choix du contenu de la lettre a été effectué de manière à assurer la représentativité des différentes formes de chacun des caractères arabes. 60 scripteurs ont participé à la collecte des données en utilisant un stylo noir, une feuille blanche avec des lignes tracées (support d'écriture), chacun des 60 scripteurs a recopié la lettre 3 fois, ce qui a donné un total de 180 pages en format A4 qui ont été numérisés en niveaux de gris à une résolution de 300 ppp. 120 pages de textes ont servi pour l'apprentissage et les 60 pages restantes ont été utilisées pour les tests.

L'objectif de l'étude présentée par Gazzah & al [GAZ 05, GAZ 06] était de déterminer un ensemble d'attributs caractérisant le style d'écriture de chaque scripteur et permettant de le différencier des autres scripteurs de la base de données. Les auteurs extraient deux types de caractéristiques (structurelles et texturales) à partir de lignes de texte :

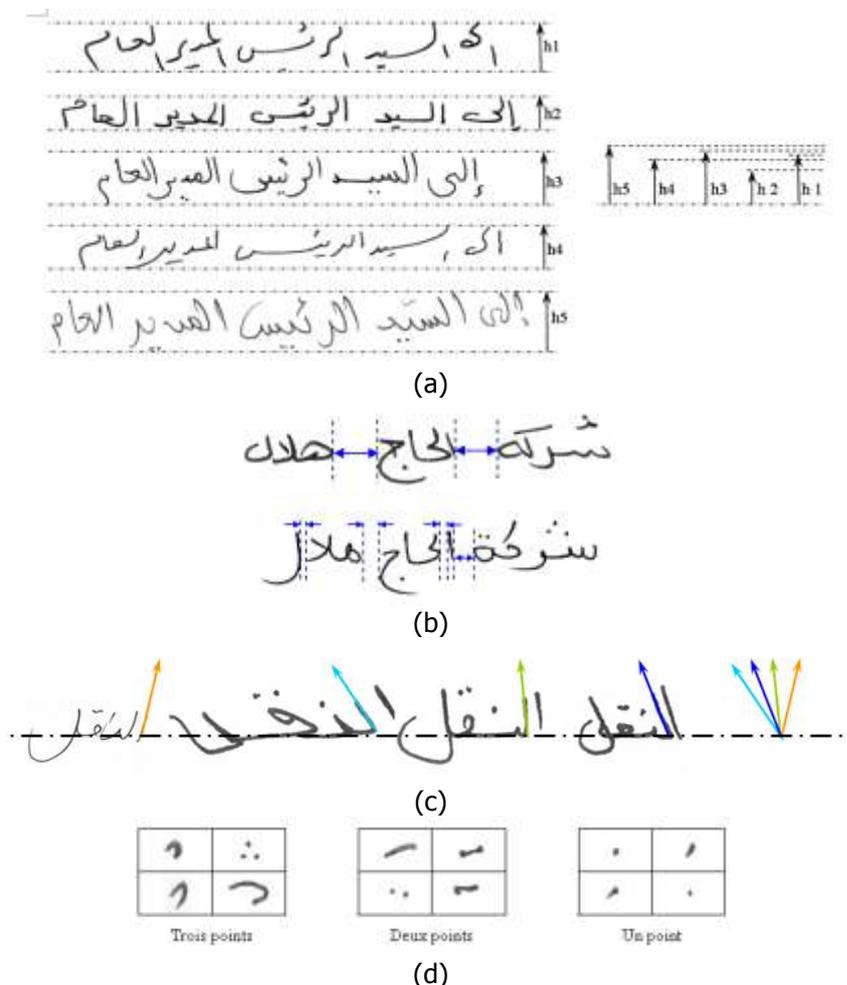


Figure 2.16. Variations structurelles de l'écriture, (a) : hauteur d'une ligne de texte, (b) : espaces entre pseudo mots, (c) : inclinaisons de l'écriture, (d) : différentes formes de points diacritiques.

Pour déterminer les caractéristiques structurelles, les auteurs ont effectué une étude morphologique sur la base de test qui a montré que la hauteur de la ligne de texte, l'espace moyen entre pseudo-mots, l'inclinaison des hampes ainsi que les caractéristiques dimensionnelles des points diacritiques, représentent des primitives discriminantes, qui marquent bien le style d'un scripteur, elles ont donc été retenues pour la caractérisation des scripteurs (voir figure 2.16). En ce qui concerne les caractéristiques texturales, les auteurs ont opté pour la description de la texture de l'écriture par la mesure de l'entropie de chacune des lignes de texte et par la considération de caractéristiques issues de l'application de la transformée en ondelettes. Les caractéristiques choisies sont l'écart type et la moyenne des matrices associées aux images de l'approximation, aux détails horizontal, vertical et diagonal.

Au total, 20 caractéristiques (12 structurelles et 8 texturales) sont considérées. Ces caractéristiques, nécessitent de disposer de plusieurs échantillons d'écriture afin de pouvoir dériver des caractéristiques statistiques suffisamment stables. Les auteurs ont utilisé, par la suite, les algorithmes génétiques pour opérer la sélection des caractéristiques non redondantes les plus discriminantes. L'application de l'algorithme d'optimisation a permis de retenir uniquement 12 caractéristiques parmi les 20 considérées. Les caractéristiques retenues sont : la hauteur de la ligne, la moyenne des inclinaisons des hampes, l'écart type des inclinaisons des hampes, la moyenne de l'épaisseur des hampes, la moyenne du ratio hauteur/largeur du rectangle englobant un point diacritique, la densité de pixels dans un rectangle englobant un point diacritique, la moyenne de la matrice de l'image d'approximation issue des ondelettes, l'écart type de la matrice de l'image d'approximation issue des ondelettes, la moyenne de la matrice du détail horizontal, l'écart type de la matrice du détail vertical, la moyenne de la matrice du détail diagonal et l'entropie

Pour identifier un scripteur, étant donné l'ensemble de caractéristiques préalablement extraites, deux classifieurs ont été implémentés et testés : Les séparateurs à vaste marge (SVM) ainsi que les réseaux de neurones artificiels de type perceptron multi-couches [ALP 04, BEL 92, BIS 06, COR 02, MIT 97]. Les meilleurs résultats obtenus sont de l'ordre de 94.73% pour les réseaux neuronaux de type perceptron multi-couches [GAZ 06] et 93.76% pour les séparateurs à vaste marge (SVM) [GAZ 05].

2.3.6. Travaux de Feddaoui & al (2006)

Dans les travaux de Feddaoui & al [FED 06], les auteurs proposent une approche de vérification du scripteur. Leur approche se base sur trois étapes principales: prétraitements, extraction de caractéristiques et décision.

Pour l'étape de prétraitements, les auteurs procèdent à la binarisation de l'image du document manuscrit avant l'estimation et la correction de l'angle de pente des lignes consécutives du texte afin d'avoir des lignes horizontales. Puis, comme tout document manuscrit peut contenir différents espaces inter-lignes et inter-mots, différentes marges et longueurs de lignes, et, dans le but de minimiser l'influence de ces facteurs sur le processus de caractérisation de la texture (extraction de caractéristiques), les auteurs commencent par la localisation des lignes de texte en utilisant une technique de projection horizontale, puis normalisent les espaces inter mots et inter-lignes par l'élimination des espaces inutiles et, enfin, ils examinent le texte manuscrit ligne par la ligne. Les espaces vides localisés sont remplis par des morceaux de texte extraits à partir du même document manuscrit (voir figure 2.17). Il est à noter que ces techniques de prétraitements sont les mêmes que celles proposées dans [SAI 00].



Figure 2.17. Prétraitements dans [FED 06], (a) Image binarisée de document; (b) Correction des angles de pente, (c) script horizontal, (d) projection horizontale, (e) normalisation des espaces inter-lignes, (f) projection verticale, (g) normalisation des espaces inter-mots et (h) remplissage de texte.

Une fois la phase de prétraitements accomplie, les auteurs extraient, à partir de chaque page, 16 blocs de textes de taille 128×128 pixels servant pour l'apprentissage et le test. Puis ils extraient des caractéristiques issues de la texture de chaque bloc de texte : ces caractéristiques sont obtenues en utilisant les filtres de Gabor décrits dans la section 2.3.2. Les fréquences utilisées sont : 4, 8, 16 et 32 et, pour chaque fréquence, quatre orientations sont considérées : 0° , 45° , 90° et 135° , pour chaque blocs un vecteur de 32 caractéristiques est obtenu. La figure 2.18 montre quelques résultats obtenus par l'application de 16 filtres de Gabor sur un bloc de texte normalisé.

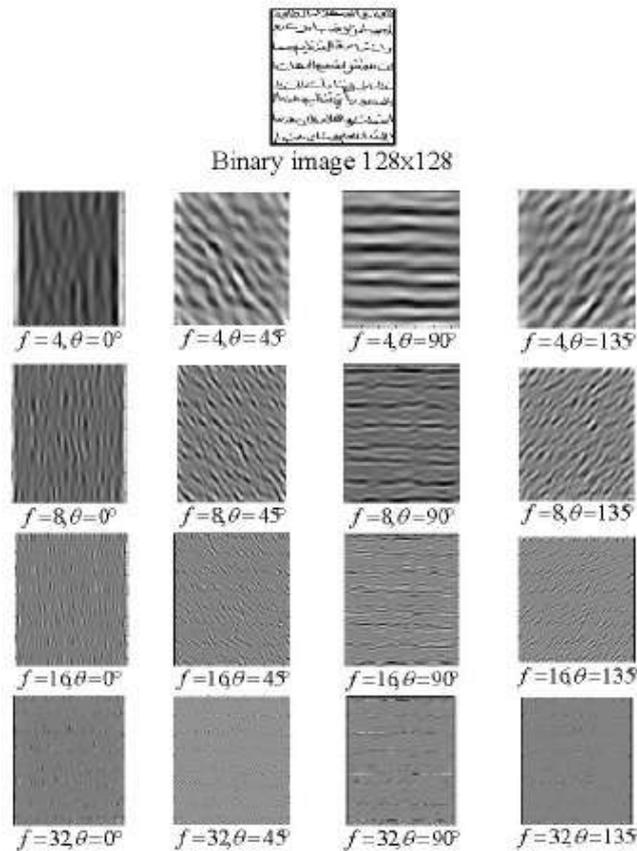


Figure 2.18. Résultats de l'application des 16 filtres de Gabor considérées avec différentes fréquences et orientations [FED 06].

Après l'étape d'extraction de caractéristique vient la dernière étape de l'approche, c'est celle de la comparaison et décision. Au cours de cette étape, chaque script est représenté par un vecteur de caractéristiques. La vérification nécessite la comparaison entre deux vecteurs afin d'estimer leur similarité en utilisant la distance euclidienne pondérée.

Cette approche a été testée sur deux bases de données différentes, la première contient 320 images de documents imprimés employant les 20 fontes les plus utilisées de l'arabe tel que Arabic Transparent, Traditional Arabic, Simple Outline, Pat, Simplified Arabic, Diwani, ... etc. La deuxième base contient des documents manuscrits écrits par 22 scripteurs différents.

Pour la première base, le taux de fausses acceptations varie entre 1.5% et 2.5% et celui de faux rejets varie entre 0.2% et 0.3% alors que pour la deuxième base, les taux de fausses acceptations enregistrés varient entre 1.9% et 5.2% et ceux de faux rejets varient entre 0.3% et 0.8%.

L'approche proposée dans [FED 06] exige des blocs de textes uniformes et l'analyse des textes ne peut se faire directement sur les documents puisqu'ils sont affectés d'espaces interlignes et inter-mots irréguliers. De ce fait, une réduction de ces facteurs par une normalisation est nécessaire. Cette approche est basée sur des caractéristiques globales (texturales), son inconvénient majeur est qu'elle nécessite une taille d'échantillons assez grande pour qu'elle puisse bien caractériser une écriture. Il est à noter que les filtres de Gabor ont été également appliqués à la reconnaissance de scripteurs à partir de documents chinois [ZHU 00, HEL 08]

2.3.7. Travaux de Schlapbach & al (2007)

Schlapbach et al. [SCL 04a, SCL 04b] ont utilisé des caractéristiques qui étaient initialement dédiées à la reconnaissance des écritures en milieu contraint et vocabulaire restreint. Dans l'un de leurs travaux de recherche, Schlapbach & al [SCL 04a, SCL 04b] ont utilisé des modèles de Markov cachés (HMMs) pour l'identification et la vérification du scripteur à partir d'une ligne de texte, en utilisant un certain nombre de modèles de Markov cachés (HMMs). Ils déterminent l'identité du scripteur en choisissant le HMM du scripteur qui fournit la meilleure mesure de confiance de la ligne de texte en question. Comme chaque HMM a été entraîné et personnalisé avec des données individuelles et spécifiques d'un seul scripteur, le scripteur du HMM qui fournit la mesure de confiance la plus élevée sera considéré comme le scripteur de la ligne de texte en question. Pour l'extraction de caractéristiques, les auteurs ont utilisé une fenêtre glissante qui est couramment pratiquée avec les HMMs. Une fenêtre d'un pixel de largeur est déplacée de gauche à droite sur la ligne de texte. A chaque position, neuf caractéristiques géométriques sont extraites : trois caractéristiques globales et six caractéristiques locales. Les caractéristiques globales représentent le nombre de pixels noirs dans la fenêtre, le centre de gravité et le moment du second ordre des pixels noirs.

Les six caractéristiques locales restantes sont la position et la direction du contour de la partie supérieure et inférieure, le nombre de transitions noir-blanc dans la fenêtre, le ratio des pixels entre contour haut et bas. Les six caractéristiques locales sont illustrées dans la figure 2.19.

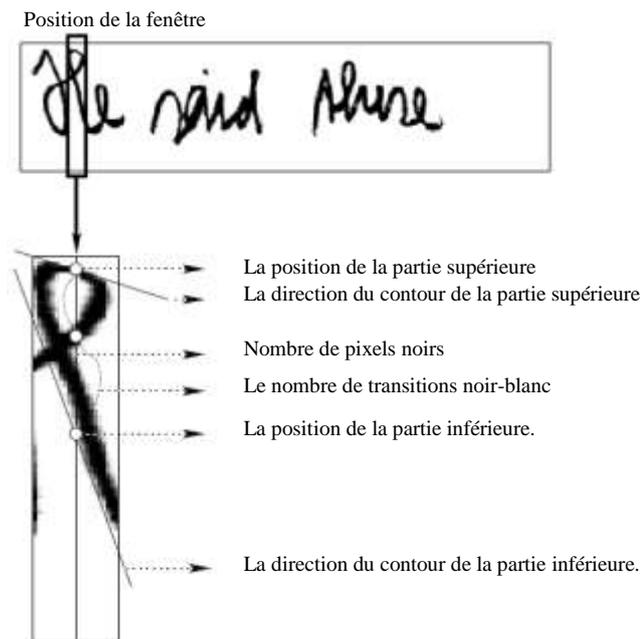


Figure 2.19. Caractéristiques locales extraites de chaque colonne de pixels dans la fenêtre glissante [SCL 07].

Pour la tâche d'identification de scripteurs, un taux d'identification correcte de l'ordre de 97,03% a été enregistré en utilisant 4103 lignes de texte provenant de 100 scripteurs différents issus de la base IAM. En ce qui concerne la vérification du scripteur, la méthode proposée donne de bons résultats pour les deux tâches considérées. Un taux d'égale erreur de 2.0% est atteint sur un total de 8100 lignes de texte provenant de 100 clients et 20 imposteurs non qualifiés. Par l'utilisation de faux qualifiés, le taux d'égale erreur augmente à 4.47% sur un

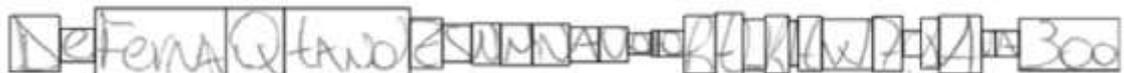
ensemble de test qui se compose de 169 lignes de texte de 20 clients et 169 lignes de texte de 20 imposteurs.

2.3.8. Travaux de Bertolini & al (2013)

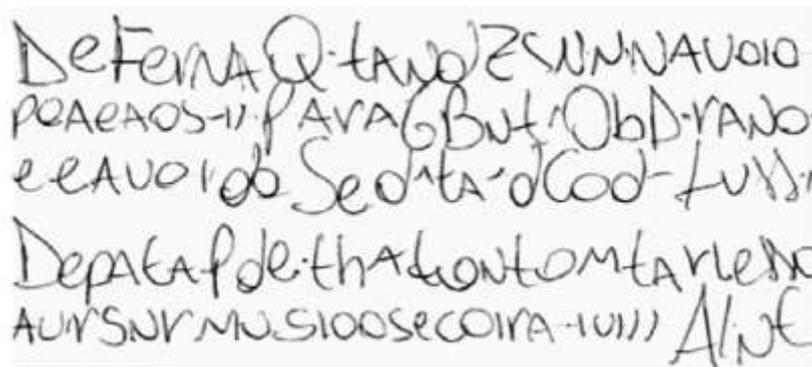
Dans [BER 13], les auteurs proposent un système de reconnaissance de scripteurs, en utilisant des descripteurs de textures. Ces descripteurs sont basés sur les motifs binaires locaux (LBP pour Local Binary Patterns) ainsi que la méthode de Quantification de phase locale (LPQ pour Local Phase Quantization).

Le concept du LBP a été proposé comme un descripteur simple, robuste face aux variations des niveaux de gris, il a prouvé aussi sa capacité à discriminer efficacement entre une large gamme de textures. Plus tard, il a été étendu afin qu'il soit uniforme et invariant à la rotation. Le concept du LPQ s'est avéré robuste aux variations de flou et ses performances sont généralement meilleures que celles des LBP.

Pour extraire ces descripteurs, l'image est binarisée en utilisant l'algorithme Otsu, puis balayée de haut en bas et de gauche à droite afin de détecter toutes ses composantes connexes en considérant un voisinage 8-connexté. Ensuite, les petites composantes, telles que les points, les virgules ainsi que le bruit, sont éliminées. Les boîtes englobant les composantes connexes restantes sont ensuite utilisées pour extraire les composantes originales de l'image en niveaux de gris. Les composantes en niveaux de gris sont ensuite alignées sur la nouvelle image en utilisant le centre de masse de la boîte englobante. Ce processus est illustré à la figure 2.20. Après le remplissage de la première ligne, la hauteur moyenne de l'ensemble des composantes connexes utilisées est calculée. Cette valeur est utilisée pour définir l'emplacement de la ligne suivante. La même opération est répétée jusqu'à l'obtention de l'image de texture (voir figure 2.21.a). L'image finale de texture représentant l'écriture du scripteur est finalement segmentée en blocs de taille $n \times m$ (voir figure 2.21.b).



(a)



(b)

Figure 2.20. Le processus de génération de texture (a) le remplissage d'une ligne et (b) la texture résultante.

Les expérimentations effectuées par les auteurs se divisent en deux parties. Celles de la première partie traitent le problème de la vérification du scripteur, et celles de la deuxième partie se concentrent sur le problème de l'identification de scripteurs. Dans toutes les

expérimentations, les séparateurs à vaste marge (SVM) ont été utilisés pour la classification. Deux bases de données ont été utilisées pour l'évaluation des caractéristiques proposées, la base d'images BFL (voir section 1.6.4) et la base d'images IAM (voir section 1.6.2), ces bases contiennent des images de textes issus de 315 et 650 scripteurs, respectivement.

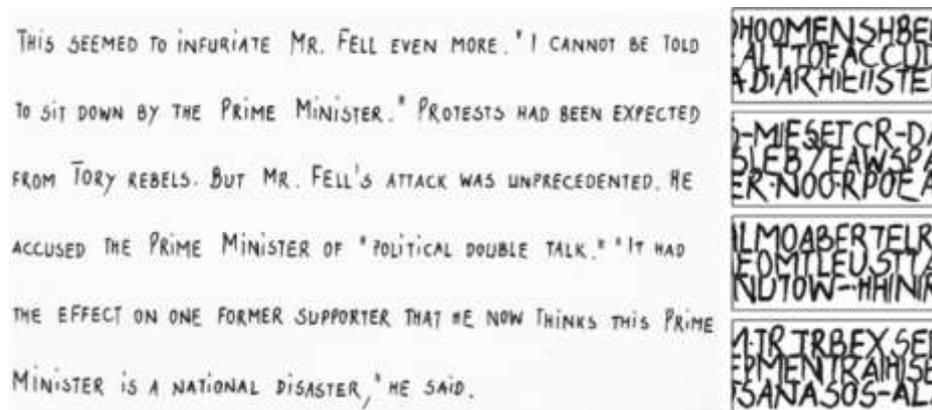


Figure 2.21. Un échantillon de la base IAM et ses blocs de texture.

Les résultats réalisés montrent que la méthode proposée pour la reconnaissance de scripteurs est capable d'atteindre une précision de l'ordre de 94.7% sur la base BFL et 94.5% sur la base IAM en utilisant les descripteurs LBP, et 99.2% en utilisant la base BFL et 96.7% sur la base IAM en considérant les descripteurs LPQ. Ces résultats se comparent favorablement à ceux de l'état de l'art et montrent que les descripteurs LBP et LPQ sont des alternatives intéressantes pour modéliser les écritures de différents scripteurs.

2.3.9. Travaux de Chaabouni & al (2014)

Chaabouni et al. [CHB 14] ont présenté une méthode pour l'identification de scripteurs. Cette méthode est basée sur l'extraction de caractéristiques multi-fractales à partir des images de mots, ainsi que leurs signaux en-ligne. L'objectif fondamental de leur étude est d'explorer l'utilité potentielle des caractéristiques multi-fractales comme une nouvelle méthode pour différencier les personnes au moyen de leurs écritures en-ligne et hors-ligne. La figure 2.22 montre le schéma du système proposé.

Pour l'extraction des caractéristiques fractales à partir des images de mots, Les auteurs ont utilisé les méthodes DLA (Diffusion Limited Aggregates), BC (Box-counting), ABCD (Average Box-counting Density). Les méthodes DLA et ABCD ont été adaptées aux signaux en-ligne. Une autre méthode a été utilisée pour l'extraction de caractéristiques à partir des signaux en-ligne des mots, c'est la méthode COMPASS.

Pour l'évaluation de la méthode proposée, les auteurs ont utilisé des images de mots provenant de la base de données ADAB [KHE 11] qui se compose de plus de 20000 mots manuscrits arabes écrits par plus de 170 scripteurs différents. Les mots de la base correspondent à 937 noms de villes et de villages tunisiens. Les auteurs ont sélectionné 110 scripteurs de la base ADAB, chacun d'eux a écrit 12 fois les noms de 24 villes ou villages tunisiens. 16 mots ont été utilisés pendant la phase d'entraînement d'un classifieur de type k-plus proches voisins (k-ppv), et les 8 mots restants ont été utilisés pour l'identification (le test).

Les taux d'identification moyens enregistrés en utilisant les images de mots sont de l'ordre de 70.3% en Top-1 et de 85.9% en Top-10, Pour les signaux de mots en-ligne, les résultats réalisés sont 73.1% en Top-1 et 86.5% en Top-10%, La combinaison des deux types de

signaux (en-ligne et hors-ligne) améliorent les performances de la méthode proposée en atteignant 83.8% et 91.3% d'identifications correctes en Top-1 et Top-10 successivement.



Figure 2.22. Architecture du système proposé par Chaabouni et al [CHB 14].

2.3.10. Travaux de Wang & al (2003, 2004)

Wang et al. [WAN 03, WAN 04] ont analysé la puissance des caractéristiques directionnelles extraites à partir des images de caractères chinois. Deux méthodes sont proposées, la première concerne l'identification de scripteurs [WAN 03] et la deuxième aborde le problème de vérification de scripteurs [WAN 04]. Les deux méthodes sont dépendantes du texte.

Avant d'extraire les caractéristiques directionnelle à partir de l'image d'un caractère les auteurs appliquent, à chaque caractère, une méthode de normalisation linéaire appelée "gravity-center normalization" afin de conserver les différents styles des différents scripteurs, Une fois les caractéristiques extraites, l'Analyse de Composante Principale (PCA) et l'Analyse Linéaire Discriminante (LDA) sont employées pour réduire la dimension du vecteur de caractéristiques et pour le rendre plus distinctif. Pour évaluer l'efficacité des méthodes proposées, les auteurs ont employé deux bases de données. La première se compose de 34 caractères chinois manuscrits, chacun a été écrit 16 fois par 25 personnes. La deuxième base se compose de 20 caractères chinois manuscrits, chacun a été écrit 16 fois par 27 personnes différentes. Des taux d'identification qui varient de 88,67% à 100% ont été atteints pour l'identification et des taux d'erreurs égales EER de l'ordre de 3.17% pour la première base de données et 4,68% pour la deuxième base de données ont été enregistrés.

2.3.11. Travaux de Garain & al (2009)

Garain et al. [GAR 09] ont proposé un système d'identification de scripteur utilisant les modèles autorégressifs bidimensionnels où chaque scripteur est représenté par un ensemble de coefficients autorégressifs bidimensionnels. Les détails de l'estimation des coefficients peuvent être trouvés dans [GAR 09]. Les auteurs ont évalué leur système sur deux bases de données contenant des images de textes en français, appartenant à 382 scripteurs, extraites de la base

de données RIMES [GRO 08] ainsi que des images de textes manuscrits bengali, provenant de 40 scripteurs différents, extraites de la base ISI. Les meilleurs résultats enregistrés sont de l'ordre de 75% sur la base ISI, de 57% sur la base RMIES et de 70.7% en combinant les deux bases de données.

2.3.12. Autres travaux

Le travail présenté par Al Zoubeidy et al. [ALZ 05] décrit, à notre connaissance, le premier système de d'identification de scripteurs pour l'écriture arabe. Les auteurs y ont adapté l'approche proposée par Said et al. [SAI 00] (voir section 2.3.2). Des caractéristiques sont extraites de l'image de l'écriture en utilisant un filtre de Gabor multi canal et un calcul de matrice de co-occurrence. Un taux d'identification de l'ordre de 92.8% a été atteint avec la distance euclidienne sur une base de 500 textes arabes (à raison de 25 pages par scripteur).

Gazzah et al [GAZ 07] proposent une approche globale en mode indépendant du texte en explorant l'écriture par l'analyse de la texture avec des ondelettes 2D utilisant le Lifting Scheme. Une évaluation comparative entre les caractéristiques extraites de la texture par neuf transformations en ondelettes différentes a été effectuée. Des expérimentations ont été menées en utilisant un perceptron multicouches sur une base de 180 échantillons de textes rédigés par 60 scripteurs. L'identification du scripteur était correcte dans 95.68% des documents considérés.

Al Dmour et al [ALD 07] présentent une technique d'extraction de caractéristiques basée sur la combinaison de mesures statistiques et spectrales, les auteurs effectuent l'extraction de deux types de caractéristiques issues de la texture, les premières sont obtenues en utilisant les filtres de Gabor, les secondes à partir des matrices de cooccurrences en niveau de gris. Les caractéristiques les plus discriminantes ont été choisies en utilisant un module hybride basé sur les algorithmes génétiques et les SVM. Quatre classifieurs ont été testés: SVM, KPPV (K plus proches voisins), un classifieur linéaire discriminant et un autre utilisant la distance euclidienne pondérée. Un taux d'identification de 90% a été atteint sur une base de 20 scripteurs.

Shahabi et al [SHA 06] déterminent la performance des caractéristiques utilisées dans [SAI 00] (voir section 2.3.2) sur une base de données de Farsi manuscrit. Les auteurs proposent une approche globale, basée sur la texture de l'image, en mode indépendant du texte. Des caractéristiques sont extraites à l'aide des filtres de Gabor et les matrices de cooccurrences. Les expérimentations effectuées sur une base de 25 scripteurs donnent un taux d'identification de 88% en Top 1 et 92% en Top 3.

Dans d'autres travaux, Shahabi et al [SHA 07] proposent une approche d'identification et de vérification de scripteurs en mode dépendant du texte, où ils choisissent parmi les filtres de Gabor ceux qui sont appropriés à la structure des textes Farsi manuscrits. Ils utilisent une nouvelle méthode d'extraction de caractéristiques qui se base sur les moments et les transformations linéaires. Deux niveaux ont été considérés dans leurs séries de tests : le niveau mot et le niveau texte. Le taux d'identification correcte est de l'ordre de 45% en Top 1 et 80% en Top 5 au niveau mot, il atteint 82,5% en Top 1 et 100% en Top 5 au niveau texte sachant que 40 scripteurs ont été considérés.

Hamid et Loewenthal [HAM 96] ont mené une étude utilisant des échantillons d'écriture en Anglais et en Urdu. Les auteurs ont présenté la finesse et l'aspect décoratif de l'écriture comme facteurs discriminants majeurs entre les scripteurs de sexe masculin et féminin. Un taux de

classification conséquent d'environ 68% a été atteint sur les échantillons d'écriture Urdu et Anglais.

Liwicki et al [LIW 11], utilisent une combinaison de caractéristiques extraites à partir des écritures manuscrites en-ligne et hors-ligne afin de permettre la détermination du sexe d'un scripteur quelconque. Un taux de bonne classification de l'ordre de 67,5% en utilisant les modèles de mélanges gaussiens (GMM : Gaussian mixture models) comme classifieurs a été enregistré. L'étude affirme également que les informations obtenues à partir des documents en-ligne donnent une meilleure classification contrairement aux caractéristiques extraites à partir d'une représentation hors-ligne des documents en-ligne. Les détails des caractéristiques utilisées dans cette étude peuvent être trouvés dans [LIW 06]

Dans une étude relativement récente [SOK 12], Sokik et al utilisent les descripteurs de Fourier, des informations sur la courbure ainsi que l'énergie de flexion pour caractériser le sexe des scripteurs d'échantillons manuscrits. Les résultats obtenus ne sont pas cependant présentés dans une forme quantifiée et seules les valeurs des différentes caractéristiques calculées à partir du même mot pour des scripteurs de sexe masculin et féminin sont présentées et discutées.

2.4. Analyse combinée pour la classification de scripteurs

Dans cette section, nous allons présenter les méthodes de classification de scripteurs basées sur la combinaison de caractéristiques globales et locales.

2.4.1. Travaux de Bulacu & al (2007)

Bulacu et al. [BUL 07a] ont présenté un travail portant sur la combinaison de caractéristiques texturales et allographiques pour la reconnaissance de scripteurs arabes. Les auteurs utilisent, pour l'évaluation des performances de l'approche proposée, la base IFN/ENIT [PEC 02] (voir section 1.6.7) contenant des documents écrits par 411 scripteurs différents. La base de données a été divisée en deux parties, la première contient les documents issus de 61 scripteurs différents et a été employée pour la construction du codebook de graphèmes utilisés dans le niveau allographique de l'approche. La plus grande partie de la base de données, contenant des documents écrits par les 350 scripteurs restants (avec 5 pages par scripteur) a été employée dans les séries de tests de reconnaissance de scripteurs.

Ainsi, cinq pages au format A4 ont été scannées avec une résolution de 300 dpi pour chaque scripteur: Les auteurs ont considéré un seul et unique corpus de 1750 documents (5 échantillons x 350 scripteurs) : à chaque fois un seul document est sélectionné pour servir de document de test. Les auteurs utilisent les caractéristiques suivantes:

Au niveau textural, l'extraction de caractéristiques commence par une convolution suivie d'un seuillage afin de générer une image binaire, où seuls les pixels des contours sont visibles. Ces derniers sont positionnés au milieu d'un carré où les voisinages sont vérifiés avec un opérateur *ET* logique, du centre vers les périphéries, pour détecter l'éventuelle présence d'un fragment entier d'un contour (voir figure 2.23). Toutes les instances vérifiées sont prises en compte dans un histogramme qui est finalement normalisé par la distribution de probabilité $P(\varphi)$, qui correspond à la probabilité de trouver dans l'image un fragment de contour orienté de l'angle φ mesuré par rapport à l'horizontale.

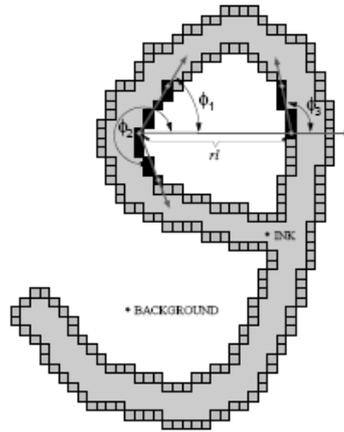


Figure 2.23. Exemple illustrant la méthode d'extraction des directions des contours à partir de la lettre arabe « g » [BUL 07a].

Les auteurs ont introduit une caractéristique supplémentaire : la distribution des charnières des contours qui estime la distribution de la probabilité jointe $P(\varphi_1, \varphi_2)$ qui quantifie la probabilité de trouver, dans l'image, deux fragments de contour orientés selon les angles φ_1 et φ_2 . Les auteurs ont également évalué d'autres caractéristiques pour la reconnaissance de scripteurs, les distributions de longueurs de segments blancs horizontaux et verticaux [ARA 77], les cooccurrences horizontales et verticales des angles des contours.

Au niveau allographique, le scripteur est considéré comme un générateur de formes appelées graphèmes. Les auteurs réalisent une segmentation de l'écriture en graphèmes en se basant sur les minimums verticaux des composantes connexes. Les fonctions de densités de probabilités de ces graphèmes sont calculées en utilisant un "codebook" (voir figure 2.24) commun obtenu en regroupant les 35000 graphèmes extraits à partir l'écriture de 61 scripteurs différents en utilisant l'une des trois méthodes de regroupement (clustering) suivantes : les k-means ainsi que les cartes de Kohonen auto-organisatrices 1D & 2D [KOH 89, ALP 04, BEL 92, BIS 06, COR 02, MIT 97]. La taille du codebook a été fixée à 400 formes (20x20). Cette valeur a été employée également dans les expérimentations des auteurs sur le script latin [BUL 05].

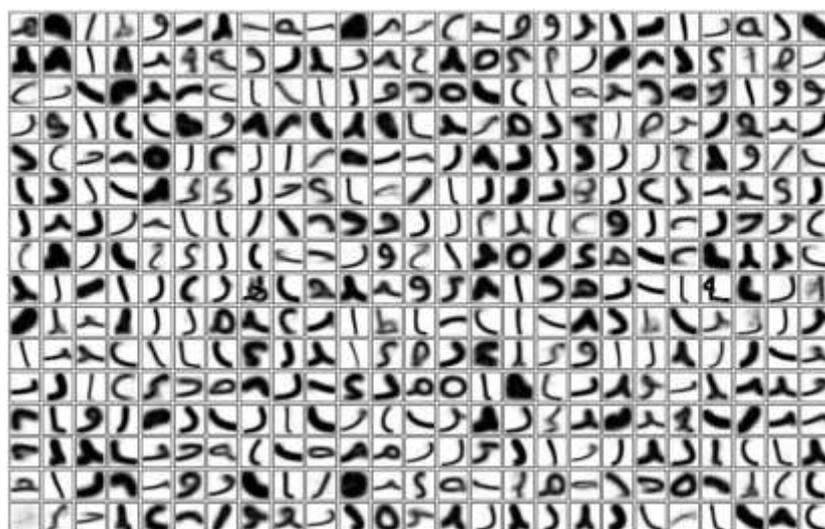


Figure 2.24. Un Codebook contenant 400 graphèmes arabes [BUL 07a].

L'évaluation de ce système d'identification a été réalisée à l'aide d'un classifieur de type plus proche voisin. Une page d'écriture est choisie parmi les 1750 récoltées (5 échantillons x 350 scripteurs), puis une distance euclidienne est calculée entre la page choisie et les pages restantes dans le corpus. Le meilleur taux d'identification atteint est de l'ordre de 88 % en Top 1 et 99% en Top 10 en utilisant de multiples caractéristiques. Pour la vérification, le meilleur taux d'erreurs égales EER (Equal Error Rate) enregistré est de l'ordre de 5.6%.

Les auteurs ont bien choisi la base de données sur laquelle ils ont effectué leurs expérimentations. En effet, la base IFN/ENIT est la base la plus utilisée par les chercheurs dans le domaine de l'analyse et la reconnaissance de documents arabes. Elle renferme des documents qui n'ont pas le même contenu textuel dans les cinq échantillons fournis par chaque scripteur. Dans leurs travaux de recherche, Schomaker et Bulacu ont abordé la reconnaissance de scripteurs à partir de documents latins [BUL 05, SCL 04a, SCL 04b] et arabes [BUL 07a]. Ils ont conclu que le script arabe est plus difficile à modéliser que le script latin.

2.4.2. Travaux de Siddiqi & al (2010)

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 imagerie de taille 13 × 13 pixels (voir figure 2.25). Ces imagerie 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 2.26).

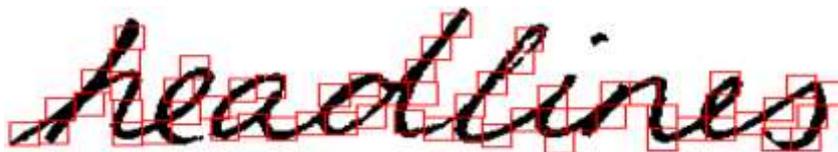


Figure 2.25. Illustration du découpage adaptatif de l'image du mot «headlines» en imagerie [SID 09].

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.

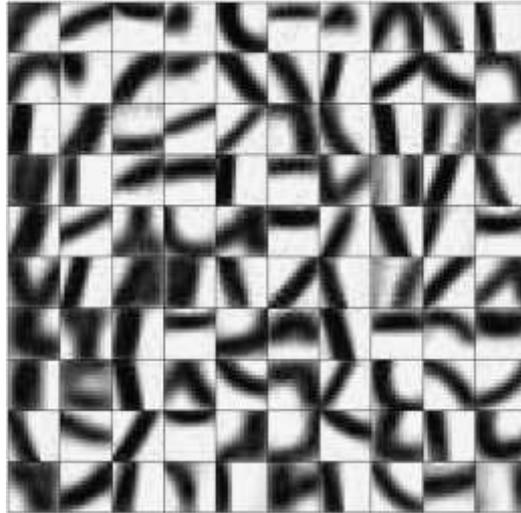


Figure 2.26. Un codebook universel de taille 100 obtenu à partir d'échantillons de la base RIMES [SID 10].

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.

Enfin, la combinaison des caractéristiques allographiques et structurelles produit des taux d'identification de l'ordre de 91% et 84% ainsi que des erreurs de vérification de l'ordre de 2.23% et 4.90% pour les bases IAM et RIMES, respectivement. Quand les deux bases de données sont combinées ensemble pour avoir une plus grande base (de 1025 scripteurs), un taux d'identification de l'ordre de 88% ainsi qu'une erreur de vérification de l'ordre de 2.86% sont enregistrés.

2.5. Compétitions sur la classification de scripteurs

L'intérêt des chercheurs à la classification de scripteurs et aux méthodes d'évaluation de ces systèmes de classification a 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 de 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 des conférences 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
ICDAR 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 [MAL 13]	Identification de scripteurs	Latin	50	6	8	Hors-ligne
ICDAR2013 - Competition on Gender Prediction from Handwriting [HAS 13]	Identification du sexe	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.

2.5.1. Compétitions sur l'identification de scripteurs

Depuis 2011, sept compétitions sur l'identification de scripteurs ont été proposées, ces différentes compétitions ont traité des aspects différents tels que l'évaluation de la capacité des systèmes soumis à rejeter des scripteurs qui n'ont pas été inclus lors de la phase d'apprentissage [HAS 11, HAS 12], l'identification de scripteurs à partir de partitions musicales [FOR 11] ainsi que l'identification de scripteurs à partir de documents multilingues (Grec, Allemand, Anglais, Français) et à partir de documents contenant un petite quantité d'écriture [LOU 11, LOU 12, LOU 13, MAL 13]. Comme décrit dans le tableau 2.1, toutes les compétitions proposées jusqu'à présent se sont concentrées sur l'écriture manuscrite hors-ligne. Aucune compétition n'a été proposée pour les documents en ligne.

2.5.1.1. Compétitions sur l'identification de scripteurs à partir de documents arabes

ICDAR2011 Arabic Writer Identification Competition [HAS 11], est la première édition d'une série de compétitions sur l'identification de scripteurs arabes, elle a été organisée en conjonction avec la conférence ICDAR 2011 par une équipe de recherche du département d'informatique et d'ingénierie de l'université de Qatar, Doha, Qatar, ainsi qu'une autre équipe de l'université de Northumbria, London, Royaume-Uni. Cette édition a attiré 30 participants dont 7 uniquement ont accepté de partager leur méthodologie et leur identité.

Le système qui a été classé en première position lors de cette édition a été proposé par Andrew Newell et Lewis Griffin du département d'informatique de l'université de London, Royaume-Uni. Ce système a réalisé un taux d'identification de l'ordre de 100% sur une base de données composée de 54 scripteurs arabes différents en utilisant des caractéristiques texturales et un classifieur de type plus proche voisin. Une description détaillée des caractéristiques utilisées par ce système est disponible dans [CRO 10]. Il est intéressant de noter que cette édition avait, comme objectif secondaire, l'évaluation de la capacité des systèmes soumis à rejeter des scripteurs qui n'ont pas été appris au préalable.

L'édition 2012 de cette compétition "The ICFHR2012 Competition on Writer Identification - Challenge 2: Arabic Scripts" [HAS 12], a été organisée en marge de la conférence ICFHR 2012. La seule différence avec l'édition précédente réside dans le fait que la base de données utilisée lors de cette édition a été élargie pour contenir des documents issus de 206 scripteurs arabes différents ce qui a entraîné une augmentation de la difficulté de la tâche à accomplir.

Pour cette deuxième édition, le nombre de participants a augmenté pour atteindre le nombre de 43 dont 16 uniquement ont accepté de partager leur méthodologie et leur identité.

Lors de cette compétition, deux systèmes ont été classés en première position en réalisant un taux d'identification de l'ordre de 95.30%. Ces systèmes ont été soumis par Andrew Newell et Lewis Griffin [CRO 10], les gagnants de la première édition, ainsi que Wayne Zhang du département de d'ingénierie de l'information de l'université chinoise de Hong Kong.

Le système proposé par Wayne Zhang combine les caractéristiques fournies par les organisateurs de la compétition avec les distributions des charnières des contours ainsi que les distributions des graphèmes introduites dans [VAN 05]. L'analyse en composantes principales avec noyau [HAS 12] est utilisée après l'application d'une analyse discriminante linéaire avec un tirage aléatoire d'un échantillon de la base de données [HAS 12], afin de réduire la dimensionnalité des vecteurs de caractéristiques. Finalement, les séparateurs à vaste marge ont été employés pour la classification.

2.5.1.2. Compétitions sur l'identification de scripteurs à partir de documents multilingues (latins et grecs)

ICDAR 2011 Writer Identification Contest [LOU 11], est la première édition d'une série destinée à l'évaluation de la capacité des systèmes d'identification de scripteurs à déterminer l'identité d'un scripteur à partir de documents multilingues et à partir d'une petite quantité d'écriture (environ deux lignes de texte). Elle s'est déroulée dans le cadre de la conférence ICDAR 2011, elle a été proposée par une équipe de recherche au Laboratoire d'intelligence computationnelle de l'institut d'Informatique et des Télécommunications du Centre National Grec de Recherche Scientifique "Demokritos", Athènes, Grèce. Sept équipes de recherche ont présenté leurs systèmes pour cette compétition. Une de ces équipes de recherche a soumis deux systèmes différents portant le nombre total de systèmes soumis à huit.

Le système proposé par Xu et al [XU 11] du département de Génie électronique de l'université de Tsinghua, Pékin, Chine, a été classé en première position lors de cette édition, il consiste à analyser les textes manuscrits en multi-lingue et à utiliser une fenêtre de grille mobile afin d'extraire un ensemble de micro-caractéristiques pour caractériser les différents styles d'écriture. L'identification de scripteurs a été effectuée à l'aide de la distance du *chi-square* pondérée. Ce système fonctionne très bien dans la plupart des expériences, même avec les images de documents contenant une petite quantité d'écriture qui demeure jusqu'à présent une tâche très difficile.

The ICFHR2012 Competition on Writer Identification Challenge 1: Latin/Greek Documents [LOU 12] est la deuxième édition de la compétition sur l'identification de scripteurs à partir de documents latins et grecs, elle a été organisée dans le cadre de la conférence ICFHR 2012. La base de données a été étendue et élargie pour contenir des paragraphes écrits par 126 scripteurs en anglais et en grec. Sept systèmes ont participé à cette édition et l'un des trois systèmes que nous avons proposé a été classé en première position. Ce système est basé sur la combinaison des distributions des longueurs de segments (voir section 3.4.4) avec les distributions des charnières des contours. La distance de Manhattan a été utilisée pour la comparaison de deux documents manuscrits. ICDAR 2013 Competition on Writer Identification [LOU 13] représente la troisième édition de compétitions sur l'identification de scripteurs à partir de documents latins et grecs, elle a été organisée dans le cadre de la conférence ICDAR 2013. Six équipes de recherche ont présenté leurs systèmes pour cette édition. Certaines équipes ont soumis plusieurs systèmes, portant le nombre total de systèmes à douze.

Le système gagnant a été soumis par Rajiv Jain et David Doermann [JAI 13] de l'Université du Maryland, College Park, Etats-Unis. Ce système calcule des gradients à partir des contours afin de capturer la forme et la courbure locale, avec une segmentation de caractères afin de créer un pseudo-alphabet pour un échantillon d'écriture donné. Une mesure de distance est alors définie entre les éléments de ces alphabets pour capter la ressemblance entre deux échantillons d'écriture. Pour plus de détails, le lecteur peut se référer à l'article de Jain et al [JAI 13].

2.5.1.3. Compétition sur l'identification de scripteurs à partir de partitions musicales

The ICDAR 2011 Writer Identification on Music Scores Competition [FOR 11], est la troisième compétition qui a été organisée en conjonction avec la conférence ICDAR 2011, elle a été proposée par une équipe de recherche du Groupe d'analyse de documents du centre de vision par ordinateur de l'université Autonoma de Barcelona, Espagne. Huit systèmes différents ont été soumis par des équipes de recherche de l'université de Tébéssa, Algérie, de l'université Badji-Mokhtar d'Annaba, Algérie, de l'université de Northumbria, London, Royaume-Uni ainsi que de l'université de Qatar, Doha, Qatar. L'objectif essentiel de cette compétition consiste à évaluer la capacité des systèmes soumis à reconnaître le scripteur à partir de partitions musicales.

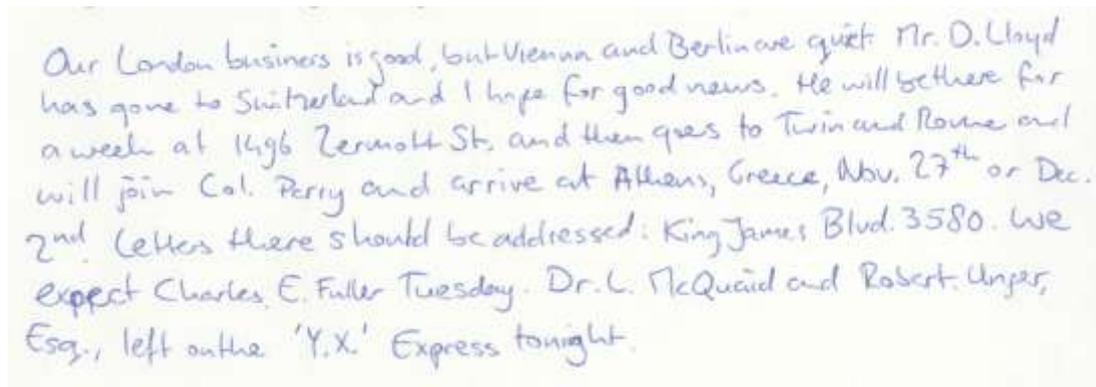
Le système classé en première position combine la distribution de probabilités des directions des contours avec des caractéristiques basées sur les graphèmes pour modéliser les scripteurs. Les organisateurs de la compétition ont proposé une base de données de 1000 partitions musicales issues de 50 musiciens différents. Sur un ensemble total de test de 500 documents, 358 documents (77%) ont été identifiés correctement par le système gagnant en utilisant un classifieur de type régression logistique. Plus de détails sur ces systèmes se trouvent dans [FOR 11].

2.5.1.4. Compétition sur l'identification de scripteurs à partir des écritures de styles différents issues du même scripteur

ICDAR 2013 Competition on Signature Verification and Writer Identification for Onand Offline Skilled Forgeries [MAL 13], est une compétition qui a été organisée en conjonction avec la conférence ICDAR 2013 par des chercheurs appartenant à plusieurs institutions scientifiques mondiales, ces institutions sont : le centre de recherche allemand sur l'intelligence artificielle, l'université suisse de Fribourg, l'institut médico-légal de La Haye, Pays-Bas, l'université

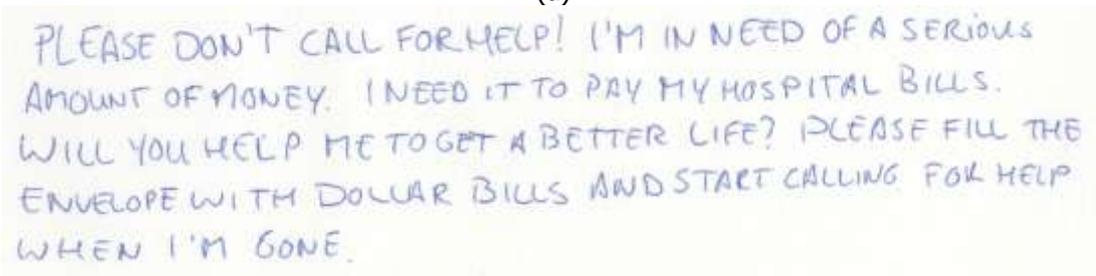
japonaise de Mie, l'Université australienne de Griffith, et le département australien des services de la police judiciaire de Victoria.

Les organisateurs considèrent un scénario souvent rencontré par les experts légaux en écriture où les morceaux de textes manuscrits à comparer et qui sont issus du même auteur sont écrits en styles complètement différents (voir figure 2.27). Huit systèmes provenant de chercheurs de trois équipes de recherches différentes ont été évalués en utilisant une base de documents manuscrits issus de 50 scripteurs. L'un des cinq systèmes que nous avons soumis à cette compétition a été classé en première position, ce système est basé sur la combinaison des distributions des longueurs de segments (voir section 3.4.4) avec les distributions des charnières des contours. La distance de Manhattan a été utilisée pour la comparaison de deux documents manuscrits.



Our London business is good, but Vienna and Berlin are quiet. Mr. D. Lloyd has gone to Switzerland and I hope for good news. He will be there for a week at 1496 Zermatt St. and then goes to Turin and Rome and will join Col. Perry and arrive at Athens, Greece, Nov. 27th or Dec. 2nd. Letters there should be addressed: King James Blvd. 3580. We expect Charles E. Fuller Tuesday. Dr. L. McQuaid and Robert Unger, Esq., left on the 'Y.X.' Express tonight.

(a)



PLEASE DON'T CALL FOR HELP! I'M IN NEED OF A SERIOUS AMOUNT OF MONEY. I NEED IT TO PAY MY HOSPITAL BILLS. WILL YOU HELP ME TO GET A BETTER LIFE? PLEASE FILL THE ENVELOPE WITH DOLLAR BILLS AND START CALLING FOR HELP WHEN I'M GONE.

(b)

Figure 2.27. Deux écritures en styles différents issues du même scripteur extraites de la base de la compétition ICDAR 2013 Competition on Signature Verification and Writer Identification for On- and Offline Skilled Forgeries [MAL 13], (a) style 1, (b) style 2.

2.5.2. Compétition sur la prédiction du sexe du scripteur

ICDAR2013 - Competition on Gender Prediction from Handwriting [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 décrite dans la section 1.6.8.

En résumé, 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. 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.

2.6. Synthèse de travaux de classification de scripteurs

Notre étude bibliographique nous a permis de constater qu'il existe plusieurs aspects qui peuvent influencer les performances des systèmes de classification de scripteurs, ces aspects sont le nombre de scripteurs, la quantité d'écriture disponible sur les échantillons, le nombre d'échantillons par scripteur, le script considéré ainsi que les caractéristiques utilisées. Avant d'effectuer des expérimentations, les chercheurs doivent faire des choix sur ces différents aspects.

Publication	Taille de la base de données	Taille de l'échantillon	Caractéristiques utilisées	Script considéré	Type de données	Tâche	Mode de reconnaissance
Audrey Seropian [SER 03]	20 scripteurs	Texte	Formes invariantes	Latin	Hors-ligne	Reconnaissance du scripteur	Indépendant
Bensefia et al. [BEN 04]	88 et 150 scripteurs	Texte	Graphèmes	Latin	Hors-ligne	Reconnaissance du scripteur	Indépendant
Pervouchine et al. [PER 07]	200 scripteurs	Caractère	Taille, largeur, allongement, présence de la boucle, structure de la barre de la lettre t, structure de la tige de la lettre h, épaisseur de traits, nombre de traits,... etc.	Latin	Hors-ligne	Reconnaissance du scripteur	Dépendant
Tan et al. [TAN 09]	120 scripteurs	Texte	Prototypes d'allographes de lettres	Latin	En-ligne	Reconnaissance du scripteur	Indépendant
Woodard et al. [WOO 10]	51 scripteurs	Texte	Descripteurs SIFT	Arabe	Hors-ligne	Reconnaissance du scripteur	Dépendant
Jain et al. [JAI 11]	302 et 650 scripteurs	Texte	Descripteurs k-AS	Latin et Arabe	Hors-ligne	Reconnaissance du scripteur	Indépendant
Zois et al. [ZOI 00]	50 scripteurs	Mot	Caractéristiques morphologiques	Latin et Grec	Hors-ligne	Reconnaissance du scripteur	Dépendant
Said et al. [SAI 00]	40 scripteurs	Bloc de texte	Filtres de Gabor et matrices de cooccurrences	Latin	Hors-ligne	Reconnaissance du scripteur	Indépendant
Marti et al. [MAR 01]	20 scripteurs	Texte	Hauteur, largeur, inclinaison et lisibilité des caractères	Latin	Hors-ligne	Reconnaissance du scripteur	Indépendant
Srihari et al. [SRI 02]	1000 scripteurs 300 scripteurs	Texte, paragraphe, mot	Micro et macro caractéristiques	Latin	Hors-ligne	Reconnaissance du scripteur et reconnaissance du sexe	Dépendant
Gazzah et al [GAZ 05, GAZ 06]	60 scripteurs	Ligne de texte	Hauteur de la ligne de texte, espace moyen entre pseudo-mots, inclinaison des hampes, caractéristiques dimensionnelles des points diacritiques, mesure de l'entropie, transformée en ondelettes	Arabe	Hors-ligne	Reconnaissance du scripteur	Indépendant
Feddaoui et al [FED 06]	22 scripteurs	Bloc de texte	Filtres de Gabor	Arabe	Hors-ligne	Reconnaissance du scripteur	Indépendant
Zhu et al. [ZHU 00]	17 scripteurs	Bloc de texte	Filtres de Gabor	Chinois	Hors-ligne	Reconnaissance du scripteur	Indépendant
Helli et al. [HEL 08]	50 scripteurs	Bloc de texte	Filtres de Gabor	Arabe	Hors-ligne	Reconnaissance du scripteur	Indépendant
Schlabach et al. [SCL 07]	100 scripteurs	Ligne de texte	Modèles de Markov cachés	Latin	Hors-ligne	Reconnaissance du scripteur	Indépendant
Bertolini et al [BER 13]	315 et 650 scripteurs	Texte	Caractéristiques texturales (LPQ et LVQ)	Latin	Hors-ligne	Reconnaissance du scripteur	Indépendant et dépendant
Chaabouni et al. [CHB 14]	170 scripteurs	Mot	Caractéristiques multi-fractales	Arabe	En et Hors-ligne	Reconnaissance du scripteur	Dépendant
Wang et al. [WAN 03, WAN 04]	25 et 27 scripteurs	Caractère	caractéristiques directionnelles	Chinois	Hors-ligne	Reconnaissance du scripteur	Dépendant
Garain et al. [GAR 09]	40 et 382 scripteurs	Texte	Coefficients autorégressifs bidimensionnels	Latin et Bengali	Hors-ligne	Reconnaissance du scripteur	Indépendant
Al Zoubeidy et al. [ALZ 05]	20 scripteurs	Bloc de texte	Filtres de Gabor et les matrices de cooccurrences	Arabe	Hors-ligne	Reconnaissance du scripteur	Indépendant
Gazzah et al [GAZ 07]	60 scripteurs	Ligne de texte	Ondelettes 2D utilisant le Lifting Scheme	Arabe	Hors-ligne	Reconnaissance du scripteur	Indépendant
Al Dmour et al [ALD 07]	20 scripteurs	Bloc de texte	Filtres de Gabor et les matrices de cooccurrences	Arabe	Hors-ligne	Reconnaissance du scripteur	Indépendant
Shahabi et al [SHA 06]	25 scripteurs	Bloc de texte	Filtres de Gabor et les matrices de cooccurrences	Arabe	Hors-ligne	Reconnaissance du scripteur	Indépendant
Shahabi et al [SHA 07]	40 scripteurs	Mot, texte	Filtres de Gabor, moments, transformations linéaires	Arabe	Hors-ligne	Reconnaissance du scripteur	Dépendant
Chanda et al. [CHA 10]	104 scripteurs	Texte	Caractéristiques directionnelles et du gradient	Bengali	Hors-ligne	Reconnaissance du scripteur	Indépendant
Purkait et al. [PUR 10]	22 scripteurs	Mot	caractéristiques de l'orientation et la courbure	Telugu	Hors-ligne	Reconnaissance du scripteur	Indépendant
Chanda et al. [CHA 12]	100 scripteurs	Texte	caractéristiques de la courbure	Oriya	Hors-ligne	Reconnaissance du scripteur	Indépendant
Bulacu et al. [BUL 07a]	350 scripteurs	Texte	Charnières des contours, directions de contours, longueurs de segments, cooccurrences des angles des contours et graphèmes	Arabe	Hors-ligne	Reconnaissance du scripteur	Indépendant
Siddiqi et al. [SID 10]	1025 scripteurs	Texte	Formes invariantes et caractéristiques de l'orientation et la courbure	Latin	Hors-ligne	Reconnaissance du scripteur	Indépendant
Hamid et al. [HAM 96]	30 scripteurs	Texte	Finesse et aspect décoratif de l'écriture	Latin & Arabe	Hors-ligne	Reconnaissance du sexe	Indépendant
Liwicki et al. [LIW 11]	150 scripteurs	Texte	Vitesse et accélération, direction et courbure	Latin	En et Hors-ligne	Reconnaissance du sexe	Indépendant
Sokic et al [SOK 12]	Inconnu	Mot	Descripteurs de Fourier, courbure et énergie de flexion	Latin Serbe	Hors-ligne	Reconnaissance du sexe	Dépendant

Tableau 2.2. Aperçu des méthodes proposées dans le domaine de la classification de scripteurs.

Nous avons présenté certains des travaux précédents afin d'illustrer les choix qui ont été faits et comment ils influencent les performances des systèmes de classification de scripteurs.

Il est évident que l'utilisation d'un grand nombre de scripteurs dans la base de données augmente la crédibilité des résultats de classification avancés. Selon cet aspect, Srihari et al. [SRI 02] et Siddiqi et al. [SID 10] ont utilisé les plus grandes bases de données (1000 et 1025 scripteurs respectivement), bien loin des bases de données de Al Dmour et al [ALD 07], Audrey Seropian [SER 03], Said et al. [SAI 00], Marti et al. [MAR 01], Feddaoui et al [FED 06], Al Zoubeidy et al. [ALZ 05] et Feddaoui et al [FED 06] qui ne comportent qu'une vingtaine de scripteurs.

La quantité d'écriture disponible sur les échantillons d'écriture est un aspect très important qui peut influencer considérablement les performances des systèmes de classification de scripteurs. Cet aspect demeure l'écueil qui empêche les chercheurs du domaine de progresser dans leurs travaux. Avoir une quantité suffisante d'écriture sur les échantillons d'écriture à utiliser, aussi bien pour l'entraînement que pour l'évaluation des systèmes de classification de scripteurs, est primordial afin de dériver des caractéristiques stables assurant une bonne caractérisation des écritures des différents scripteurs. Siddiqi et al. [SID 10] ont effectué des expérimentations afin de déterminer la quantité d'écriture nécessaires pour la reconnaissance de scripteur. Les résultats ont révélé que trois lignes de texte d'environ une trentaine de mots sont suffisantes pour la réussite de la reconnaissance de scripteur.

Notre étude bibliographique nous a amenés à une remarque très importante : le nombre d'échantillons disponibles par scripteur sont d'une importance considérable pour les systèmes de classification de scripteurs en général, et, plus particulièrement pour les systèmes dépendants du texte utilisant une quantité d'écriture de l'ordre d'un caractère, d'un mot ou d'une ligne de texte.

Le système de Pervouchine et al. [PER 07] a été évalué en utilisant entre 24 et 30 échantillons de caractères par scripteur, 45 mots étaient nécessaires pour le système de Zois et al. [ZOI 00] alors que les systèmes d'identification et de vérification de scripteurs de Wang et al. [WAN 03, WAN 04] nécessitaient 16 échantillons de caractères par scripteur afin de donner des performances acceptables.

Nous avons constaté aussi que la plupart des travaux consacrés à la classification de scripteurs sont basés sur l'utilisation d'échantillons d'écriture latine [BUL 07, KEM 06, NOS 04, SCH 07, BEN 05, SID 10, GAR 09, BER 13, SCL 04a, SCL 04b, SRI 02, MAR 01, SAI 00, JAI 11, TAN 09, PER 07, BEN 04, SER 03]. Ce n'est que depuis une décennie que les chercheurs se sont orientés vers la proposition de méthodes pour les scripts grec [ZOI 00], chinois [ZHU 00, WAN 03, WAN 04], arabe [BUL 07a, ALD 07, GAZ 07, ALZ 05, CHB 14, FED 06, GAZ 05, GAZ 06, WOO 10], Bengali [GAR 09, CHA 10], Telugu [PUR 10] et l'Oriya [CHA 12]. Nous pouvons confirmer que tous les systèmes développés jusqu'à présent sont des systèmes mono script, qui ont été développés pour être testés uniquement dans un environnement mono script, et personne n'a pensé à élargir ces systèmes afin qu'ils soient opérationnels et performants dans des environnements multi-scripts.

Les caractéristiques utilisées pour la modélisation des scripteurs est l'un des aspects les plus distinctifs entre les systèmes de classification de scripteurs. Nous avons constaté que les caractéristiques utilisées, sont majoritairement globales, elles sont extraites de l'ensemble du texte représentant le scripteur. Ces caractéristiques se classent en trois grandes catégories :

- Caractéristiques texturales ou fractales où le contenu du document est vu simplement comme une image et non comme une écriture. L'application de plusieurs techniques d'analyse de textures a été envisagée par plusieurs chercheurs, on peut citer par exemple les filtres de Gabor et / ou des matrices de cooccurrences [ALD 07, ALZ 05, FED 06, SHA 06, SHA 07, SAI 00], les coefficients autorégressifs bidimensionnels [GAR 09], les descripteurs LPQ et LVQ [BER 13], les Charnières des contours, les directions de contours, les longueurs de segments, les cooccurrences des angles des contours [BUL 07a] ainsi que les ondelettes 2D utilisant le Lifting Scheme [GAZ 07]. L'analyse multi fractale a été envisagée par Chaabouni et al. dans [CHB 14]
- Caractéristiques structurelles qui s'attachent à décrire les particularités morphologiques de l'écriture. On peut citer par l'exemple des caractéristiques telles : la hauteur de la ligne, la moyenne des inclinaisons des hampes, l'écart type des inclinaisons des hampes, la moyenne de l'épaisseur des hampes, la moyenne du ratio hauteur/largeur du rectangle englobant un point diacritique et la densité de pixels dans un rectangle englobant un point diacritique [GAZ 05, GAZ 06, GAZ 07].

D'autres chercheurs du domaine ont présenté des études ne prenant pas en compte les caractéristiques globales du texte représentant le scripteur, mais ils ont utilisé des caractères, des graphèmes et parfois même des petits fragments des caractères pour caractériser les écritures des scripteurs, l'inconvénient majeur des caractéristiques locales réside dans le fait que leur extraction dépende fortement de processus de segmentation employé, l'échec de ce dernier peut influencer considérablement l'étape d'extraction de caractéristiques.

Nous tenons enfin à faire une dernière remarque qui est que la majorité des travaux publiés dans le domaine de la classification de scripteurs ont été consacrés au problème de la reconnaissance de scripteurs à l'inverse du problème de la détermination du sexe du scripteur qui a été malheureusement très peu abordé. En effet, seulement quatre travaux ont été recensés, il s'agit des méthodes proposées par Srihari et al [SRI 02] et Hamid et Loewenthal [HAM 96] et celles de Liwicki et al [LIW 11] et Sokic et al [SOK 12].

RECONNAISSANCE DU SCRIPTEUR EN MODE INDEPENDANT DU TEXTE EN UTILISANT DES DOCUMENTS MULTI-SCRIPTS

Ce chapitre présente l'une de nos contributions principales, qui consiste en la proposition d'une méthode indépendante du texte pour la reconnaissance du scripteur dans un environnement multi-scripts [DJE 13a]. Des textes manuscrits écrits en grec et anglais sont considérés dans le présent chapitre. L'objectif essentiel est de reconnaître le scripteur d'un texte écrit dans un script à partir d'échantillons du même scripteur dans un script différent et, par conséquent, valider l'hypothèse que le style d'écriture d'une personne demeure constant à travers différents scripts. Un autre aspect intéressant étudié dans ce chapitre est l'utilisation de textes courts, ce qui ressemble à des scénarios de la vie réelle, où les experts légaux, en général, ne trouvent que des petits morceaux de textes pour identifier ou vérifier l'identité d'un scripteur donné. La méthode proposée est basée sur la distribution de longueurs de segments qui est comparée avec les méthodes les plus connues et les plus performantes de l'état de l'art. La classification est réalisée en utilisant les K plus proches voisins (k-PPV) ainsi que les séparateurs à vaste marge (SVM). Les résultats expérimentaux obtenus sur une base de données de 126 scripteurs avec 4 échantillons par scripteur montrent que la méthode proposée permet d'obtenir des performances intéressantes en mode identification ainsi qu'en mode vérification dans un environnement multi-scripts.

3.1. Introduction

La reconnaissance du scripteur offre plusieurs applications utiles qui en font un domaine de recherche attrayant et prometteur en reconnaissance des formes [SCL 06]. La tâche de classification en reconnaissance de formes consiste à attribuer une forme à une classe parmi un ensemble de classes connues. Dans le cas de la reconnaissance du scripteur, une forme représente un échantillon de texte manuscrit tandis que chaque scripteur représente une classe, la tâche consiste alors à affecter l'échantillon en question à l'une des classes possibles (scripteurs).

Les progrès notables dans le domaine de la reconnaissance de scripteurs à partir de textes manuscrits ont ouvert de nouvelles directions de recherches durant la dernière décennie. En effet, plusieurs chercheurs proposent, dans des travaux assez récents, de combiner différents types de caractéristiques [BUL 07, DJE 10], d'introduire de nouvelles caractéristiques [BUL 07, SID 09a, SID 09b, SID 09c], d'examiner l'influence de la taille des caractères sur les performances des systèmes existants [OZA 06], d'évaluer les systèmes existants en utilisant des documents dégradés [CHE 11], d'explorer l'utilité de différencier les personnes par le biais de leurs traces écrites en-ligne et hors-ligne [CHB 11, CHB 14], d'identifier le scripteur à partir de notes musicales manuscrites [FOR 08] et, enfin d'utiliser des modèles immunologiques dans la tâche d'identification du scripteur [DJE 11].

La reconnaissance du scripteur à partir de documents monolingue est devenue un domaine de recherche qu'on pourrait qualifier de mature. Un grand nombre de scripts ont été considérés dans la littérature à savoir : le chinois [WAN 04], le japonais [KEM 06], l'arabe [BUL 07a, DJE 11, DJE 10], le bengali [CHA 10], le Telugu [PUR 10], l'Oriya [CHA 12] et le latin [BUL 07, KEM 06, NOS 04, SCH 07, BEN 05, SID 10]. La reconnaissance du scripteur dans un environnement multi-scripts reste, cependant, un domaine de recherche relativement inexploré avec une seule contribution à notre connaissance [GAR 09].

Le système d'identification de scripteurs proposé dans [GAR 09] a été évalué en combinant des images de textes manuscrits issues de la base de données RIMES [GRO 08] (contenant des textes en français appartenant à 382 scripteurs) avec d'autres images de textes manuscrits extraites de la base de données ISI (comprenant des textes bengali provenant de 40 scripteurs différents). Les deux bases de données n'ont, cependant, aucun scripteur commun entre elles et le système proposé par Garain et al ne se concentre pas sur l'étude du cas d'une personne écrivant plus d'un script (pour quelqu'un qui écrit, par exemple, en bengali et en français) pour déterminer si certaines caractéristiques communes existent dans l'écriture de cette même personne à travers différents scripts. Par conséquent, nous pouvons affirmer que tous les systèmes existants, à travers la littérature de la reconnaissance de scripteur, se limitent à reconnaître le scripteur de documents manuscrits rédigés dans une seule langue (ou script).

L'objectif essentiel de cette partie de notre travail est d'étendre la recherche en reconnaissance de scripteurs à un véritable environnement multi-scripts, en analysant s'il est possible de reconnaître le scripteur d'un texte écrit dans un script à partir d'échantillons du même scripteur dans un script différent.

Afin de développer un système de reconnaissance du scripteur à partir de documents multi-scripts, il est indispensable d'utiliser des caractéristiques indépendantes des scripts considérés. Les caractéristiques allographiques se sont révélées être très efficace pour la reconnaissance du scripteur à partir de documents monolingue. Les allographes sont toutefois très liés au script considéré et ne semblent donc pas être un choix intéressant dans un environnement multi-scripts. Nous croyons que l'utilisation de caractéristiques globales telles que celles basées sur la texture peut être très intéressante, car ces caractéristiques sont plus liées à l'écriture du scripteur qu'au script étudié. La méthode proposée est basée sur la distribution de longueurs de segments [GAL 75, DJE 10] qui représente une voie prometteuse pour caractériser les scripteurs de documents écrits dans différents scripts et langues. La performance de ces caractéristiques est également comparée à un certain nombre de méthodes de reconnaissance de scripteurs récentes [BUL 07, SID 10, GAR 09].

Le second objectif de notre étude est d'investiguer la reconnaissance du scripteur en utilisant des documents contenant une petite quantité de texte. En effet, d'après notre étude bibliographique, nous avons constaté que dans les systèmes traditionnels de reconnaissance du scripteur, une grande quantité d'écriture est requise pour caractériser et reconnaître correctement le scripteur d'un échantillon de texte donné. La performance de ces systèmes, en utilisant de petites quantités de texte, est susceptible à une dégradation importante. Pour des problèmes pratiques, en particulier, l'expertise judiciaire, la quantité de texte disponible est généralement très limitée. Dans de tels scénarios, les systèmes de reconnaissance du scripteur de l'état de l'art évalués sur de grandes quantités de texte ne peuvent pas réaliser des résultats acceptables lorsqu'on ne dispose que de quantités limitées de texte. La reconnaissance du scripteur à partir de documents manuscrits courts représente donc un domaine d'étude assez intéressant.

Le reste de ce chapitre est organisé comme suit: la section 3.2 donne une brève description de la base de données utilisée dans nos expérimentations. La section 3.3 décrit la méthode d'extraction de caractéristiques proposée ainsi que les expérimentations que nous avons entreprises pour montrer l'indépendance de ces caractéristiques par rapport au script utilisé. Nous présentons ensuite les classifieurs utilisés suivis par les résultats expérimentaux détaillés ainsi que leur analyse. Nous terminons ce chapitre par des conclusions relatives à cette partie de notre travail.

3.2. Base de données utilisée

Dans cette section, nous allons décrire la base de données que nous avons utilisée pour toutes les expérimentations que nous avons effectuées dans cette partie de notre travail. Nous allons présenter ensuite les similitudes et les différences entre les scripts latins et grecs.

3.2.1. Description de la base de données

Dans notre étude, nous avons utilisé la base de données fournie par les organisateurs de la Compétition de l'identification du scripteur "ICFHR2012 Competition on Writer Identification Challenge 1: Latin/Greek Documents" [LOU 12] (voir section 2.5.1.2). Cette base de données a été développée par une équipe de recherche à l'Institut d'informatique et de télécommunications au Centre National de la Recherche Scientifique "Demokritos", en Grèce. La base de données utilisée comprend des échantillons de 126 scripteurs grecs, chacun a contribué par quatre pages, deux en grec et deux en anglais. Les paragraphes manuscrits ont été numérisés et binarisés et ne comportent aucun élément non textuel. Les documents grecs correspondent à la langue maternelle des scripteurs. Deux échantillons de paragraphes grecs et anglais écrits par la même personne sont présentés dans la figure 3.1.

Socrates was a Classical Greek philosopher. Credited as one of the founders of western philosophy, he is an enigmatic figure known only

(a)

Ο Σωκράτης διδάσκει ότι η αρετή κατίζεται με την βοήθεια που απ'αυτην απορρέουν όλες οι άλλες αρετές, γιατί αυτές είναι το υπέρτατο αγαθό και την

(a)

Figure 3.1. Images de deux paragraphes du même scripteur.

3.2.2. Similitudes et différences entre les scripts latins et grecs

Après avoir présenté la base de données, nous résumons les principales similitudes et différences entre le grec et l'anglais. Cela permettra de soutenir l'argument que la méthode proposée est indépendante du script étudié.

En comparant les textes grecs et anglais (figure 3.2a), on peut voir qu'il ya des similitudes entre les deux. L'anglais et le grec sont lus et écrits de gauche à droite, ont des lettres majuscules et

minuscules (figure 3.2b), des espaces entre les mots (figure 3.2a) et partagent quelques lettres majuscules communes (figure 3.2b). En dépit de ces similitudes, il existe des différences importantes entre les deux. L'alphabet grec est seulement utilisé pour écrire le grec alors que l'alphabet latin est utilisé pour écrire un grand nombre de langues, l'anglais étant l'une de ces langues. La plupart des lettres ont des formes complètement différentes à travers les deux scripts. Beaucoup de lettres majuscules de l'alphabet anglais ressemblent à leurs lettres minuscules "imprimées", ce qui n'est pas du tout le cas en grec (figure 3.2b). L'alphabet anglais comprend vingt-six lettres majuscules ainsi que vingt-six lettres minuscules, contrairement à l'alphabet grec qui contient vingt-quatre lettres majuscules et vingt-cinq lettres minuscules (figure 3.2b). Alors que toutes les lettres de l'alphabet anglais conservent leurs formes originales, indépendamment de leur position dans un mot, une lettre de l'alphabet grec prend une forme différente quand elle apparaît à la fin d'un mot. Enfin, les accents et les signes diacritiques sont rares en anglais, mais sont un phénomène courant en grec (figure 3.2a).

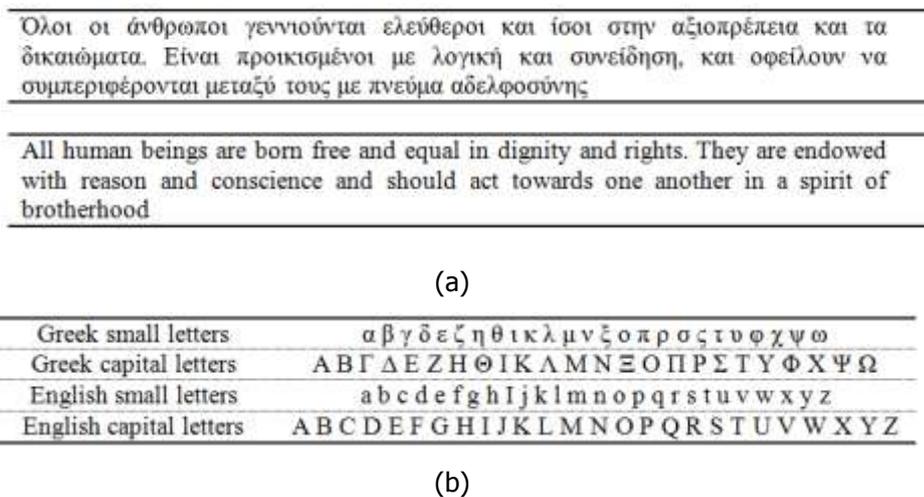


Figure 3.2. Textes grecs et anglais: (a) le même paragraphe écrit en grec et en anglais
(b) les lettres minuscules et majuscules des alphabets grec et anglais.

3.3. Extraction de caractéristiques

La méthode proposée est basée principalement sur un ensemble de caractéristiques extraites à partir des matrices de longueurs de segments [GAL 75, DJE 10]. Nous avons également implémenté quelques méthodes de l'état de l'art qui ont donné de très bons résultats en reconnaissance du scripteur à partir de documents monolingue.

3.3.1. Distribution de longueurs de segments (Run-length)

Nous caractérisons le style d'écriture d'un scripteur en calculant la distribution des séquences de pixels (Run-length). Ces caractéristiques sont déterminées à partir d'une image binaire de l'écriture, où les pixels noirs correspondent à la trace d'encre et les pixels blancs correspondent à l'arrière-plan. Les séquences de pixels sont calculées directement à partir de l'image complète et ne nécessitent pas de segmentation du texte en mots ou en caractères. Afin de calculer les séquences de pixels, l'image est balayée dans quatre directions principales: horizontale, verticale, diagonale gauche et diagonale droite. L'histogramme normalisé de ces séquences de pixels est interprété comme une probabilité de distribution afin de caractériser le scripteur.

Nous commençons par la définition d'un «run» comme un ensemble de pixels consécutifs ayant des niveaux de gris similaires, dans une direction donnée. Si $A_i A_j$ est un segment composé de pixels $A_i, A_{i+1}, \dots, A_{j-1}, A_j$ de couleurs identiques, le pixel A_{i-1} doit être de couleur différente de celle du pixel A_i , le pixel A_j doit être aussi de couleur différente de celle du pixel A_{j+1} .

Nous définissons ensuite une méthode statistique de caractérisation de la texture qui consiste à effectuer le comptage du nombre de segments de pixels de même intensité dans une direction donnée et représenter les résultats dans une matrice appelée matrice de longueur de segments P. Pour cela, une direction ($0^\circ, 45^\circ, 90^\circ$ ou 135°) et un nombre de niveaux de gris sont préalablement fixés. La valeur contenue dans la case (l, n) de la matrice est égale au nombre de segments de longueur l et de niveaux de gris n. Donc le nombre de colonnes de la matrice est dynamique car elle dépend de la longueur du plus long segment. De par sa conception, ce calcul est symétrique, il est par conséquent inutile de la calculer dans les quatre directions complémentaires ($180^\circ, 225^\circ, 270^\circ$ ou 315°) sachant qu'on considère ici huit directions possibles entre le pixel étudié et ses voisins). Ainsi, en général, la matrice est établie pour les quatre directions principales ($0^\circ, 45^\circ, 90^\circ$ et 135°).

La méthode d'extraction de caractéristiques proposée est illustrée par un exemple dans la figure 3.3. Nous considérons une image de 8×6 avec deux couleurs $C = \{0, 1\}$. Cette image représente le chiffre '2'.

0	1	1	1	1	0
1	0	0	0	0	1
1	0	0	0	0	1
0	0	0	0	1	0
0	0	0	1	0	0
0	0	1	0	0	0
0	1	0	0	0	1
1	1	1	1	1	1

(a)

2	1	2	2	1	1
9	2	0	0	0	1

(b)

3	2	3	2	1	0	0	0
11	4	0	0	0	0	0	0

(c)

7	3	2	1	0	1
15	2	0	0	0	0

(d)

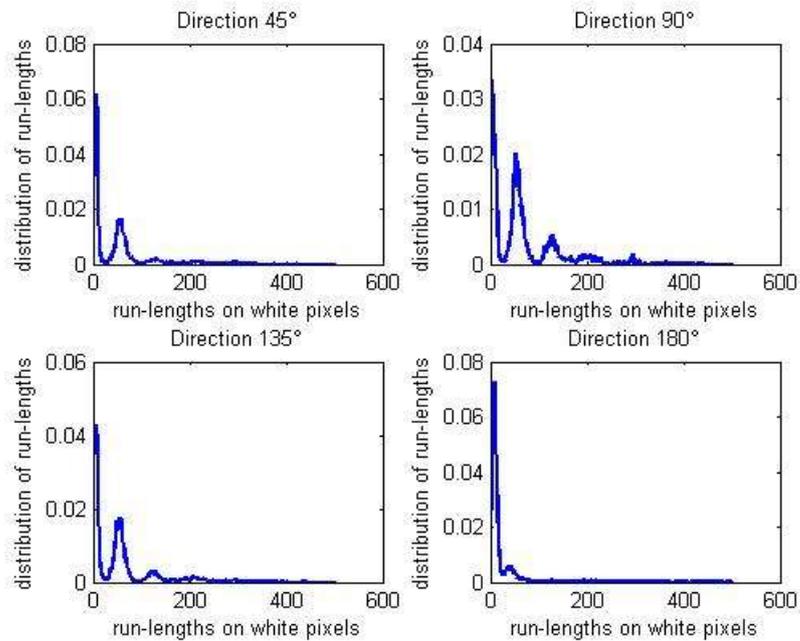
4	2	3	3	0	0
9	0	0	1	0	1

(e)

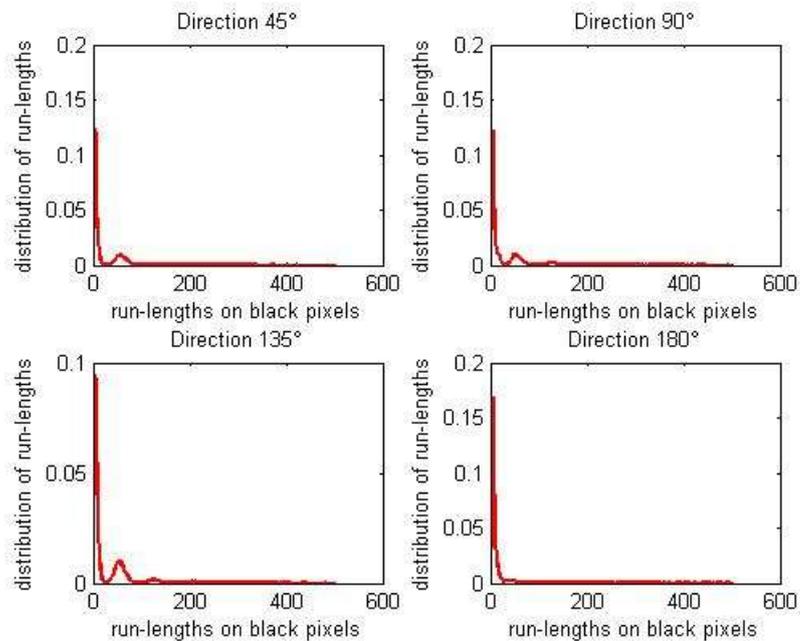
Figure 3.3. Calcul de la matrice de longueur de segments (a) Une image 8×6 avec deux valeurs de couleurs (0 et 1) (b) matrice de longueur de segments pour la direction 45° (c) matrice de longueur de segments pour la direction 90° (d) matrice de longueur de segments pour la direction 135° (e) matrice de longueur de segments pour la direction 180° .

Chaque élément de la matrice indique le nombre de fois où l'image contient un segment d'une longueur $\{1, 2, 3, 4, 5, 6\}$ dans les directions $45^\circ, 135^\circ$ et 180° et des segments d'une

longueur de $\{1, 2, 3, 4, 5, 6, 7, 8\}$ dans la direction 90° . Le premier élément de la première ligne de la matrice correspond au nombre de fois où la couleur 0 apparaît isolément, le deuxième élément est le nombre de fois où la couleur 0 apparaît en segments de deux pixels et ainsi de suite... La ligne suivante capte les mêmes informations à partir de l'image pour la couleur 1.



(a)



(b)

Figure 3.4. Distributions des longueurs des segments dans les directions (0° , 45° , 90° et 135°) sur (a) Pixels blancs. (b) Pixels noirs.

Les quatre matrices à longueurs de segments sont converties en vecteurs normalisés qui sont ensuite concaténés pour obtenir un vecteur unique caractérisant le scripteur d'un document. Cela conduit naturellement au problème de la grande dimension du vecteur de caractéristiques. Cependant, on peut remarquer que la plupart des valeurs non nulles sont contenues dans les premières colonnes de la matrice (voir figure 3.4). Nous gardons donc uniquement un sous-ensemble de colonnes (choisi empiriquement) pour chacune des matrices. Pour les longueurs de segments extraits à partir de pixels noirs, nous ne gardons que les 100 premières colonnes pour chacune des quatre directions alors que pour les pixels blancs nous gardons les 50 premières colonnes. Cela nous donne un total de 600 éléments ($100 \times 50 \times 4 + 4$) pour chaque document manuscrit.

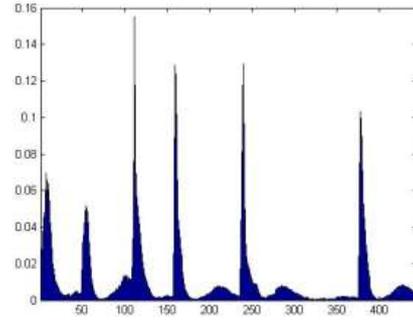
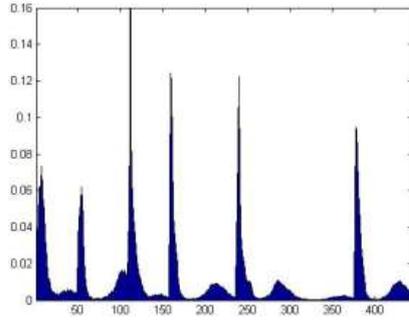
Ces caractéristiques de longueurs de segments sont similaires à celles que nous avons utilisées dans le système avec lequel nous avons participé à la compétition d'identification de scripteurs (ICDAR2011 Writer Identification Contest) [LOU 11].

Nous avons également participé, avec une partie de ces caractéristiques aux compétitions d'identification de scripteurs suivantes :

1. The ICDAR2011 Arabic Writer Identification Competition [HAS 11],
2. The ICDAR 2011 Writer Identification on Music Scores Competition [FOR 11],
3. The ICFHR2012 Competition on Writer Identification - Challenge 2: Arabic Scripts [HAS 12],
4. The ICFHR2012 Competition on Writer Identification Challenge 1: Latin/Greek Documents [LOU 12].
5. The ICDAR 2013 Competition on Writer Identification [LOU 13].
6. The ICDAR 2013 Competitions on Signature Verification and Writer Identification for On- and Offline Skilled Forgeries (SigWiComp2013) [MAL 13].

Le système d'identification du scripteur basé sur ces caractéristiques a réalisé la troisième meilleure performance à la compétition d'identification du scripteur [LOU 11] et le deuxième meilleur résultat à la compétition d'identification du scripteur en utilisant des partitions musicales [FOR 11]. Ces résultats sont un bon indicateur sur l'efficacité de ces caractéristiques dans la caractérisation des scripteurs. Ces caractéristiques n'utilisent aucune connaissance sur le script utilisé et leur performance est comparable à celle de la méthode proposée dans [SID 10] alors qu'elles sont plus performantes que les caractéristiques de directions des contours proposées dans [ALM 08a], les caractéristiques basées sur les graphèmes présentées dans [ALM 08b] ainsi que les caractéristiques basées sur les K-Segments Adjacents décrites dans [JAI 11]. L'indépendance du script des caractéristiques proposées a été aussi vérifiée sur la base de données arabe IFN/ENIT [PEC 02] dans nos études précédentes [DJE 12a, DJE 12b, DJE 11, DJE 10].

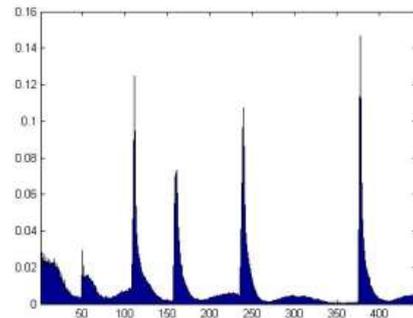
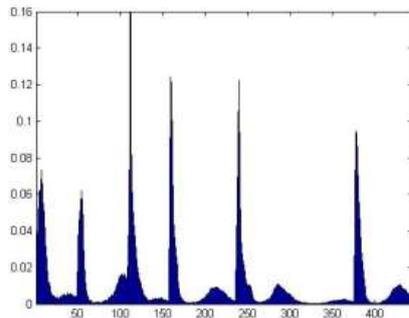
Le pouvoir discriminant des caractéristiques proposées est illustré visuellement par la figure 3.5. La figure 3.5a montre la distribution des longueurs des segments sur deux échantillons issus du même scripteur indiquant clairement une grande similarité intra-scripteur. D'autre part, la figure 3.5b illustre les distributions la distribution des longueurs des segments de textes issus de deux scripteurs différents mettant en évidence la variabilité inter-scripteurs.



Socrates was a Classical Greek philosopher. Credited as one of the founders of western philosophy, he is an enigmatic figure known only through the classical accounts of his students. Plato's dialogues are the most comprehensive accounts of Socrates to survive from antiquity. Forming an accurate picture of the historical Socrates and his philosophical viewpoints is problematic at best. This issue is known as the Socratic problem. The knowledge of the man, his life, and his philosophy is based on writings by his students and contemporaries. Foremost among them is Plato, however, works by Xenophon, Aristotle, and Aristophanes also provide important insights. The difficulty of finding the real Socrates arises because these works are often philosophical or dramatic texts rather than straightforward histories. Aside from Thucydides who makes no mention of Socrates or philosophers in general, there is in fact no such thing as a straightforward history contemporary with Socrates that deals with his own time and place.

Sokrates war ein für das abendländische Denken grundlegender griechischer Philosoph, der in Athen lebte und wirkte. Seine herausragende Bedeutung zeigt sich u.a. darin, dass alle griechischen Denker vor ihm als Vorsokratiker bezeichnet werden. Sokrates entwickelte die philosophische Methode eines strukturierten Dialogs, die er *Maieutik* nannte. Diese Kunst der Gesprächsführung und ihre philosophischen Inhalte sind nur indirekt überliefert worden, da Sokrates selbst nichts Schriftliches hinterlassen hat. Mehrere seiner Schüler, der berühmteste unter ihnen Platon, haben sokratische Dialoge verfasst und unterschiedliche Züge seiner Lehre betont. Die unbeugsame Haltung des Sokrates in dem gegen ihn wegen angeblich verderblichen Einflusses auf die Jugend und wegen Missachtung der Griechischen Götter geführten Prozess hat zu seinem Nachruhm wesentlich beigetragen. Das Todesurteil nahm er als gültiges Fehlurteil gelassen hin, bis zur Hinrichtung durch den Schierlingsteech beschäftigt ihn und die zu Besuch im Gefängnis weilenden Freunde und Schüler philosophische Fragen.

(a)



Socrates was a Classical Greek philosopher. Credited as one of the founders of western philosophy, he is an enigmatic figure known only through the classical accounts of his students. Plato's dialogues are the most comprehensive accounts of Socrates to survive from antiquity. Forming an accurate picture of the historical Socrates and his philosophical viewpoints is problematic at best. This issue is known as the Socratic problem. The knowledge of the man, his life, and his philosophy is based on writings by his students and contemporaries. Foremost among them is Plato, however, works by Xenophon, Aristotle, and Aristophanes also provide important insights. The difficulty of finding the real Socrates arises because these works are often philosophical or dramatic texts rather than straightforward histories. Aside from Thucydides who makes no mention of Socrates or philosophers in general, there is in fact no such thing as a straightforward history contemporary with Socrates that deals with his own time and place.

Sokrates war ein für das abendländische Denken grundlegender griechischer Philosoph, der in Athen lebte und wirkte. Seine herausragende Bedeutung zeigt sich u.a. darin, dass alle griechischen Denker vor ihm als Vorsokratiker bezeichnet werden. Sokrates entwickelte die philosophische Methode eines strukturierten Dialogs, die er *Maieutik* nannte. Diese Kunst der Gesprächsführung und ihre philosophischen Inhalte sind nur indirekt überliefert worden, da Sokrates selbst nichts Schriftliches hinterlassen hat. Mehrere seiner Schüler, der berühmteste unter ihnen Platon, haben sokratische Dialoge verfasst und unterschiedliche Züge seiner Lehre betont. Die unbeugsame Haltung des Sokrates in dem gegen ihn wegen angeblich verderblichen Einflusses auf die Jugend und wegen Missachtung der Griechischen Götter geführten Prozess hat zu seinem Nachruhm wesentlich beigetragen. Das Todesurteil nahm er als gültiges Fehlurteil gelassen hin, bis zur Hinrichtung durch den Schierlingsteech beschäftigt ihn und die zu Besuch im Gefängnis weilenden Freunde und Schüler philosophische Fragen.

(b)

Figure 3.5. Distributions des longueurs des segments de (a) Deux textes manuscrits issus du même scribeur, (b) Deux textes manuscrits de deux scribeurs différents.

3.3.2. Caractéristiques de l'état de l'art

Dans une tentative pour étudier l'efficacité des distributions des longueurs de segments dans un environnement multi-scripts, la performance de ces caractéristiques est comparée avec certaines des méthodes les plus récentes de l'état de l'art qui ont montré de bons résultats en reconnaissance du scripteur monolingue. Celles-ci comprennent les distributions des directions des contours et de charnières des contours proposées par Bulacu et al [BUL 07], une combinaison des caractéristiques du codebook avec des caractéristiques visuelles de l'écriture extraites à partir des chaînes de Freeman et un ensemble de polygones approximant les contours présentée par Siddiqi et al [SID 10] et les coefficients autorégressifs (AR) proposés par Garain et al [GAR 09]. Une description détaillée de ces caractéristiques et des résultats obtenus peut être trouvée dans les documents respectifs relatifs à ces travaux. Le tableau 3.1 résume, pour chacune des caractéristiques utilisées, le numéro correspondant, la description et la dimension.

Caractéristique	Description	Dimension
f_1	Distribution des longueurs de segments blancs dans les quatre directions	200
f_2	Distribution des longueurs de segments noirs dans les quatre directions	400
f_3	Distribution des longueurs de segments noirs et blancs dans les quatre directions	600
f_4	Distribution des directions des contours [BUL 07]	16
f_5	Distribution des charnières des contours avec un fragment de 7 pixels [BUL 07]	2304
f_6	Distribution des approximations polygonales des contours [SID 10]	42
f_7	Distribution des chaînes de Freeman globales [SID 10]	314
f_8	Distribution des chaînes de Freeman locales [SID 10]	230
f_9	Codebook universel [SID 10]	200
f_{10}	Coefficients autorégressifs [GAR 09]	24

Tableau 3.1. Une vue d'ensemble des caractéristiques implémentées et leurs dimensionnalités.

3.4. Indépendance des caractéristiques proposées du script considéré

Avant d'évaluer les caractéristiques proposées dans un environnement multi-script, nous avons mené un certain nombre d'expérimentations afin d'étudier le comportement de la méthode d'extraction de caractéristiques proposée (Run-lengths) quand on change de script. Autrement dit, l'objectif de cette section est de montrer son indépendance vis-à-vis du script considéré. Des bases de données expérimentales contenant des échantillons d'écriture de scripts différents

issues de quelques scripteurs (par exemple 26 scripteurs) jusqu'à mille scripteurs ont été utilisées dans nos expérimentations. Il est intéressant de noter que lors de ces expérimentations, deux protocoles d'évaluations différents ont été considérés.

3.4.1. Expérimentations sur la base IFN-ENIT

Dans cette série d'expérimentations [DJE 12a], nous avons utilisé un seul et unique corpus de 1375 échantillons d'écritures arabes extraits de la base de données IFN-ENIT décrite dans la section 1.6.7 où, à chaque fois, un échantillon est sélectionné pour servir de document de test.

L'évaluation des caractéristiques proposée a été réalisée à l'aide d'un classifieur du type plus proche voisin. Un échantillon d'écriture est choisi parmi les 1375 considérés, puis la distance de Manhattan (Cityblock) est calculée entre l'échantillon choisi et les échantillons restants dans le corpus (1375-1=1374). Ensuite, les distances sont triées de l'échantillon le plus proche au moins proche. Les résultats obtenus sont résumés dans le tableau 3.2.

Caractéristiques	Top 1	Top 5	Top 10
Longueurs de segments blancs	47,20%	76.14%	86,54%
Longueurs de segments noirs	75.42%	90.25%	93.82%
Longueurs de segments noirs et blancs	88.07%	96.87%	98.54%

Tableau 3.2. Résultats de l'identification de scripteurs en utilisant la base IFN-ENIT.

3.4.2. Expérimentations sur la base KHATT

Dans cette série d'expérimentations [DJE 14], nous avons utilisé 3000 échantillons d'écritures arabes issus de 1000 scripteurs différents extraits de la base KHATT décrite dans la section 1.6.6. L'évaluation des caractéristiques proposée a été réalisée à l'aide des séparateurs à vaste marge multi-classes (SVM). Les échantillons d'écritures choisis sont divisés en deux ensembles, un pour l'apprentissage (2000 échantillons) et l'autre pour le test (1000 échantillons). Le tableau 3.3 présente les résultats d'identification enregistrés.

Caractéristiques	Top 1	Top 5	Top 10
Longueurs de segments blancs	25.80%	42.60%	48.10%
Longueurs de segments noirs	58.20%	71.90%	73.20%
Longueurs de segments noirs et blancs	70.60%	82.90%	84.80%

Tableau 3.3. Résultats de l'identification de scripteurs en utilisant la base KHATT.

3.4.3. Expérimentations sur la base CVL

Dans cette série d'expérimentations [KLE 13], nous avons utilisé un seul et unique corpus de 1555 échantillons d'écritures extraits de la base de données CVL décrite dans la section 1.6.5.

L'évaluation des caractéristiques proposée a été réalisée en utilisant le protocole d'évaluation adapté dans la section 3.4.1. Les résultats de cette série d'expérimentation sont montrés dans le tableau 3.4.

Caractéristiques	Top 1	Top 2	Top 5	Top 10
Longueurs de segments noirs et blancs	96.00%	97.00%	97.80%	98.10%

Tableau 3.4. Résultats de l'identification de scripteurs en utilisant la base CVL.

3.4.4. Expérimentations sur les bases des compétitions ICDAR et ICFHR contenant des documents latins et grecs

Dans cette série d'expérimentations [LOU 11, LOU 12, LOU 13], nous avons utilisé les bases de données de compétitions ICDAR 2011, ICFHR 2012 et ICDAR 2013 décrites dans la section 2.5.1.2. L'évaluation des caractéristiques proposées a été réalisée en utilisant le protocole d'évaluation adapté dans les sections 3.4.1 et 3.4.3. Les résultats enregistrés sont présentés dans le tableau 3.5.

Base	Top 1	Top 2	Top 5	Top 10
Base ICDAR 2011	98.60%	100%	100%	100%
Base ICFHR 2012	89.80%	94.30%	97.80%	98.80%
Base ICDAR 2013	90.10%	93.40%	97.00%	97.90%

Tableau 3.5. Résultats de l'identification de scripteurs en utilisant les bases des compétitions ICDAR et ICFHR.

Les résultats expérimentaux obtenus sur les différentes bases de données montrent que les caractéristiques proposées (Run-lengths) permettent d'obtenir des performances intéressantes en mode indépendant du texte en utilisant des textes de différents scripts (arabe, grec et latin). Ces résultats nous ont incités à utiliser les caractéristiques proposées (Run-lengths) pour la caractérisation de scripteurs afin de valider l'hypothèse énoncée dans cette partie de notre travail (au début de ce chapitre) : le style d'écriture d'une personne demeure constant à travers différents scripts.

3.5. Reconnaissance du scripteur

Une fois que les échantillons d'écriture sont représentés par l'ensemble de leurs caractéristiques, nous procédons à l'identification ou à la vérification du scripteur comme indiqué dans les sous sections suivantes.

3.5.1. Identification du scripteur

L'identification du scripteur est effectuée en utilisant deux classifieurs traditionnels: Les K-plus Proches Voisins (k-NN) et les séparateurs à vaste marge multi-classes (SVM). Une brève introduction à ces classifieurs est donnée ci-dessous :

3.5.1.1. k Plus Proche Voisins (k-NN)

Le classifieur de type k Plus Proche Voisins (k-NN pour k Nearest Neighbor) est un algorithme qui classe les objets selon leurs proximités aux données utilisées pour l'entraînement dans l'espace de caractéristiques. C'est une approche de modélisation simple qui approxime la frontière de décision localement. Il n'y a pas d'entraînement proprement dit et les calculs sont seulement effectués lors de la classification. La classification est effectuée par un vote majoritaire du voisinage et l'objet est assigné à la classe la plus présente parmi les k plus proche voisins. Bien que plusieurs distances puissent être utilisées pour définir le voisinage, la mesure que nous avons utilisée est la distance de Manhattan. Le classifieur kNN est une méthode non-paramétrique qui ne nécessite pas d'établir une hypothèse au préalable sur la nature des distributions de données (contrairement à une régression linéaire, par exemple). Le seul paramètre à déterminer est la taille du voisinage (k) et il est défini à partir des données. Une grande valeur de k réduit l'effet du bruit sur les données, mais définit des frontières de décisions sans tenir compte de particularités locales.

3.5.1.2. Séparateurs à vaste marge multi-classes (SVM)

Pour tester la méthode SVM, nous avons utilisé la librairie « *SVM and Kernel Methods Matlab Toolbox* », librement disponible sur [CAN 05]. Les SVM représentent une classe d'algorithmes appartenant, avec la méthode des k plus proches voisins, à la catégorie des méthodes d'apprentissage supervisé.

La méthode SVM est une méthode de classification linéaire qui repose sur l'hypothèse que, étant donné un espace approprié, il existe un classificateur linéaire (appelé hyperplan) permettant de distinguer les deux classes de l'espace (+/-). Le but de cette méthode est d'apprendre, à partir d'un ensemble d'exemples d'apprentissage (apprentissage supervisé), une fonction qui prédit les classes pour de nouveaux objets. Plus concrètement, il s'agit de trouver l'hyperplan optimal, qui sépare les données et maximise la distance entre les deux classes.

L'hyperplan optimal est celui, parmi tous les hyperplans valides, qui réalise la marge maximale entre les points des deux classes. C'est la raison pour laquelle on parle de séparateur à vaste marge. Les points les plus proches de la frontière entre les deux classes et qui sont utilisés pour la détermination de l'hyperplan optimal sont appelés vecteurs supports. L'hyperplan optimal est celui qui permettra au mieux de classer les nouveaux exemples. La classification d'un nouvel exemple inconnu est donnée par sa position par rapport à l'hyperplan optimal. Face à un cas non linéairement séparable (c'est-à-dire la plupart des problèmes réels), les méthodes SVM recourent à une fonction noyau pour effectuer une transformation non linéaire des données. Le résultat de cette transformation, appelé espace de re-description, est un espace de dimension plus grande.

3.5.2. Vérification du scripteur

Pour la vérification de scripteur, on calcule la distance de Manhattan entre deux échantillons donnés et on considère qu'ils sont écrits par la même personne si la distance est inférieure à un seuil prédéfini. Au-delà de la valeur de ce seuil, on considère que les échantillons sont écrits par des scripteurs différents. En faisant varier le seuil d'acceptation, des courbes ROC sont calculées et la performance de la vérification est quantifiée par le «Equal Error Rate» (EER), le point de la courbe où le taux de fausses acceptations (FAR) est égal au taux de faux rejets (FRR) (voir section 1.3.2).

3.6. Résultats expérimentaux et discussion

Cette section présente les expérimentations réalisées pour étudier l'efficacité des caractéristiques proposées pour la reconnaissance du scripteur et de valider l'hypothèse que le style d'écriture d'un individu reste plus ou moins le même à travers les différents scripts. Les mesures de performances utilisées sont le taux d'identification en Top-1 et le taux d'erreur égale (EER) pour tâche de vérification. La section 3.6.1 présente les évaluations dépendantes de la langue, la section 3.6.2 présente les résultats de la reconnaissance du scripteur dans un environnement multi-scripts tandis que la dernière section présente une étude sur la stabilité des caractéristiques.

3.6.1. Expérimentations dépendantes de la langue

Dans cette section, nous présentons et analysons les performances des caractéristiques proposées dans l'identification et la vérification du scripteur en utilisant des échantillons de textes de chaque langue (anglaise ou grecque) indépendamment de l'autre. Pour chacune des deux langues considérées dans notre étude, une image de chacun des 126 scripteurs est utilisée pour l'apprentissage tandis que l'autre est utilisée dans le test. Nous avons envisagé deux scénarios d'évaluation différents. Dans le premier, nous choisissons la première image de chaque scripteur pour l'apprentissage et la deuxième image pour le test. Dans le deuxième scénario, nous choisissons la deuxième image pour l'apprentissage et la première image pour le test.

Le tableau 3.6 présente les taux d'identification des deux classifieurs ainsi que les taux d'erreurs égales (EER) sur les échantillons grecs à l'aide des caractéristiques de longueurs de segments proposés ainsi que les caractéristiques de l'état de l'art présentés dans la section 3.3.2 (voir tableau 3.1). Le tableau 3.7 présente les mêmes résultats pour des échantillons d'écriture en anglais.

Caractéristique	SVM		KNN		EER
	Scénario 1	Scénario 2	Scénario 1	Scénario 2	
$f1$	58.73%	61.11%	61.11%	69.84%	8.38%
$f2$	81.75%	79.36%	84.92%	82.54%	3.63%
$f3$	92.06%	88.09%	92.06%	91.27%	2.78%
$f4$	48.41%	47.62%	67.46%	61.11%	6.76%
$f5$	73.81%	82.54%	82.54%	78.57%	5.15%
$f6$	57.94%	55.56%	67.46%	65.08%	5.16%
$f7$	64.29%	69.84%	62.70%	60.32%	6.73%
$f8$	47.62%	46.82%	58.73%	63.49%	9.13%
$f9$	65.08%	67.46%	63.49%	73.02%	9.22%
$f10$	61.11%	64.29%	68.25%	65.08%	3.97%

Tableau 3.6. Résultats de l'identification et la vérification du scripteur en utilisant les textes grecs.

Caractéristique	SVM		KNN		ERR
	Scenario 1	Scenario 2	Scenario 1	Scenario 2	
<i>f1</i>	49.21%	49.21%	42.86%	46.03%	11.9%
<i>f2</i>	77.78%	80.95%	76.98%	73.81%	5.23%
<i>f3</i>	83.33%	87.30%	80.95%	82.54%	3.57%
<i>f4</i>	46.03%	50.00%	62.70%	54.76%	7.54%
<i>f5</i>	79.36%	80.16%	73.02%	80.16%	4.76%
<i>f6</i>	59.52%	53.17%	69.05%	65.08%	5.62%
<i>f7</i>	66.67%	69.84%	59.52%	60.32%	7.14%
<i>f8</i>	44.44%	50.00%	54.76%	53.17%	7.54%
<i>f9</i>	71.43%	78.57%	70.63%	76.19%	9.07%
<i>f10</i>	63.49%	54.76%	68.25%	68.25%	4.42%

Tableau 3.7. Résultats de l'identification et la vérification du scripteur en utilisant les textes anglais.

Une analyse des résultats présentés dans les tableaux 3.6 et 3.7 conduit aux observations suivantes :

1. La combinaison des distributions des longueurs de segments noirs et blancs (*f3*) donne des performances meilleures que celles enregistrées par toutes les autres caractéristiques. Elle a donné les meilleures performances dans toutes les expérimentations de l'identification du scripteur et la même tendance peut être observée pour la tâche de vérification du scripteur aussi.
2. Les méthodes de l'état de l'art qui ont enregistrés des taux d'identification très élevés sur de grandes bases de données telles que IAM [MAR 02] et RIMES [GRO 08], ont montrées une baisse significative des performances dans nos expérimentations. Cela est certainement dû au fait que les images utilisées lors de nos expérimentations contiennent de petites quantités de texte. Par conséquent, ces méthodes ne sont pas très efficaces quand une quantité limitée de texte est disponible.
3. La combinaison des distributions des longueurs de segments noirs et blancs (*f3*) permet de réaliser d'assez bons résultats sur de petites quantités de texte et elle est donc moins sensible à la quantité de texte, contrairement aux autres caractéristiques traditionnelles.
4. Il convient de noter également que les longueurs de segments noirs (*f2*) sont plus informatives que les longueurs de segments blancs (*f1*).
5. Les longueurs de segments blancs (*f1*) seuls ne sont pas très discriminatifs mais contribuent à améliorer les performances lorsqu'ils sont combinés avec les longueurs de segments noirs (*f2*).
6. Les résultats de l'identification et de la vérification enregistrés sur des textes grecs sont meilleurs que ceux réalisé sur des textes anglais pour toutes les caractéristiques à l'exception des caractéristiques du codebook universel (*f9*). Cela peut être justifié par le fait que ces échantillons étaient écrits dans la langue maternelle des scripteurs. Il est possible que les particularités et les caractéristiques des écritures des différents

scripteurs impliqués dans nos expérimentations se reflètent mieux en écrivant dans leur langue maternelle.

- Il peut également être observé dans les tableaux 3.6 et 3.7 que les caractéristiques du codebook universel (f_9) sont dépendantes du script. Étant donné que le codebook a été généré sur un ensemble de documents en anglais, les taux d'identification diminuent de 71,43% et 78,57% (pour les scénarios 1 et 2, respectivement) sur des textes anglais à 65,08% et 73,02% pour les deux scénarios en utilisant des documents grecs.

3.6.2. Expérimentations Multi-scripts (Grec vs Anglais et inversement)

Cette section présente les évaluations menées pour valider l'hypothèse disant que le style d'un scripteur reste constant à travers différents scripts. Nous avons effectué une série d'expérimentations en utilisant des échantillons de textes dans la base d'apprentissage qui sont écrits en un script différent de ceux de la base de test. Deux scénarios différents sont considérés dans notre étude: dans le premier, les classifieurs sont entraînés en utilisant des échantillons grecs et les tests sont effectuées en utilisant des échantillons en anglais alors que le contraire est considéré dans le deuxième scénario. Les résultats de l'identification et de la vérification dans ces expérimentations sont présentés dans le tableau 3.8.

Caractéristique	SVM		KNN		ERR
	Scenario 1	Scenario 2	Scenario 1	Scenario 2	
f_1	38.89%	42.46%	41.27%	37.70%	14.55%
f_2	60.32%	61.11%	60.32%	62.30%	7.80%
f_3	73.02%	76.59%	73.41%	76.19%	5.75%
f_4	32.54%	42.06%	38.49%	38.49%	12.57%
f_5	62.30%	57.94%	61.51%	61.11%	8.07%
f_6	30.56%	29.76%	39.29%	38.09%	11.77%
f_7	49.21%	50.00%	32.94%	32.54%	14.28%
f_8	30.56%	34.52%	33.33%	30.95%	14.75%
f_9	55.16%	51.98%	50.40%	44.05%	14.35%
f_{10}	40.08%	42.86%	39.68%	39.68%	9.99%

Tableau 3.8. Les résultats de l'identification et de la vérification du scripteur en utilisant les documents grec en apprentissage et les documents anglais en test et inversement.

Une analyse des résultats présentés dans le tableau 3.8 conduit aux observations suivantes:

- On peut remarquer que les performances des caractéristiques évaluées à la fois en identification et en vérification est moins impressionnante par rapport à celles obtenues lors des expérimentations dépendantes de la langue. La différence entre les performances de ces caractéristiques est également plus importante.
- Comme dans le cas des expérimentations dépendantes de la langue, la combinaison des distributions des longueurs de segments noirs et blancs (f_3) continue de représenter le meilleur ensemble de caractéristiques en réalisant des taux d'identification de l'ordre de 73,02% et 76,59% (pour les scénarios 1 et 2, respectivement avec les SVM multi-classes comme classifieurs) et un taux d'erreur égale (EER) de l'ordre de 5,75%. Avec l'utilisation de classifieur KNN, le même

ensemble de caractéristiques donne des taux d'identification de l'ordre de 73,41% et 76,19% pour les deux scénarios. Considérant que les expérimentations ont été réalisées dans un environnement multi-scripts avec une quantité de texte très limitée, ces résultats sont très prometteurs.

3. Les coefficients autorégressifs (f_{10}), qui ont été évalués dans un environnement multi-scripts partiel [GAR 09] réalisent des taux d'identification de l'ordre de 40,08% et 42,86% et un taux d'erreur égale d'environ 10%. Par conséquent, ils ne semblent pas être un choix approprié dans un vrai environnement multi-scripts.
4. Il est également intéressant de noter que la plupart des caractéristiques de l'état de l'art sont plus dépendantes du script ou de l'alphabet considéré que les caractéristiques des longueurs des segments. Elles modélisent uniquement les particularités du script ou de l'alphabet tandis que les caractéristiques des longueurs des segments (f_3) comprennent également des informations telles que la largeur moyenne des lettres, la densité de l'écriture, la structure de lettres, la taille moyenne des lettres, la position de caractères, les régions incluses à l'intérieur des lettres, les espaces intra et inter-mots et la régularité globale de l'écriture manuscrite. Ces attributs sont liés au style d'écriture d'un individu (qui reste le même à travers différents scripts) plutôt qu'au script considéré. C'est pour cela que ces caractéristiques sont plus performantes que les caractéristiques de l'état de l'art, surtout quand les deux écritures à comparer ne sont pas écrites en utilisant le même script.
5. Enfin, les résultats obtenus pour l'identification et la vérification du scripteur démontrent clairement le potentiel des caractéristiques des longueurs des segments (f_3) pour la reconnaissance du scripteur à partir d'images d'écriture contenant une petite quantité de texte. Ces résultats soutiennent également l'idée proposée dans cette partie de notre travail : «les scripteurs partagent certaines caractéristiques à travers les différents scripts qu'ils produisent et il est possible d'identifier un scripteur à partir d'échantillons écrits dans un script différent de ceux utilisés pour l'apprentissage».

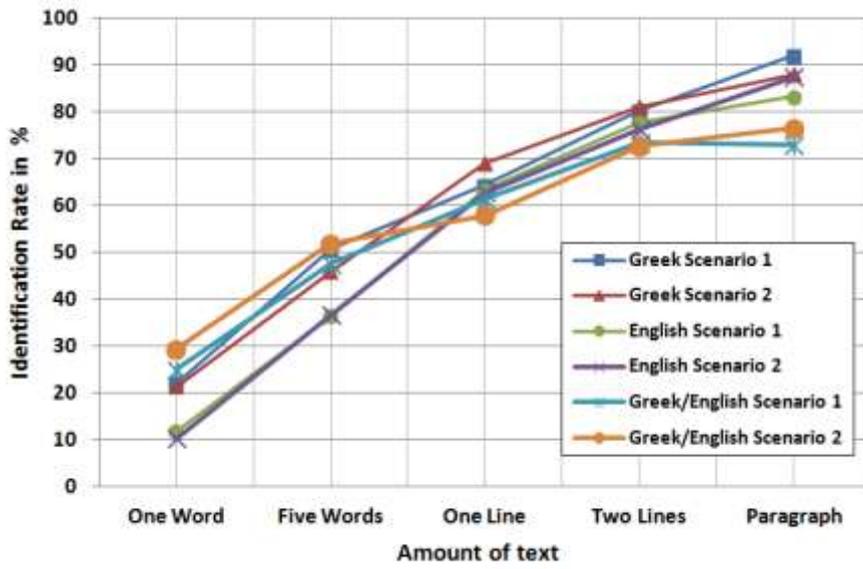
3.6.3. Stabilité des caractéristiques proposées

Dans cette section, nous allons étudier la stabilité des caractéristiques proposées en fonction de la quantité du texte disponible ainsi que de la présence du bruit dans les images considérées dans les différentes séries d'évaluations.

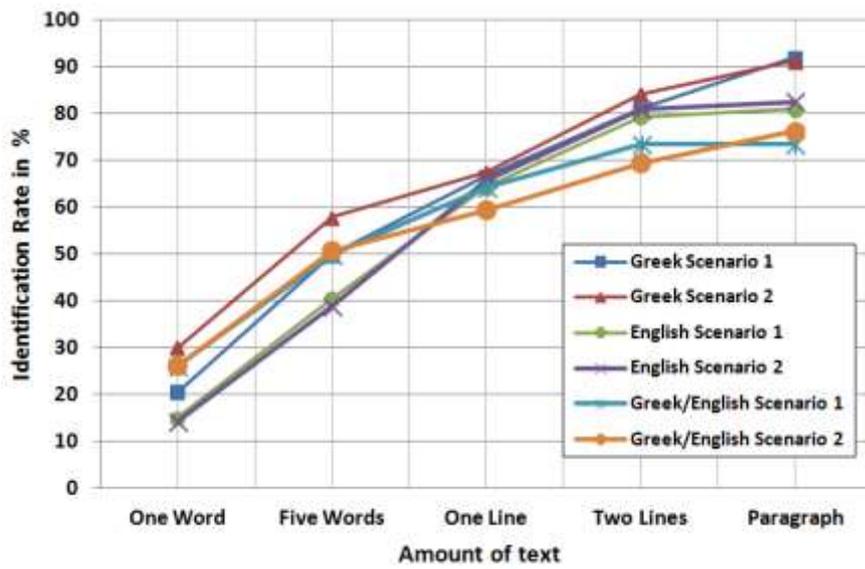
3.6.3.1. Performances en fonction de la quantité de texte

Un des paramètres les plus importants qui influencent les performances des systèmes d'identification et de vérification du scripteur est la quantité disponible du texte dans les échantillons d'écriture pour chaque scripteur.

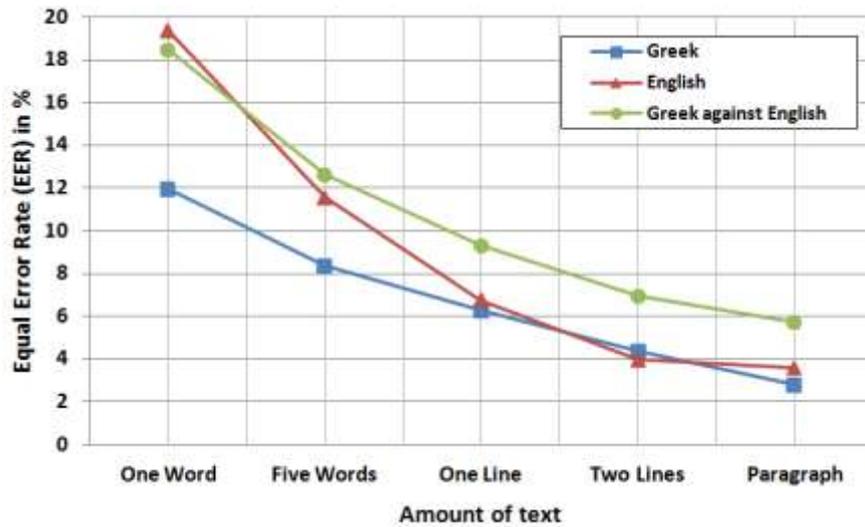
Pour étudier comment les performances des caractéristiques des longueurs des segments proposées varient en fonction de la quantité du texte, nous avons évalué notre système en utilisant un mot, cinq mots, une ligne, deux lignes et un paragraphe complet (3 lignes ou plus) pour chacun des scripteurs. Les résultats de ces expérimentations sont résumés dans la figure 3.6. Naturellement, les performances se dégradent considérablement lorsqu'on utilise de faibles quantités d'écriture. Dans certains cas, les courbes des performances commencent à se stabiliser à partir de deux lignes de texte. Ces observations sont cohérentes avec celles de Siddiqi et al [SID 10] où les auteurs rapportent les résultats «acceptables» avec trois lignes de texte en utilisant la base des données IAM [MAR 02].



(a)



(b)



(c)

Figure 3.6. Performances en utilisant différentes quantités de texte: (a) Taux d'identification avec le classifieur SVM, (b) Taux d'identification avec la classifieur KNN, (c) Taux d'erreur égale.

3.6.3.2. Influence du bruit sur les performances des caractéristiques proposées

La présence du bruit dans les documents étudiés est également un paramètre important qui peut affecter, de manière significative, les performances des caractéristiques des longueurs des segments pour l'identification et la vérification du scripteur. Pour étudier la sensibilité de ces caractéristiques au bruit dans l'image, nous avons ajouté un bruit de type sel et poivre de densités variées aux images de la base de données. La densité du bruit varie de 1% à 5% de l'ensemble des pixels, ce qui a permis d'obtenir cinq ensembles de données pour les expérimentations. Des échantillons extraits à partir de ces ensembles de données bruitées sont illustrés dans la figure 3.7, tandis que les résultats de l'identification et de la vérification sur des images bruitées sont résumés dans la figure 3.8.

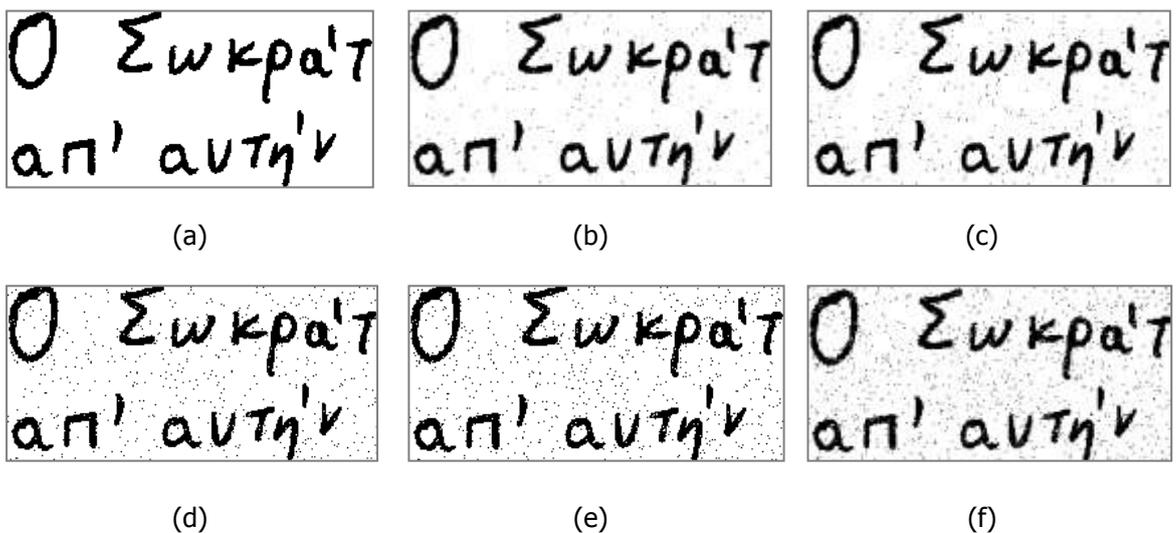
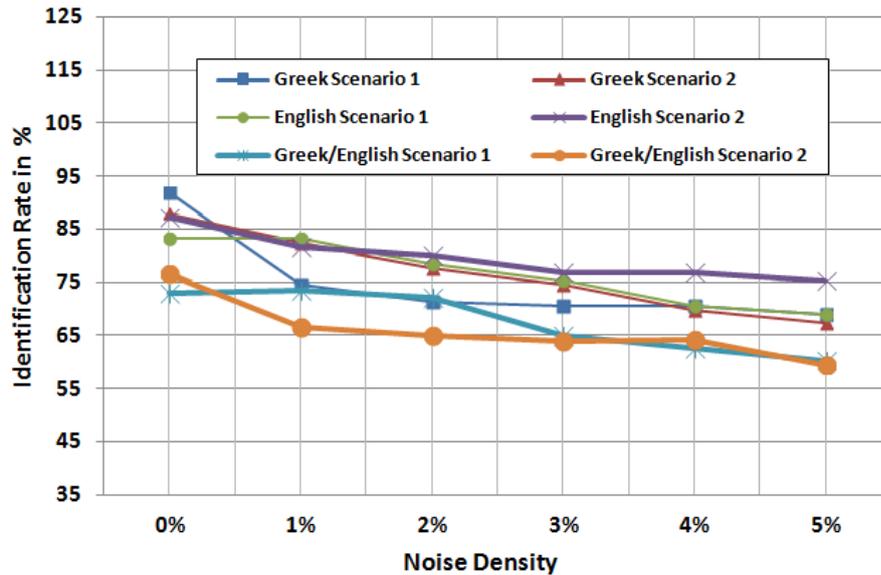
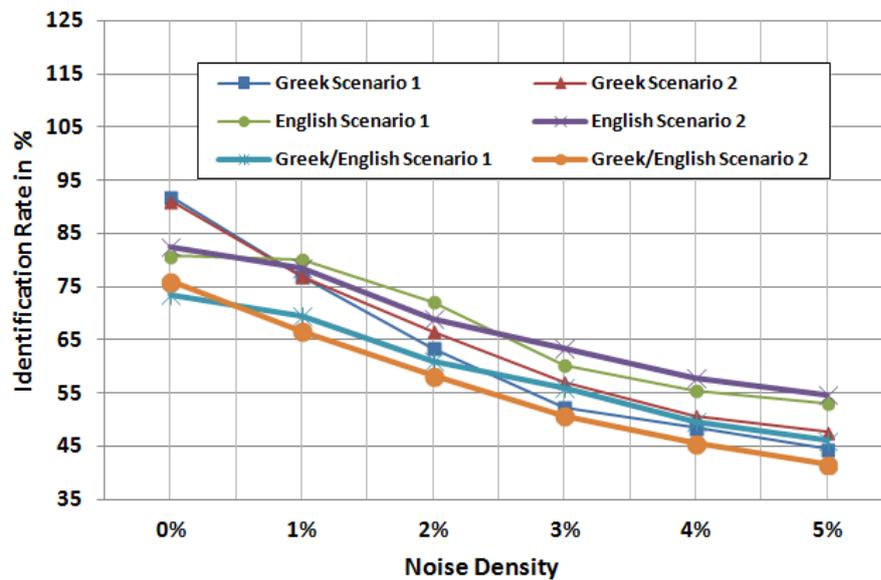


Figure 3.7. (a) Image Originale, (b-f) Images bruitées avec densités de 1%, 2%, 3%, 4% et 5% respectivement.

Naturellement, une dégradation progressive des performances est enregistrée quand la densité de bruit augmente indiquant la sensibilité des caractéristiques des longueurs des segments au bruit. Le système ne parvient pas à reconnaître un nombre considérable de scripteurs en présence d'un bruit qu'on pourrait qualifier de considérable. Il convient toutefois de noter que ces expérimentations visaient essentiellement à étudier les effets du bruit sur ces caractéristiques. Pour des applications pratiques, le bruit peut être éliminé durant une étape de prétraitements avant l'extraction de ces caractéristiques.



(a)



(b)

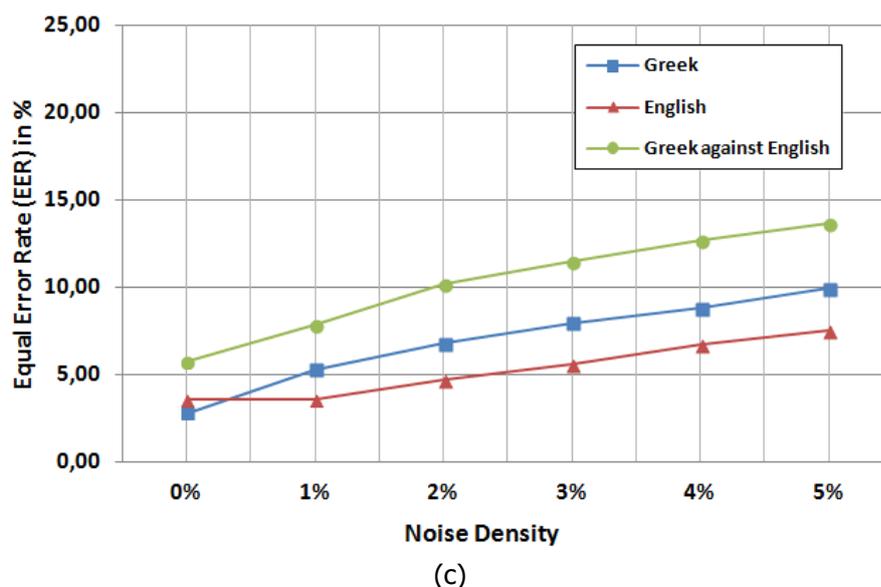


Figure 3.8. Performances du système en fonction de la densité du bruit: (a) Taux d'identification avec le classifieur SVM, (b) Taux d'identification avec le classifieur KNN, (c) Taux d'erreur égale.

3.7. Conclusion

Ce chapitre avait pour objectif de présenter notre méthode indépendante du texte pour l'identification et la vérification du scripteur à partir de documents multi-scripts [DJE 13a]. Nous avons utilisé un ensemble de caractéristiques qui ont montré des résultats prometteurs sur une base de données de documents manuscrits écrits en deux langues différentes : grecque et anglaise. A notre connaissance, cette étude est la première de son genre qui implique des textes multi-scripts dans le vrai sens du terme. Les évaluations ont été effectuées sur la seule base de données existante de ce type contenant des échantillons de textes très courts collectés auprès de 126 scripteurs différents [LOU 12].

Les résultats obtenus pour l'identification et la vérification de scripteurs sont très encourageants. Ils reflètent l'efficacité des caractéristiques de longueur de segments (run-lengths) dans un environnement multi-scripts et valident l'hypothèse avancée dans cette thèse : le style d'écriture d'une personne reste approximativement le même à travers différents scripts. Il est également intéressant de noter que, contrairement à la plupart des études qui utilisent des pages complètes de texte, nos résultats sont basés sur une quantité limitée de texte manuscrit, ce qui est, en fait, plus proche des scénarios du monde réel.

Un autre aspect intéressant de cette étude est l'évaluation et la comparaison d'un certain nombre de méthodes de l'état de l'art sur cette base de données. Les caractéristiques utilisées dans ces méthodes présentent naturellement une diminution des performances lorsqu'elles sont exposées à un environnement multi-scripts. Dans tous les cas, les caractéristiques de longueurs de segments se sont montrées plus performantes que les autres caractéristiques. Il serait intéressant d'évaluer toutes ces caractéristiques sur des bases de données relativement importantes contenant un grand nombre de scripteurs et de nombreux scripts par scripteur. C'est cependant une tâche difficile de trouver des personnes qui sont familières avec plusieurs scripts et, par conséquent, de construire des bases de données réelles pour ce genre d'évaluation.

ANALYSE AUTOMATIQUE DE L'ÉCRITURE MANUSCRITE POUR LA DÉTERMINATION DU SEXE D'UN INDIVIDU

Ce chapitre présente l'une de nos contributions principales qui consiste en une étude pour la détermination du sexe des individus à partir d'images scannées de leurs traces écrites [SID 14]. La méthode proposée est basée sur l'extraction d'un ensemble de caractéristiques de l'écriture à partir d'échantillons de scripteurs de sexes masculin et féminin et l'entraînement d'un classifieur afin qu'il puisse distinguer entre les deux. Des attributs de l'écriture comme l'orientation, la courbure, la texture et la lisibilité sont estimés en calculant des caractéristiques locales et globales. La classification est effectuée à l'aide des réseaux de neurones artificiels (ANN) ainsi que des séparateurs à vaste marge (SVM). La méthode proposée a été évaluée en utilisant deux bases de données sous un certain nombre de scénarios où des résultats intéressants ont été enregistrés pour la détermination du sexe à partir l'écriture manuscrite.

4.1. Introduction

L'écriture est l'un des plus anciens modes de communication dans notre civilisation qui s'est développée et évoluée au fil du temps. Un individu apprend à écrire en copiant des formes à partir d'un cahier d'écriture standard qui, lui aussi, diffère selon la localisation géographique, les circonstances temporelles, sociales et culturelles. Le style appris à partir du cahier d'écriture, cependant, s'atténue avec le temps et la personne développe ses préférences personnelles d'écriture. Dans ce contexte, contrairement au texte électronique ou imprimé, un texte manuscrit apporte des informations supplémentaires sur l'individu qui a produit le texte. Ceci rend l'analyse de l'écriture un domaine de recherche attractif pour les psychologues, les examinateurs de documents, les paléographes, les graphologues et les analystes judiciaires. 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 du sexe à partir de l'écriture manuscrite a été un thème de recherche intéressant depuis les premières décennies du siècle dernier [BRO 29, LOP 09, TEN 34, NEW 26] et a beaucoup évoluée depuis ce temps. Avec les progrès dans les techniques de l'analyse d'images et de la reconnaissance de formes, l'analyse manuelle de l'écriture manuscrite est remplacée par des systèmes automatisés.

Un grand nombre de systèmes ont été proposés et développés pour l'analyse automatique de documents manuscrits ciblant principalement des applications telles que la reconnaissance de l'écriture, le repérage de mots (word spotting en anglais), l'identification du scripteur et la vérification de signatures manuscrites. 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 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 [SUN 01, KAR 05, HAM 96, LIW 06, LIW 11, SOK 12].

Dans ce chapitre, nous présentons notre proposition d'un système de classification automatique de documents manuscrits par le sexe de leur scripteur en utilisant un ensemble de caractéristiques visant à déterminer un sous-ensemble d'attributs discriminants définis par les psychologues. Chaque échantillon d'écriture est représenté par un ensemble de caractéristiques qui sont utilisées par un classifieur afin qu'il puisse d'apprendre à distinguer entre les deux catégories: masculine et féminine. La classification est effectuée à l'aide des réseaux de neurones ainsi que des séparateurs à vaste marge. La méthode proposée, évaluée sur deux ensembles de données donne des résultats intéressants. En plus de l'objectif primaire qui consiste à déterminer le sexe d'un scripteur à partir de son écriture, nous avons également étudié l'influence du contenu textuel, du script et de la localisation géographique des scripteurs sur les performances de classification.

Ce chapitre est organisé comme suit. Dans la section suivante, nous discutons certains attributs de l'écriture qui servent à distinguer les écrits des scripteurs masculins et féminins. Nous présentons ensuite les caractéristiques proposées dans la section 4.3, suivies par la description des techniques de classification utilisées dans la section 4.4. Les résultats expérimentaux et leur analyse sont présentés dans la section 4.5, tandis que la dernière section conclut ce chapitre.

4.2. Différences de sexe dans l'écriture manuscrite

Comme indiqué précédemment, plusieurs études ont montré que le genre d'un scripteur (son sexe) 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 genre d'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.

Typiquement, les psychologues suggèrent qu'une écriture féminine soit caractérisée par le fait d'être soignée, bien organisé, arrondie, petite et symétrique. D'autre part, les écrits pressés, irréguliers, désordonnés, sinueux et inclinés sont plutôt susceptibles d'appartenir à un scripteur masculin [JAM 91, BUR 02].

Quelques exemples d'échantillons d'écriture de scripteurs masculins et féminins soutiennent ces conclusions et sont illustrés sur la figure 4.1. Les examinateurs de documents ont recensé un ensemble de 21 caractéristiques distinctives (qualitatives et quantitatives) qui peuvent être

efficacement utilisées pour l'analyse de l'écriture [ROY 99]. Ces caractéristiques sont généralement qualifiées comme des caractéristiques classiques [SRI 02]. Certaines de ces caractéristiques qui peuvent être calculées algorithmiquement à partir d'images scannées de l'écriture sont connues comme des caractéristiques computationnelles. Elles comprennent, par exemple, l'inclinaison, les espaces intra et inter-mots, la pression de la plume et des informations sur le gradient, etc. Certaines de ces caractéristiques ont été également appliquées avec succès à la détermination du sexe d'un scripteur à partir des caractéristiques de son écriture [KAR 05].

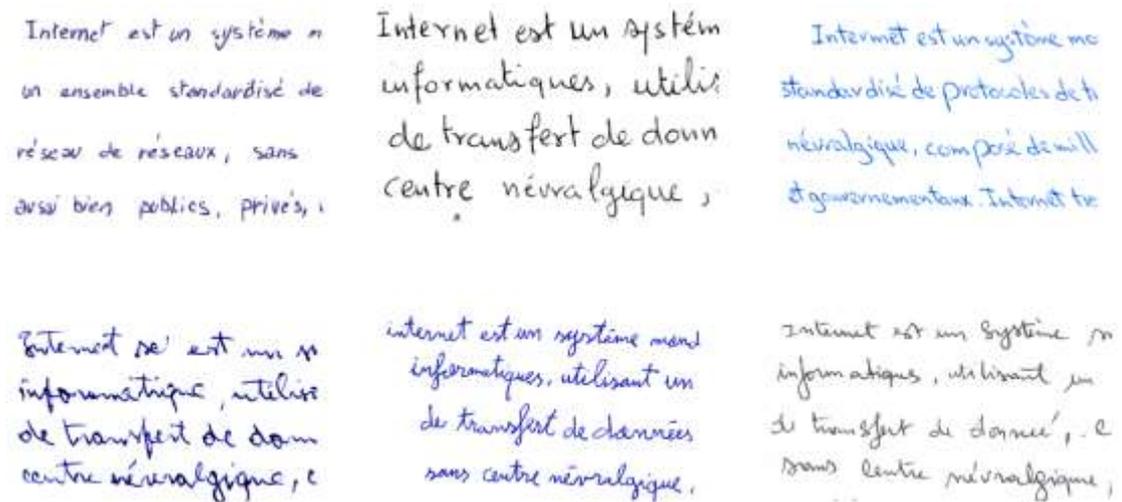


Figure 4.1. Échantillons d'écriture de scripteurs féminins (en haut) et masculins (en bas)

La tâche de détermination du sexe d'un scripteur à partir de documents manuscrits est étroitement liée à celle de l'identification du scripteur. La première différence réside dans le fait que l'identification du scripteur est problème à N classes tandis que la classification en genre est un problème à deux classes. Les caractéristiques qui ont été appliquées efficacement à l'identification du scripteur sont donc susceptibles de donner de bonnes performances pour la détermination du sexe d'un scripteur aussi [LIW 11]. Les bases de données utilisées ainsi que les caractéristiques proposées seront présentées et discutées en détails dans les sections 4.3 et 4.4 respectivement.

4.3. Bases de données

Dans notre étude, nous avons utilisé des échantillons d'écriture de deux bases de données différentes, la base QUWI [ALM 12] qui a été développée à l'université de Qatar ainsi que la base de données MSHD qui a été créée au niveau du Laboratoire de Mathématiques, d'Informatique et des Systèmes (LAMIS) de l'université de Tébessa en algérie.

Dans la base de données QUWI, chaque scripteur a contribué à la production de 4 pages, deux en arabe et deux en anglais. La première et la troisième page de chaque scripteur contiennent un texte arbitraire choisi par le scripteur en arabe et en anglais respectivement, tandis que la deuxième et la quatrième page contiennent les deux mêmes textes (arabe et anglais) recopiés par tous les scripteurs. Ceci permet d'utiliser la base de données en mode dépendant du texte ainsi qu'en mode indépendant du texte. La performance du système sur différents scripts

(arabe et anglais) peut également être étudiée. Nous avons utilisé des échantillons d'écriture de 475 scripteurs différents (1900 échantillons) dans notre étude.

Dans nos expérimentations, nous avons divisé la base de données QUWI en 3 parties. Des échantillons provenant de 300 scripteurs sont utilisés comme ensemble d'apprentissage (d'entraînement), ceux de 75 scripteurs constituent l'ensemble de validation tandis que ceux de 100 scripteurs sont utilisés comme ensemble de test. La répartition des scripteurs sur les ensembles d'apprentissage (d'entraînement), de validation et de test est résumée dans le tableau 4.1. La répartition des scripteurs reste le même dans les différentes expérimentations. La répartition des échantillons varie naturellement d'une expérimentation à une autre (voir la section 4.6).

Ensemble de données	Apprentissage	Validation	Test
Scripteurs	300	75	100
Echantillons	1200	300	400

Tableau 4.1. Répartition de la base de données QUWI.

La base de données MSHD comprend des échantillons d'écritures arabes et françaises de 87 scripteurs différents. Chaque scripteur a contribué à la production de 12 pages, 6 en français et 6 en arabe, soit un total de 1044 échantillons d'écriture. Le texte de l'échantillon i est le même ($i = 1, 2, \dots, 12$) pour l'ensemble des scripteurs. Les informations sur le sexe de trois scripteurs ne sont pas disponibles, c'est pour cela que les écritures de 84 scripteurs sont prises en compte dans nos évaluations. Pour la plupart des évaluations, les échantillons d'écriture de 42 scripteurs constituent l'ensemble d'apprentissage alors que ceux des 42 scripteurs restants composent l'ensemble de test. Nous avons utilisé la même base de données pour évaluer un système d'identification du scripteur [DJE 13]. La figure 4.2 illustre des échantillons d'écritures extraits des bases de données QUWI et MSHD.

L'internet est un système mondial d'interconnexion de réseaux informatiques, utilisant un ensemble standardisé de protocoles de transfert de données. C'est donc un réseau de réseaux, sans centre hiérarchique, composé de milliers de réseaux aussi bien publics, privés, universitaires, commerciaux et gouvernementaux. Internet transporte un large spectre d'information et permet l'élaboration d'applications et de services variés comme le courrier électronique, la messagerie instantanée et le World Wide Web. Internet ayant été popularisé par l'apparition du World Wide Web, les deux sont parfois confondus par le public non averti. Le World Wide Web n'est pourtant que l'un des applications d'internet.

الانترنت هو نظام عالمي من الترابط بين الشبكات الحاسوبية، باستخدام مجموعة قياسية من البروتوكولات لنقل البيانات. إنه بذلك شبكة من الشبكات، بدون مركز هرمي، تتكون من آلاف الشبكات سواء كانت عامة، خاصة، أكاديمية، تجارية أو حكومية. إنترنت تنقل مجموعة واسعة من المعلومات وتتيح تطوير تطبيقات وخدمات متنوعة مثل البريد الإلكتروني، المراسلة الفورية، وWorld Wide Web. إنترنت أصبحت شائعة بفضل ظهور World Wide Web، وغالباً ما يتم الخلط بين الإثنين من قبل الجمهور غير المتخصص. إنترنت هي في الحقيقة واحدة من تطبيقات إنترنت.

(a)

we could experiment on human for good reason. In compliance with ethical restrictions, cancer experiments usually adopt two approaches, both proved invaluable in acquiring basic knowledge but defective when applied to drug discovery. The first approach uses human cell lines cultured in petri dishes. Scientists expose cells with known characteristics to various circumstances or potential drugs, observe

وإنه تماماً كذلك عندما نتبع مع المرضى الذين يعانون من الأمراض والمخاطر التي يرافقها، وإرضاع درجة حرارة الأذن، وما شابه من ذلك. من دوائهم في العيادات، وإرضاع مسوب مياه الدمار ومن الكثير من السوائل وتطلع مياه الشرب، مصفاة من قمامة تأكل الحبة الأوروبية التي كانت تسمح لمجموعة فوق النسبية الهائلة الدنوي إلى حواله الكورة الأمريكية، وأثره على ما استقرت الكثير من أنواع الحياة الدقيقة في الأرض.

(b)

Figure 4.2. Échantillons extraits (a) de la base MSHD, (b) de la base QUWI.

4.4. Extraction de caractéristiques

Parmi les différents attributs discriminants des écritures masculines et féminines, nous nous concentrons, dans notre étude, sur l'inclinaison et l'orientation, la rondeur et la courbure, la propreté et la lisibilité ainsi que la texture de l'écriture. Afin de calculer algorithmiquement ces attributs à partir d'images numérisées d'écriture, nous utilisons un ensemble de caractéristiques calculées à différents niveaux d'observation. L'orientation et la courbure sont estimées à partir des contours de l'écriture en représentant les contours à l'aide: 1) chaînes de Freeman et 2) un ensemble de polygones approximant les contours.

La netteté ou la lisibilité de l'écriture, bien qu'étant un attribut très subjectif, est estimée en calculant la dimension fractale de l'écriture. Dans des travaux antérieurs [SID 09a, SID 09b, SID 09c, SID 10], les caractéristiques basées sur l'orientation et la courbure ont montré des résultats très encourageants pour les tâches d'identification et de vérification de scripteurs tandis que les dimensions fractales ont été employées avec succès pour la classification des écritures dans [BOU 98]. Les caractéristiques basées sur la texture sont également connues pour caractériser le scripteur d'un document [SAI 00]. Dans cette étude, nous sommes intéressés à étudier l'efficacité de ces caractéristiques pour la tâche de détermination du sexe à partir de l'écriture manuscrite.

4.4.1. Orientation et courbure

L'information de l'orientation et de la courbure dans une écriture est capturée par un ensemble de caractéristiques calculées à partir des contours de l'écriture. La représentation du contour est choisie en se basant sur l'hypothèse que la forme des caractères dans un document manuscrit peut être encapsulée par ses contours. L'utilisation des contours élimine également la sensibilité à l'instrument d'écriture tout en conservant la forme des caractères.

Nous extrayons les contours intérieurs et extérieurs de l'écriture et, comme nous l'avons discuté précédemment, ces contours sont représentés par une séquence des codes de chaîne de Freeman et par un ensemble de polygones obtenus en appliquant un algorithme de polygonisation aux contours. Ces deux représentations correspondent à deux différentes échelles d'observations et les caractéristiques calculées à partir de ces différentes représentations se complètent mutuellement. Nous discutons l'extraction de caractéristiques de l'orientation et de la courbure de chacune de ces représentations dans les sous-sections suivantes.

4.4.1.1. Caractéristiques basées sur les chaînes de codes de Freeman

Les chaînes de codes de Freeman ont été appliquées efficacement à des problèmes tels que la reconnaissance de mots et de caractères [HIR 96, KIM 97, BLU 03, BLU 07], la classification des styles d'écriture [DEH 03] et l'identification du scripteur [SID 09a, SID 10]. Étant donné que notre tâche de détermination du sexe concerne également des documents manuscrits, nous pensons qu'une représentation basée sur les codes de la chaîne Freeman peut être efficace pour l'extraction de caractéristiques. Nous représentons les contours de l'écriture par des chaînes de codes de Freeman. Chaque contour est une séquence de pixels de frontière avec $\{c_j | 1 < j \leq M_{i-1}\}$ où $c_j \in \{0,1, \dots, 7\}$ et M_i est la longueur du contour i . Un exemple de caractère avec son contour et les codes associés à chacune des directions est illustré à la figure 4.3.

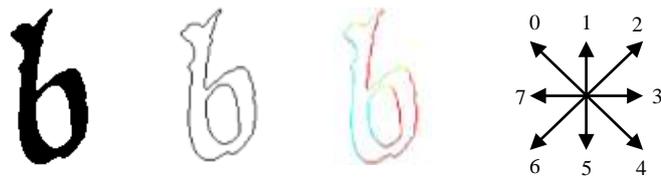


Figure 4.3. Une image d'un caractère avec ses contours et la représentation en chaîne de codes de Freeman.

Dès que l'image de l'écriture est représentée par une chaîne de codes, nous calculons l'histogramme (normalisé) de la chaîne de codes, généralement appelé distribution de codes de Freeman ($f1$). Les (huit) cases de l'histogramme représentent la contribution relative de chacune des huit directions principales dans une écriture tandis que les orientations dominantes dans l'écriture sont représentées par des pics dans l'histogramme.

Cependant, il est important de noter que, puisque les images sont hors-ligne, nous ne pouvons pas discriminer entre les traits avant et les traits arrière, ainsi, la séquence de code affectée à un contour particulier dépend de la manière de le traverser. Une solution serait de diviser l'histogramme en quatre cases représentant les quatre directions principales des traits : horizontale, verticale, diagonale gauche et diagonale droite. Cependant, notre expérience a montré que le meilleur choix serait de conserver les 8 cases et d'être consistant dans la manière de tracer un contour.

La figure 4.4 illustre la distribution des codes de Freeman calculée à partir de deux échantillons d'écriture. Il peut être constaté que l'orientation verticale globale dans l'échantillon 'a' se traduit par deux pics aux cases respectives de l'histogramme correspondant. De même, pour l'échantillon 'b' où l'écriture est inclinée vers la droite, les pics peuvent être observés au niveau des cases correspondant aux directions diagonales droites.

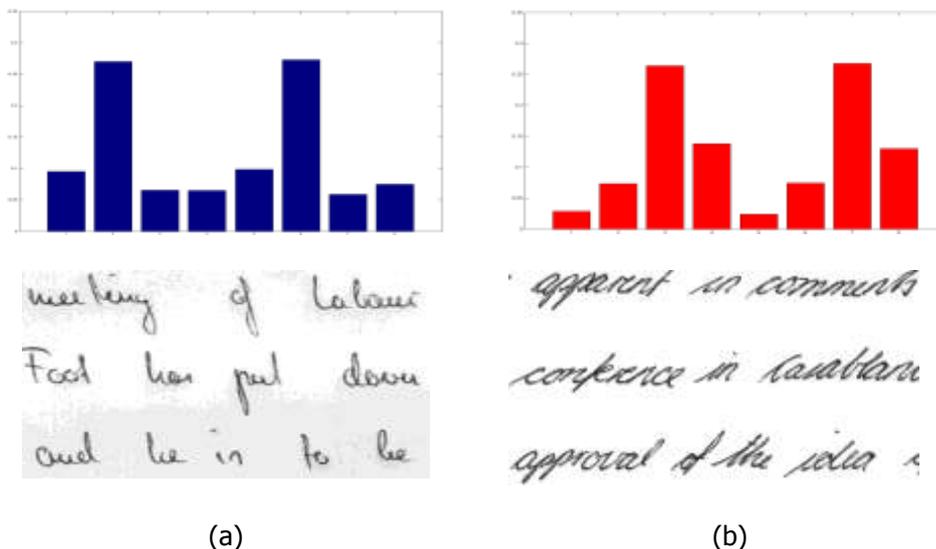


Figure 4.4. Des échantillons d'écriture et leurs distributions de codes de Freeman respectives.

Pour estimer la courbure au niveau du pixel, nous calculons l'histogramme de paires de codes de Freeman. Nous initialisons une matrice de 8×8 avec la mise à zéro de tous les éléments. Pour chaque paire (i, j) dans la représentation en chaîne de codes de Freeman d'une écriture, nous incrémentons l'élément correspondant de la matrice (histogramme). La distribution est finalement normalisée pour être indépendante de la quantité du texte. Cette distribution ($f2$)

pourrait être considérée comme une mesure de l'angle (courbure) entre les vecteurs représentant les directions de la chaîne de codes comme illustré à la figure 4.5.

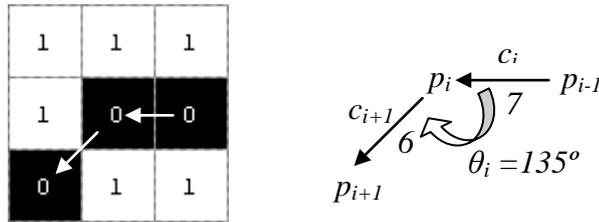


Figure 4.5. La paire de chaîne de codes de Freeman (7, 6) représentant un angle de 135° au pixel de position p_i .

Les attributs de l'orientation et de la courbure, dans une écriture, calculés à partir des chaînes de codes de Freeman sont efficaces mais, étant donné que ces attributs sont calculés au niveau du pixel, ils pourraient être sensibles au bruit dans l'écriture. Pour compléter ces caractéristiques, nous calculons des caractéristiques similaires en estimant d'abord les contours par un ensemble de polygones. Ceci, en fait, correspond à une échelle d'observation lointaine et les caractéristiques calculées sont également plus robustes au bruit. Ces caractéristiques sont discutées dans la sous section suivante.

4.4.1.2. Caractéristiques basées sur les polygones

En utilisant l'algorithme séquentiel de polygonisation [KAR 84], nous estimons les contours par un ensemble de segments de ligne (polygones). Un exemple de contours polygonisés est illustré dans la Figure 4.6.



(a)



(b)

Figure 4.6. Polygonisation a) image originale b) Contours polygonisés.

Pour chaque segment dans l'écriture, nous calculons sa pente et nous utilisons la distribution de ces pentes comme notre prochaine caractéristique (f_3). L'intervalle -90° à 90° est quantifié en 8 cases et les pentes des lignes approxinant les contours d'écriture sont comptées dans leurs cases respectives. L'histogramme est finalement normalisé et est utilisé comme caractéristique. Pour estimer la courbure, nous calculons l'angle entre chaque paire de segments connectés comme suit:

$$\alpha_i = \pi - \cos^{-1} \frac{V_i \cdot V_{i+1}}{|V_i||V_{i+1}|}$$

Avec V_i et V_{i+1} sont les vecteurs de (x_{i-1}, y_{i-1}) à (x_i, y_i) et de (x_i, y_i) à (x_{i+1}, y_{i+1}) respectivement comme illustré à la figure 4.7.

De manière similaire à la distribution des pentes, les angles (0° à 180°) sont quantifiés en 8 cases et leur distribution (f4) est utilisée pour caractériser l'écriture. Les détails d'implémentation de ces caractéristiques peuvent être trouvés dans [SID 10].

Après avoir présenté les caractéristiques de l'orientation et de la courbure calculées à partir de deux échelles d'observation et deux niveaux de détails différents, nous présentons et discutons les caractéristiques basées sur les dimensions fractales dans la section suivante.

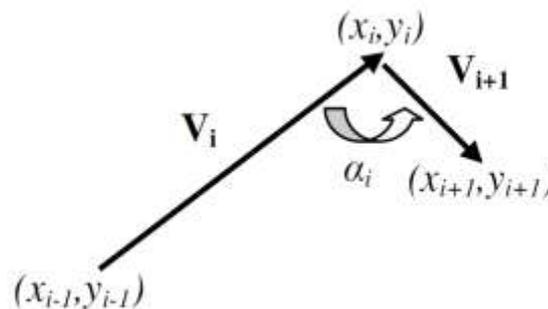


Figure 4.7. Angle entre les segments voisins des contours polygonisés.

4.4.2. Caractéristiques fractales

Le comportement fractal de l'écriture a d'abord été prouvé par Nicole Vincent [VIN 95]. Des études ultérieures ont révélé l'efficacité des caractéristiques fractales pour la caractérisation de scripteurs [SAB 00] ainsi que la classification des écritures [BOU 98]. Les auteurs dans la référence [BOU 98] calculent un graphe de lisibilité à partir de caractéristiques fractales et regroupent les écritures en classes en fonction de leur lisibilité. Les caractéristiques fractales ont également été appliquées à l'identification de scripteurs avec des taux de succès acceptables sur des petites bases de données [CHB 10]. Dans le but de capturer la régularité et la lisibilité de l'écriture, nous avons choisi de calculer la dimension fractale des écritures et, éventuellement, s'en servir comme caractéristique afin de discriminer entre les échantillons d'écriture des scripteurs masculins et féminins.

La dimension fractale peut être estimée par plusieurs méthodes et une étude détaillée de ces méthodes peut être trouvée dans [LOP 09]. Les catégories populaires de ces méthodes comprennent les méthodes de comptage de boîtes, les méthodes de mouvements browniens fractionnaires et les méthodes de mesure des surfaces. Dans notre implémentation, nous utilisons la méthode la plus connue de comptage de boîtes (Box-counting method) pour calculer la dimension fractale d'un texte manuscrit donné.

L'idée de base de la méthode de comptage de boîtes (Box-counting method) est de diviser l'objet (l'écriture dans notre cas) en un certain nombre de boîtes de dimensions r et compter le nombre de boîtes contenant des informations. Le processus est répété en faisant varier la taille de la boîte et la dimension fractale est estimée comme défini dans l'équation suivante.

$$D = \lim_{r \rightarrow 0} \frac{\log N(r)}{\frac{1}{r}}$$

où $N(r)$ est le nombre de boîtes de taille r nécessaire pour couvrir l'objet.

La dimension fractale est un nombre réel et une seule valeur peut ne pas être suffisamment discriminante pour être utilisée comme caractéristique d'un problème complexe comme la détermination du sexe à partir de documents manuscrits. Par conséquent, nous introduisons également l'analyse multi-fractale et calculons les dimensions fractales généralisées D_q en fonction des moments d'ordre q [VIC 90]. Le calcul de D_q s'appuie sur N points choisis aléatoirement appartenant à l'objet et sur le comptage du nombre de pixels $M_i(r)$ à l'intérieur des boîtes de dimension r pour chaque point i . Les dimensions généralisées D_q sont calculées en utilisant la moyenne de $M(r)$ pour les différentes valeurs de r [LOP 09].

Dans notre implémentation, nous calculons les dimensions généralisées D_q pour $q = 1, 2, 3, 5, 10$ qui sont combinées avec la dimension fractale calculée par la méthode de comptage de boîtes et ce afin d'avoir un vecteur de caractéristiques de six éléments pour estimer la régularité et la lisibilité des écritures.

Dans la section suivante, nous discutons de la troisième catégorie de caractéristiques, les caractéristiques texturales.

4.4.3. Caractéristiques texturales

L'analyse de texture de l'écriture considère chaque écriture comme une texture visuelle distincte. La texture est liée à l'aspect général de l'écriture et peut être représentée de plusieurs façons. Parmi les méthodes d'analyse la texture de l'écriture les plus significatives, Said et al. [SAI 00] ont employé les filtres de Gabor multi-canaux et les matrices de co-occurrence (GLCM) pour proposer une solution basée sur l'analyse de la texture au problème d'identification du scripteur. Certaines études récentes aussi ont utilisé les caractéristiques basées sur l'analyse de la texture de l'écriture pour l'identification du scripteur [SID 08] et la vérification du scripteur [HAN 12]. D'autres mesures de texture appliquées à l'écriture comprennent les motifs binaires locaux (Local Binary Patterns LBP) [BER 13] et les Coefficients autorégressifs [GAR 09]. La performance de ces descripteurs sur des tâches telles que l'identification du scripteur a été jugée meilleure que celle des filtres de Gabor multi-canaux ou des matrices de co-occurrence (GLCM). Par conséquent, nous avons choisi d'employer les motifs locaux binaires (LBP) ainsi que les coefficients autorégressifs comme descripteurs de texture pour une éventuelle discrimination entre les écritures des scripteurs masculins et féminins. Ces caractéristiques sont discutées ci-après.

4.4.3.1. Motifs binaires locaux (LBP)

Les motifs binaires locaux ont été introduits par Ojala et al dans [OJA 94, OJA 96] et ont été employés de manière très efficace jusqu'à nos jours pour différentes applications en analyse et classification de textures [HUA 11, BER 13, XIA 09].

La méthode originale des motifs locaux binaires (LBP) proposée dans [OJA 94, OJA 96] consiste à générer un nombre restreint d'unités de textures. En considérant un ensemble de voisinages $V = \{V_0, V_1, \dots, V_8\}$, on compare les pixels voisins au pixel central afin de générer un motif binaire. L'affectation binaire est réalisée comme suit : pour $i = 1, \dots, 8$, si $V_i < V_0$ on assigne au pixel voisin i la valeur 0, sinon on lui assigne la valeur 1. Ensuite, les valeurs binaires de ce motif sont multipliées par des poids et sommées pour calculer le code LBP du pixel central. Ce

processus est illustré par la figure 4.8. L'histogramme des codes LBP constitue le descripteur caractérisant la texture. En 2002, les auteurs ont étendu leur méthode originale pour inclure des voisinages de différentes tailles et pour capturer les caractéristiques dominantes à différentes échelles [OJA 02]. Les auteurs ont également introduit le concept de motifs binaires uniformes et non uniformes qui est basé sur le nombre de transitions spatiales dans le code LBP binaire (passage de 0 à 1 ou de 1 à 0).

Si le nombre de transitions est supérieur à deux, alors le motif est considéré comme non-uniforme. Sinon le motif est considéré comme uniforme. Le code LBP binaire peut alors être vu comme une chaîne circulaire. Par exemple, le code 00100100 n'est pas uniforme, car il contient 4 transitions mais les codes 00000000 et 00100000 sont uniformes, car ils contiennent 0 et 2 transitions respectivement.

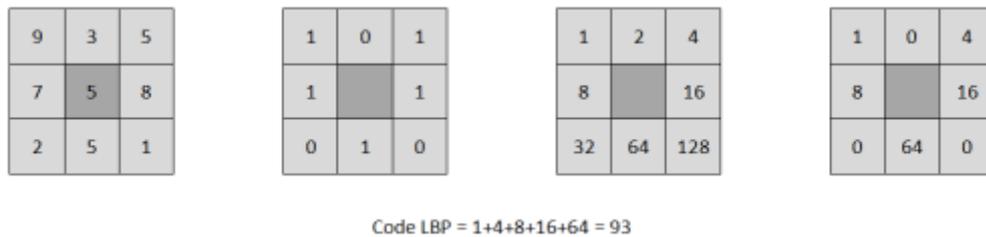


Figure 4.8. Construction d'un motif binaire et calcul du code LBP pour un pixel central.

Dans notre implémentation, nous calculons les LBP à partir d'images binaires de l'écriture. Pour p points voisins, nous pouvons avoir un maximum de $p \times (p - 1) + 2$ motifs uniformes. Nous utilisons un voisinage de $p = 16$ pixels avec un total de $242 = (16 \times (16 - 2) + 2)$ motifs uniformes possibles. Le descripteur (histogramme de motifs binaires locaux LBP) comprend donc 242 éléments pour les motifs uniformes et un élément pour tous les motifs non-uniformes, ce qui nous donne un vecteur de caractéristiques de dimension 243.

4.4.3.2. Coefficients autorégressifs

Les modèles autorégressifs bidimensionnels ont été introduits par K. Deguchi [DEG 86] pour la représentation des images ainsi que la caractérisation des textures. Ils ont été appliqués aussi avec succès à la segmentation de textures [SHU 93] et la modélisation de textures [MHI 93]. Récemment, les modèles autorégressifs ont été adaptés pour caractériser et identifier les auteurs de textes manuscrits [GAR 09].

Pour notre tâche de détermination de sexe, nous caractérisons une écriture donnée par un ensemble de coefficients autorégressifs bidimensionnels extraits à partir des images binaires de textes manuscrits. Pour estimer ces coefficients, la valeur de chaque pixel $x_{i,j}$ dans l'image est prédite par une combinaison linéaire des valeurs de ses pixels voisins.

$$x_{i,j} = \sum_{p,q \in D} \theta_{pq} x_{i-p} y_{j-q}$$

Lorsque D représente le contexte de voisinage qui est généralement une fenêtre rectangulaire $D = \{(p, q) | -m \leq p \leq m, -n \leq q \leq n, (p, q) \neq (0, 0)\}$. θ_{pq} sont les coefficients autorégressifs tandis que $p \times q$ est l'ordre du modèle. Les coefficients sont estimés en minimisant l'erreur quadratique entre la valeur prédite et la valeur réelle du pixel. Les détails de l'estimation des coefficients peuvent être trouvés dans [GAR 09]. Dans notre implémentation, nous avons utilisé un voisinage de 5×5 pixels qui donne un total de 24 coefficients autorégressifs.

En résumé, les échantillons d'écritures manuscrites sont représentés par trois types de caractéristiques : l'orientation et la courbure, les dimensions fractales ainsi que des caractéristiques basées sur la texture. Le tableau 4.2 résume ces caractéristiques avec leurs dimensionnalités.

Catégorie	Caractéristique	Description de la caractéristique	Dimension
Orientation et Courbure	f1	Distribution de codes de Freeman	8
	f2	Distribution de paires de codes de Freeman	64
	f3	Distribution des pentes de segments	8
	f4	Distribution des courbures	8
Dimension fractale	f5	Dimension fractale par la méthode de comptage de boîtes	1
	f6	Dimension fractale généralisée	5
Texture	f7	Motif local binaire	243
	f8	Coefficients autorégressifs	24
Total			361

Tableau 4.2. Récapitulatif des caractéristiques.

4.5. Classification

La classification est réalisée en utilisant deux classifieurs de l'état de l'art, les réseaux de neurones artificiels (ANN) et les séparateurs à vaste marge (SVM). Les classifieurs sont entraînés à l'aide des trois ensembles de caractéristiques extraites de l'ensemble de données d'apprentissage (d'entraînement) tandis que les différents paramètres des deux classifieurs sont déterminés empiriquement sur l'ensemble des données de validation.

Le réseau de neurones artificiels (ANN) est un réseau à trois couches, la couche d'entrée ayant le même nombre de neurones que la dimension d'un ensemble particulier de caractéristiques, la couche de sortie comprend deux neurones correspondant aux deux classes (masculine et féminine), tandis que le nombre de neurones dans la couche cachée est déterminé en fonction de la dimension du vecteur de caractéristiques (à l'aide de l'ensemble de données de validation).

Le classifieur SVM est basé sur un noyau polynomial. Comme pour le classifieur ANN, les paramètres du SVM sont déterminés de façon empirique sur l'ensemble de données de validation. Pour la mise en œuvre, nous avons utilisé la boîte à outils Matlab des SVM décrite dans [CAN 05]. Les performances de ces deux classifieurs ainsi que les différentes évaluations effectuées sont discutées en détails dans la section suivante.

4.6. Résultats expérimentaux

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 les bases de données QUWI et MSHD présentées dans la section 4.3. Nous présentons d'abord les taux de classification sur les ensembles de données complètes en utilisant les deux classifieurs (SVM et ANN) et nous décrivons plus loin quelques scénarios intéressants. 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 4.3 présente les taux de classification sur les deux bases de données. Pour la base QUWI, des échantillons de 300 scripteurs sont utilisées pour l'apprentissage (l'entraînement), tandis que ceux de 100 scripteurs (400 échantillons) sont utilisés pour les tests. Pour la base MSHD, les échantillons des 42 premiers scripteurs sont utilisés comme base d'apprentissage et les échantillons des 42 scripteurs restants sont utilisés comme base de test. Des taux de bonne classification de l'ordre de 68,75% et 73,02% sont atteints sur les bases de données QUWI et MSHD respectivement. Ces résultats sont comparables à ceux de l'état de l'art discutés dans la section 4.1. Il 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 des deux classifieurs (ANN et SVM), on peut constater d'après le tableau 4.3 qu'il n'y a pas une différence importante entre les deux. Parmi les trois types de caractéristiques, celles de l'orientation et de la courbure sont plus performantes que celles extraites à partir de la texture et des dimensions fractales sur les deux bases de données. Une autre observation aussi intéressante est que la combinaison des trois types de caractéristiques entraîne des améliorations marginales dans les taux globaux de classification.

Pour les expérimentations ultérieures, nous allons donc discuter les résultats des différentes catégories de caractéristiques seulement. 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, notamment, en modes dépendant et indépendant du texte et du script ainsi que des évaluations inter-bases de données (cross database evaluations).

Base de données	QUWI		MSHD	
	SVM	ANN	SVM	ANN
Orientation et courbure	68.75%	67.00%	72.82%	69.25%
Texture	59.75%	61.50%	68.65%	64.88%
Dimensions fractales	61.50%	62.50%	62.30%	61.90%
Combinaison	68.75%	67.50%	73.02%	69.44%

Tableau 4.3. Taux de classification sur les bases QUWI et MSHD.

4.6.1. Evaluations en modes dépendant et indépendant du texte

Ces expérimentations visent à étudier la façon dont les performances des différentes caractéristiques varient en fonction du contenu textuel des échantillons d'écriture dans les ensembles d'apprentissage et de test. Pour la base QUWI, nous utilisons la page 2 de chaque scripteur pour les textes arabes ainsi que la page 4 de chaque scripteur pour les textes anglais pour les évaluations dépendantes du texte. La page 2 (page 4) de 300 scripteurs est utilisée pour l'entraînement tandis que la même page de 100 autres scripteurs est utilisée pour les tests. Pour les évaluations indépendantes du texte, nous considérons les pages 1 et 3 pour les textes arabes et anglais, respectivement.

Étant donné que tous scripteurs de la base de données MSHD ont copié 6 échantillons en français et 6 autres en arabe, nous évaluons le système 12 fois en utilisant l'échantillon i des 42

premiers scripteurs dans l'entraînement (l'apprentissage) et le même échantillon des 42 derniers scripteurs dans les tests avec $i = 1, 2, \dots, 12$.

Le taux moyen de classification pour $i = 1, 2, \dots, 6$ représente la performance sur les échantillons français en mode dépendant du texte, tandis que la même valeur pour $i = 7, 8, \dots, 12$ représente le taux de classification sur des échantillons arabes en mode dépendant du texte. Pour les évaluations en mode indépendant du texte, les 3 premiers échantillons français (arabes) des 42 premiers scripteurs sont utilisés pour l'entraînement (l'apprentissage) et les 3 derniers échantillons des 42 scripteurs restants sont utilisés pour les tests. Plus tard, les 3 dernières images de textes français (arabes) sont utilisées comme ensemble d'entraînement (d'apprentissage) et les 3 premiers comme ensemble de test. Les taux globaux de classification pour ces expérimentations sont évalués selon la moyenne des deux exécutions.

Il convient également de noter que, dans tous les sous-ensembles de données mentionnés ci-dessus, il n'y a aucun scripteur appartenant à la fois à l'ensemble d'apprentissage et à la base de test. Les résultats de ces évaluations sont résumés dans le tableau 4.4. En général, les taux de classification des expérimentations en modes dépendants et indépendants du texte sont du même ordre que ceux présentés au tableau 4.3, surtout pour les attributs de l'orientation et la courbure qui donnent de meilleures performances par rapport aux caractéristiques texturales et fractales dans la plupart des cas.

Base de données	Mode Caractéristique	Dépendant du texte		Indépendant du texte	
		SVM	ANN	SVM	ANN
QUWI-Anglais	Orientation & courbure	68.00%	70.00%	70.00%	66.00%
	Texture	63.00%	62.00%	62.00%	61.00%
	Fractale	65.00%	65.00%	65.00%	64.00%
QUWI-Arabe	Orientation & courbure	69.00%	71.00%	63.00%	62.00%
	Texture	65.00%	63.00%	63.00%	63.00%
	Fractale	66.00%	66.00%	62.00%	65.00%
MSHD-Français	Orientation & courbure	68.25%	67.06%	67.46%	66.27%
	Texture	66.67%	66.27%	66.27%	65.48%
	Fractale	64.68%	66.27%	63.09%	65.87%
MSHD-Arabe	Orientation & courbure	73.41%	72.62%	68.65%	69.44%
	Texture	74.20%	72.22%	72.22%	71.43%
	Fractale	65.08%	65.87%	64.28%	65.08%

Tableau 4.4. Taux de classification des évaluations dépendantes et indépendantes du texte sur les bases de données QUWI et MSDB.

En comparant les taux de classification enregistrés en modes dépendants et indépendants du texte, à l'exception d'une seule expérimentation (la base QUWI-Anglais avec le classifieur SVM), les taux de classification des évaluations dépendantes du texte sont meilleurs (bien que légèrement, dans la plupart des cas) que ceux des évaluations indépendantes du texte. Une autre observation très intéressante est que les caractéristiques de l'orientation et de la courbure, sont, en général, plus sensibles au contenu textuel des images, contrairement aux caractéristiques texturales et fractales. Ceci semble très naturel parce que l'orientation et la courbure représentent des caractéristiques locales de l'écriture et sont donc plus sensibles au contenu de l'image. Les caractéristiques texturales et fractales étant des attributs globaux de l'écriture sont relativement moins sensibles au contenu textuel et donc présentent moins de variations dans les taux de classification en modes dépendants et indépendants du texte. Dans

la section suivante, nous présentons les résultats des évaluations dépendantes et indépendantes du script.

4.6.2. Evaluations en modes dépendant et indépendant du script

L'objectif de ces expérimentations est d'étudier la façon dont les taux de classification varient si des documents du même ou de différents scripts sont utilisés comme ensembles d'entraînement (d'apprentissage) et de test. Dans les expérimentations dépendantes du script, des échantillons d'écriture du même script (anglais, français ou arabe) sont utilisés aussi bien dans les ensembles d'entraînement (d'apprentissage) et de test alors que pour les évaluations indépendantes de script, les ensembles d'apprentissage et de test contiennent des échantillons d'écriture de scripts différents. Les résultats de ces expérimentations sont résumés dans les tableaux 4.5 et 4.6 respectivement.

Ensemble d'apprentissage	Ensemble de Test	Caractéristique	Taux de classification	
			SVM	ANN
QUWI-Anglais	QUWI-Anglais	Orientation & courbure	68.50%	66.50%
		Texture	60.00%	61.50%
		Fractale	63.50%	63.00%
QUWI-Arabe	QUWI-Arabe	Orientation & courbure	68.50%	65.00%
		Texture	61.50%	62.00%
		Fractale	61.50%	61.50%
MSHD-Français	MSHD-Français	Orientation & courbure	67.06%	69.44%
		Texture	70.63%	68.25%
		Fractale	62.30%	61.51%
MSHD-Arabe	MSHD-Arabe	Orientation & courbure	76.98%	73.41%
		Texture	70.63%	71.41%
		Fractale	61.51%	62.30%

Tableau 4.5. Taux de classification des évaluations dépendantes du script sur les bases de données QUWI et MSHD

Sur la base de données QUWI, Lors des évaluations dépendantes du script, en utilisant les textes arabes, les pages 1 et 2 issues de 300 scripteurs différents sont utilisées pour l'apprentissage et ces mêmes pages de 100 autres scripteurs sont utilisées dans les tests. Pour les expérimentations en utilisant les textes anglais, les pages 3 et 4 sont utilisées. Ce qui fait un total de 600 échantillons d'apprentissage et 200 échantillons de test pour chacun des deux scripts (anglais et arabe).

Sur la base de données MSDB, 6 échantillons de 42 scripteurs sont utilisés comme ensembles d'apprentissage et de test pour les textes français et arabes. Les expérimentations indépendantes du script impliquent des échantillons d'apprentissage dans un script différent de celui des échantillons du test.

En comparant les taux de classification dans les tableaux 4.5 et 4.6, on peut constater que les évaluations dépendantes du script donnent de meilleures performances que les évaluations indépendantes de script. Naturellement, quand des échantillons d'écriture du même script sont utilisés pour l'entraînement et pour le test, le système réalise de meilleurs taux de classification par rapport à ceux obtenus quand des échantillons de différents scripts sont utilisés pour l'apprentissage et le test. Comme nous l'avons déjà constaté lors des évaluations dépendantes et indépendantes du texte (section 4.6.1), les caractéristiques de l'orientation et de la courbure

donnent de meilleures performances comparées à celles calculées à partir de la texture ou des dimensions fractales, mais sont également plus sensibles au script étudié (considéré).

Ensemble d'apprentissage	Ensemble de Test	Caractéristique	Taux de classification	
			SVM	ANN
QUWI-Anglais	QUWI-Arabe	Orientation & courbure	60.00%	64.00%
		Texture	60.00%	62.50%
		Fractale	67.00%	65.00%
QUWI-Arabe	QUWI-Anglais	Orientation & courbure	60.50%	65.00%
		Texture	54.00%	60.00%
		Fractale	62.50%	63.00%
MSHD-Français	MSHD-Arabe	Orientation & courbure	69.05%	69.44%
		Texture	69.05%	68.65%
		Fractale	70.63%	69.84%
MSHD-Arabe	MSHD-Français	Orientation & courbure	57.14%	61.90%
		Texture	52.38%	60.71%
		Fractale	57.94%	61.11%

Tableau 4.6. Taux de classification des évaluations dépendantes du script sur les bases de données QUWI et MSHD

4.6.3. Evaluations inter-bases de données

Dans la dernière série d'expérimentations, nous utilisons les échantillons d'écriture d'une base de données dans l'apprentissage et ceux de l'autre base de données dans les tests. Étant donné que les deux bases de données ont été développées dans deux pays différents, il serait très intéressant d'analyser si des scripteurs masculins/féminins de différentes régions géographiques partagent certaines caractéristiques communes.

Une inspection des résultats du tableau 4.7 révèle que des taux de classification acceptables sont obtenus lorsque le système est entraîné en utilisant des échantillons d'écriture d'une base de données et testé en utilisant une base totalement différente.

Les taux de classification sont meilleurs lorsque les ensembles de données QUWI (arabe et anglais) sont utilisés pour l'entraînement (l'apprentissage) et les ensembles de données MSHD (français et arabe) sont utilisés pour les tests par rapport au cas inverse où les ensembles de données MSDH (arabe et français) sont utilisés pour l'entraînement et les ensembles de données QUWI (arabe et anglais) sont utilisés pour les tests. Ceci peut être attribué au fait que la taille de l'ensemble de données de test est environ deux fois la taille de l'ensemble de données d'apprentissage dans le dernier cas, ce qui entraîne des taux de classification relativement bas. Compte tenu du fait que les deux bases de données ont été produites par des scripteurs de régions géographiques et de circonstances culturelles totalement différentes, des taux de classification pouvant aller jusqu'à 72% pour le script arabe et à 63,5% pour le script latin (anglais et français) sont très encourageants. Ils représentent également une indication sur le fait que les scripteurs appartenant à un genre particulier (masculin/féminin) partagent certaines caractéristiques communes qui sont, dans une certaine mesure, constantes entre les individus de différentes origines.

En résumé, les séries d'expérimentations que nous avons menées valident l'hypothèse qu'une corrélation existe entre une écriture manuscrite et le sexe de son scripteur. Parmi les trois catégories de caractéristiques que nous avons employées, les caractéristiques de l'orientation et

de la courbure se sont avérées les plus efficaces dans un certain nombre de scénarios d'évaluation.

Les évaluations dépendantes et indépendantes du texte et du script ont montré l'efficacité des caractéristiques proposées pour déterminer le sexe dans les scénarios mentionnés ci-dessus. Enfin, les taux de classification des évaluations inter-bases de données (cross-database evaluations) reflètent que le sexe peut être déterminé avec un taux de succès acceptable indépendamment du contexte du scripteur en question.

Ensemble d'apprentissage	Ensemble de Test	Caractéristique	Taux de classification	
			SVM	ANN
QUWI-Arabe	MSHD-Arabe	Orientation & courbure	72.22%	70.04%
		Texture	68.85%	69.44%
		Fractale	57.14%	60.00%
MSHD-Arabe	QUWI-Arabe	Orientation & courbure	55.13%	58.88%
		Texture	58.13%	58.37%
		Fractale	53.13%	56.25%
QUWI-Anglais	MSHD-Français	Orientation & courbure	57.74%	60.52%
		Texture	61.51%	63.49%
		Fractale	62.90%	62.50%
MSHD-Français	QUWI-Arabe	Orientation & courbure	56.75%	58.13%
		Texture	57.87%	57.87%
		Fractale	54.25%	54.75%

Tableau 4.7. Taux de classification des évaluations inter bases de données.

4.7. Conclusion

Ce chapitre présente une méthode efficace pour la reconnaissance du sexe d'un individu à partir de l'analyse automatique de son écriture manuscrite. Bien qu'il fût un domaine de recherche populaire en psychologie pour de nombreuses décennies, ce problème est relativement moins exploré par les chercheurs en informatique. Nous avons identifié un sous-ensemble d'attributs d'écriture discriminatoires parmi ceux proposés dans les différentes études psychologiques, ces attributs ont été estimés à partir d'images de documents manuscrits. Les trois catégories de caractéristiques que nous avons considérées dans notre étude incluent des caractéristiques extraites à partir de l'orientation et la courbure, de la texture ainsi que des dimensions fractales. Ces caractéristiques ont été utilisées pour entraîner deux classificateurs, des réseaux de neurones artificiels (ANN) et des séparateurs à vaste marge (SVM). L'efficacité de ces caractéristiques dans la détermination du sexe du scripteur d'un échantillon donné a été évaluée sur deux bases de données, les bases QUWI et MSHD. Le système réalise des taux de classification encourageants selon divers scénarios d'évaluation.

CONCLUSION ET PERSPECTIVES

Le travail de recherche mis en évidence dans ce mémoire est axé sur deux contributions principales qui se placent dans le cadre de l'analyse et la caractérisation de l'écriture manuscrite. Dans notre première contribution, nous exploitons l'écriture manuscrite pour la reconnaissance de scripteurs. Ce problème a connu un regain d'intérêt, au cours des dernières années, de la part de la communauté des chercheurs du domaine de l'analyse et la reconnaissance de documents. Ceci peut être justifié par ses applications diverses, telles que les utilisations juridiques classiques où il est fondamental d'authentifier l'auteur d'un document tel qu'un testament ou un acte de vente, l'identification ou l'authentification des auteurs dans le cadre de systèmes d'aide à l'analyse de manuscrits anciens, ou encore la diminution de la complexité de reconnaissance des textes multi-scripteurs en la ramenant à des problématiques mono-scripteurs.

Contrairement aux méthodes classiques proposées jusqu'à présent et qui sont conçues pour la reconnaissance de scripteurs dans un environnement monolingue, nous avons proposé une méthode permettant de reconnaître le scripteur d'un texte écrit dans un script (anglais, par exemple) à partir d'échantillons du même scripteur dans un script différent (tel que le Grec) et par conséquent valider l'hypothèse que le style d'écriture d'une personne demeure constant à travers différents scripts. Afin de caractériser les scripteurs de documents écrits dans différents scripts et langues, nous avons utilisé les distributions de longueurs de segments (Run-length) qui sont des caractéristiques indépendantes du script et qui ne nécessitent aucune connaissance préalable du script considéré. Les résultats que nous avons enregistrés sont très encourageants, reflètent l'efficacité des caractéristiques proposées dans un environnement multi-scripts et valident l'hypothèse précédemment avancée.

Dans notre deuxième contribution principale, nous avons exploité l'écriture manuscrite afin de déterminer si l'on pouvait, ou pas, prédire le sexe (genre) d'un individu d'une manière fiable à partir de cette dernière (son écriture). Cette hypothèse a fait l'objet de plusieurs travaux de recherche effectués par les psychologues à l'opposé de leurs homologues informaticiens qui n'ont pas accordé à ce problème l'importance qu'il mérite. A cet effet, nous avons proposé l'utilisation de l'inclinaison, de l'orientation, la rondeur, de la courbure, de la propreté, de la lisibilité ainsi que la texture de l'écriture afin d'avoir une caractérisation plus au moins acceptable des écritures masculines et féminines.

Des scénarios d'évaluations différents ont été considérés afin:

1. d'étudier l'influence du contenu textuel des échantillons utilisés en apprentissage et en test sur la tâche de détermination du sexe,
2. d'étudier si l'utilisation en test d'un script différent de celui utilisé en apprentissage peut affecter la détermination du sexe.
3. d'examiner l'influence de la localisation géographique des scripteurs sur la tâche de détermination du sexe.

Deux classifieurs classiques ont été utilisés pour l'évaluation de la capacité des caractéristiques proposées à prédire le sexe d'un individu à partir de son écriture, et les résultats enregistrés sur

deux bases de données contenant 475 et 84 scripteurs respectivement semblent intéressants et prometteurs.

Les deux contributions que nous avons proposées dans le cadre de notre travail de recherche nous ont permis d'aboutir à des résultats prometteurs, mais nous ont aussi ouvert plusieurs voies pouvant être exploitées dans le futur. Certaines des perspectives envisagées sont des extensions relatives aux deux approches proposées dans ce mémoire alors que d'autres sont plus générales.

Pour l'approche de reconnaissance de scripteurs et concernant les techniques de classification utilisées, nous pensons qu'il serait intéressant d'envisager l'utilisation d'autres techniques de classification que celles que nous avons adoptées dans la présente étude et d'étudier leur comportement dans un environnement multi-script. . Il serait intéressant aussi d'envisager et d'expérimenter des possibilités de combinaison de techniques de classification, en commençant par celles intégrant les deux types de classifieurs proposés dans cette étude.

Il est nécessaire de rappeler que la performance d'un système d'identification de scripteurs 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 scripteurs.

Enfin, et dans le but de confirmer les résultats présentés dans cette étude, une grande base de données devrait être créée, où chaque scripteur doit écrire dans plusieurs scripts: latin, grec, chinois, japonais, hébreu, arabe, etc. Évidemment, le scripteur doit être familiarisé avec ces scripts afin d'avoir son propre style d'écriture (sinon, le scripteur copie sans aucun style le caractérisant). C'est toutefois une tâche très difficile, car un scripteur ne connaît généralement que peu de langues et de scripts.

Un autre aspect très important qui fera l'objet de nos futures travaux dans le domaine de l'identification de scripteurs, c'est l'identification en ensemble-ouvert (open-set en anglais), il serait intéressant d'étendre le système d'identification de scripteurs proposé pour inclure un seuil de rejet afin de rejeter les scripteurs qui ne font pas partie de la base de données, autrement dit, intégrer au système actuel une option lui permettant de rejeter toute écriture n'appartenant à aucun scripteur de la base d'apprentissage.

La capacité de l'approche de reconnaissance de scripteurs proposé à analyser des écritures manuscrites de différentes résolutions et avec différentes épaisseurs des traits peut également être examinée. Aussi, il serait très intéressant d'évaluer le pouvoir discriminant de l'approche proposée et de montrer sa généralité en l'employant pour la vérification de signatures manuscrites avec la détection de falsifications et pour la classification de manuscrits anciens.

Concernant l'approche proposée pour la détermination du sexe (genre) d'un scripteur à partir de son écriture manuscrite, 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.

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.
- [ALD 07] Al-Dmour. A., Abu-Zitar. R., "Arabic writer identification based on hybrid spectral-statistical measures", *Journal of Experimental & Theoretical Artificial Intelligence*, Vol. 19, pp. 307-332, 2007.
- [ALM 08a] Al-Ma'adeed. S., Mohammed. E., Al-Kassis. D., "Writer identification using edge-based directional probability distribution features for Arabic words", In: *Proceedings of the 6th ACS/IEEE International Conference on Computer Systems and Applications (AICCSA 2008)*, Doha, Qatar, pp. 582 – 590, 2008.
- [ALM 08b] Al-Ma'adeed. S., Al-Kurbi. A.A., Al-Muslih. A., Al-Qahtani. R., Al-Kubisi. H., "Writer identification of Arabic handwriting documents using grapheme features", In: *Proceedings of the 6th ACS/IEEE International Conference on Computer Systems and Applications (AICCSA 2008)*, Doha, Qatar, pp : 923 – 924, 2008.
- [ALM 12] Al-Ma'adeed. S., Ayouby. W., Hassaine. A., and Aljaam. J.M., "QUWI: An Arabic and English handwriting dataset for offline writer identification". In: *Proceedings of 13th International Conference on Frontiers in Handwriting Recognition (ICFHR 2012)*, Bari, Italy, pp. 746 - 751, 2012.
- [ALP 04] Alpaydin. E., "Introduction to machine learning", MIT Press, 2004.
- [ALZ 05] Al-Zoubeidy L. M., Al-Najar H. F., "Arabic writer identification for handwriting images", In: *Proceedings of the International Arab Conference on Information Technology (ACIT'2005)*, Amman, Jordan, 2005, pp. 111 - 117.
- [ARA 77] Arazi. B., "Handwriting identification by means of run-length measurements". In *IEEE Transactions on systems, man and cybernetics*, vol. 7, N°. 12, pp. 878 - 881, 1977.
- [AUG 06] Augustin. E., Brodin. J.M., Carré. M., Geoffrois. E., Grosicki. E., Prêteux. F., "RIMES evaluation campaign for handwritten mail processing", In: *Proceedings of the 6th Workshop on Frontiers in Handwriting Recognition (IWFHR 2006)*. La Baule, France, p. 231- 235, 2006.
- [BAL 06] Ball. G.R., Srihari. S., Srinivasan. H., "Versatile search of scanned arabic handwriting", In: *Summit on Arabic and Chinese Handwriting Recognition (SACH'06)*, College Park, MD, USA, pp. 57 - 69, 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.
- [BEL 01] Belaïd. A., "Reconnaissance automatique de l'écriture et du document". *Pour la Science*, 22 pages, 2001.
- [BEL 92] Belaid. A., Belaid. Y., "Reconnaissance des formes: Méthodes et applications", InterEditions, 1992.
- [BEN 04] Bensefia. A., "Analyse de documents manuscrits: Identification et vérification de scripteur ", Thèse de doctorat de l'université de Rouen, France, 2004.

- [BEN 05] Bensefia. A., Paquet. T., Heutte. L., "A writer identification and verification system", In Pattern Recognition Letters, Vol. 26, N^o. 13, pp. 2080 – 2092, 2005.
- [BER 13] Bertolini. D., Oliveira. L.S., Justino. E., Sabourin. R., "Texture-based descriptors for writer identification and verification", In: Expert Systems with Applications Journal, Vol. 40, N^o. 6, pp. 2069 – 2080, 2013.
- [BIS 06] Bishop. C. M., "Pattern recognition and machine learning", p. 304, Springer, 2006
- [BLU 03] Blumenstein. M., Verma. B., Basli. H., "A novel feature extraction technique for the recognition of segmented handwritten characters", In: Proceedings of the 7th International Conference on Document Analysis and Recognition (ICDAR 2003), Edinburgh, Scotland, UK, pp. 137 - 141, 2003.
- [BLU 07] Blumenstein. M., Liu. X.Y., Verma. B., "An investigation of the modified direction feature for cursive character recognition", In: Pattern Recognition, Vol. 40, N^o. 2, pp. 376 - 388, 2007.
- [BOU 98] Boultraeu. V., Vincent. N., Sabourin. R., Emptoz. H., "Handwriting and signature: One or two personality identifiers?", In; Proceedings of the 14th International Conference on Pattern Recognition (ICPR 1998), Brisbane, Australia, pp. 1758 - 1760, 1998.
- [BRO 29] Broom. M. E., Thompson. B., Bouton. M. T., "Sex differences in handwriting". In: Journal of Applied Psychology, Vol. 13, N^o. 2, pp. 159 - 166, 1929.
- [BUL 03] Bulacu. M., Schomaker. L., Vuurpijl. L., "Writer Identification Using Edge-Based Directional Features". In: Proceedings of the 7th International Conference on Document Analysis and Recognition (ICDAR 2003), Edinburgh, Scotland, UK, pp. 937 – 941, 2003.
- [BUL 05] Bulacu M., Schomaker. L., "A Comparison of Clustering Methods for Writer Identification and Verification", In: Proceedings of the 8th 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^o. 4, pp. 701 - 717.
- [BUL 07a] Bulacu. M., Schomaker. L., Brink. A., "Text-Independent Writer Identification and Verification on Offline Arabic Handwriting", In: Proceedings of the 9th International Conference on Document Analysis and Recognition (ICDAR 2007), Brazil, pp. 769 – 773, 2007.
- [BUR 02] Burr. V., "Judging gender from samples of adult handwriting: accuracy and use of cues", In: Journal of Social Psychology, Vol. 142, N^o. 6, pp. 691 - 700, 2002.
- [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.
- [CEDAR] http://www.cedartech.com/products_cedarfox.html
- [CHA 10] Chanda. S., Franke. K., Pal. U., Wakabayashi. T., "Text independent writer identification for Bengali script". In: Proceedings of the 19th International Conference on Pattern Recognition (ICPR 2008). Tampa, Florida, USA, pp. 2005 – 2008.

- [CHA 12] Chanda. S., Franke. K., Pal. U., "Text independent writer identification for Oriya script", In Proceedings of the International Workshop on Document Analysis Systems, Gold Coast, Queensland, Australia, pp. 369 – 373, 2012.
- [CHB 10] Chaabouni. A., Boubaker. H., Kherallah. M., Alimi. A.M., El-Abed. H., "Fractal and multi-fractal for arabic offline writer identification", In: Proceedings of the 20th International Conference on Pattern Recognition (ICPR 2010), Istanbul, Turkey, pp. 3793 - 3796, 2010.
- [CHB 11] Chaabouni. A., Boubaker. H., Kherallah. M., Alimi. A.M., El-Abed. H., "Combining of Off-line and On-line Feature Extraction Approaches for Writer Identification", In Proceedings of the 11th International Conference on Document Analysis and Recognition (ICDAR 2011), Beijing, China, pp. 1299 – 1303, 2011.
- [CHB 14] Chaabouni. A., Boubaker. H., Kherallah. M., El-Abed. H., Alimi. A.M., "Static and Dynamic Features for Writer Identification Based on Multi-fractals", In: International Arab Journal of Information Technology, Vol. 11, N°. 4, 2014.
- [CHE 11] Chen. J., Cheng. W., Lopresti. D., "Using perturbed handwriting to support writer identification in the presence of severe data constraints", In: Proceedings of the 18th International Conference on Document Recognition and Retrieval (DRR 2011), San Francisco, CA, USA, pp. 1 – 10, 2011.
- [CHU 12] Chung. S.J., "Diagnosis and treatment of hand tremor", In: Journal of the Korean medical association, Vol. 55, N°. 10, pp. 987 – 995, 2012.
- [COR 02] Cornuéjols. A., Miclet. L., Kodratoff. Y., "Apprentissage artificiel : Concept et algorithmes", Eyrolles, pp. 638, 2002.
- [CRE 95] Crettez. J. P., "Premier Degré de Caractérisation des Ecritures Manuscrites", Dans : Traitement du Signal, Vol. 12, N°. 6, pp. 567 - 573, 1995.
- [CRE 98] Crettez. J.P., Lorette. G., "Reconnaissance de l'écriture manuscrite", pp: 15, 1998.
- [CRO 10] Crosier. M., Griffin. L., "Using Basic Image Features for Texture Classification", In: International Journal of Computer Vision, Vol. 88, N°. 3, pp. 447 – 460, 2010.
- [DAL 08] Dalitz. C., Droettboom. M., Pranzas. B., Fujinaga. I., "A Comparative Study of Staff Removal Algorithms", In: IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 30, N°. 5, pp. 753 – 766, 2008.
- [DEG 86] Deguchi. K., "Two-dimensional auto-regressive model for analysis and sythesis of gray-level textures", In: Proceedings of the 1st International Symposium for Science on Form, pp. Tokyo, Japan, 441 - 449, 1986.
- [DEH 03] Dehkordi. M.E., Sherkat. N., Allen. T., "Handwriting style classification", In: International Journal of Document Analysis and Recognition, Vol. 6, N°. 1, pp. 55 - 74, 2003.
- [DIL 98] Dillon. K. M., "Narcissism and embellishments of signature", In: Psychological Reports, vol. 62, pp. 152–154, 1998.
- [DJE 08] Djeddi. C., Souici-Meslati. L., "Une approche locale en mode indépendant du texte pour l'identification de scripteurs : Application à l'écriture arabe", Dans Actes du Colloque International Francophone sur l'Ecrit et le Document (CIFED 2008), Rouen, Fance, pp 151 – 156, 2008.

- [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 11] Djeddi. C., Souici-Meslati. L., "Artificial immune recognition system for Arabic writer identification", In: Proceedings of the 5th International Symposium on Innovation in Information and Communication Technology (ISIICT 2011), Amman, Jordan, pp. 159 – 165, 2011.
- [DJE 12a] Djeddi. C., Souici-Meslati. L., Ennaji. A., "Writer recognition on Arabic handwritten documents". In : Proceedings of the 5th International Conference on Image and Signal Processing. Agadir, Morocco, pp. 493 – 501, 2012.
- [DJE 12b] Djeddi. C., Siddiqi. I., Souici-Meslati. L., Ennaji. A., "Multi-script writer identification optimized with retrieval mechanism". In : Proceedings of the 13th International Conference on Frontiers in Handwriting Recognition (ICFHR 2012). Bari, Italy, pp. 507 – 512, 2012.
- [DJE 13] Djeddi. C., Siddiqi. I., Souici-Meslati. L., Ennaji. A., "Codebook for writer characterization: A vocabulary of patterns or a mere representation space?", In: Proceedings of the 12th International Conference on Document Analysis and Recognition, (ICDAR 2013), Washington, USA, pp. 423 - 427.
- [DJE 13a] Djeddi. C., Siddiqi. I., Souici-Meslati. L., Ennaji. A., "Text-independent writer recognition using multi-script handwritten texts", In Pattern Recognition Letters, Vol. 34, N° 10, pp: 1196 – 1202.
- [DJE 14] Djeddi. C., Souici-Meslati. L., Siddiqi. I., Ennaji. A., El-Abed. H., Gattal. A., "Evaluation of Texture Features for Offline Arabic Writer Identification", To appear in Proceedings of the International Workshop on Document Analysis Systems (DAS 2014), France, 2014.
- [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.
- [FED 06] Feddaoui. N., Hamrouni. K., "Personal identification based on texture analysis of Arabic handwriting text", In: Proceedings of the 2nd IEEE International Conference On Information & Communication Technologies: From Theory To Applications, Damascus, Syria, pp. 1302 – 1307, 2006.
- [FER 08] Ferrari. V., Fevrier. L., Jurie. F., Schmid. C., "Groups of adjacent contour segments for object detection", In: IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 30, N°. 1, pp. 36 – 51, 2008.
- [FIE 13] Fiel. S., Sablatnig. R., "Writer Identification and Writer Retrieval using the Fisher Vector on Visual Vocabularies", In: Proceedings of the 12th International Conference on Document Analysis and Recognition (ICDAR 2013), Washington, USA, pp: 545 – 549, 2013.
- [FIS 12] Fisher. J., Maredia. A., Nixon. A., Williams. N., Leet. J., "Identifying Personality Traits, and Especially Traits Resulting in Violent Behavior through Automatic Handwriting Analysis", In: Proceedings of the Student-Faculty Research Day, CSIS, Pace University, May 4th, 2012.
- [FOR 08] Fornes. A., Lladós. J., Sanchez. G., "Writer identification in old handwritten music scores", In: The IAPR Workshop on Document Analysis Systems (DAS 2008), Nara, Japan, pp. 347 – 353, 2008.

- [FOR 11] Fornés. A., Dutta. A., Gordo. A., Lladós. J., "The ICDAR 2011 Music Scores Competition: Staff Removal and Writer Identification", In: Proceedings of the 11th International Conference on Document Analysis and Recognition (ICDAR 2011), Beijing, China, pp. 1511 – 1515, 2011.
- [FOR 12] Fornés. A., Dutta. A., Gordo. A., Lladós. J., "CVC-MUSCIMA: A Ground-truth of Handwritten Music Score Images for Writer Identification and Staff Removal", In International Journal on Document Analysis and Recognition, Vo. 15, N°. 3, pp. 243 - 251, 2012.
- [FRA 04] Franke. K., Schomaker. L., Vuurpijl. L., Van Herp. M., Guyion. I., "WANDA : A Common Ground for Forensic Handwriting Examination and Writer Identification", In: European Network of Forensic Handwriting Experts (ENFHEX) news, pp: 23-47, 2004.
- [FRE 08] Freitas. C., Oliveira. L.S., Sabourin. R., Bortolozzi. F., "Brazilian forensic letter database", In: Proceedings of the 11th International Conference on frontiers on handwriting recognition (ICFHR 2008), Montréal, Québec, Canada, 2008.
- [FRI 01] Friedman. J.H., "Greedy function approximation: a gradient boosting machine," In: the Annals of Statistic, vol. 29, N°. 5, pp. 1189 – 1232, 2001.
- [FRI 02] Friedman. J.H., "Stochastic gradient boosting," In: Journal of Computational Statistics and Data Analysis, vol. 38, N°. 4, pp. 367 – 378, 2002.
- [GAL 75] Galloway. M.M., "Texture analysis using gray level run lengths," In: Computer Graphics Image Processing, Vol. 4, N°. 2, pp. 172 – 179, 1975.
- [GAR 09] Garain. U., Paquet. T., "Off-line multi-script writer identification using AR coefficients," In: Proceedings of the 10th International Conference on Document Analysis and Recognition (ICDAR 2009). Barcelona, Spain, pp. 991 – 995, 2009.
- [GAZ 05] Gazzah. S., Essoukri Ben Amara. N., "Writer Identification using SVM Classifier and Genetic Algorithm for Optimal features selection", In : Proceedings of the International Arab Conference on Information Technology (ACIT 2005), Amman, Jordan, pp. 461-466, 2005.
- [GAZ 06] Gazzah. S., Essoukri Ben Amara. N., "Une approche a priori pour l'identification du scripteur en reconnaissance optique de l'écriture arabe", Dans Actes du 9^{ème} Colloque International Francophone sur l'Ecrit et le Document, Fribourg, Suisse, 18-21 septembre 2006, pp : 241 – 246.
- [GAZ 07] Gazzah. S., Essoukri Ben Amara. N., "Arabic Handwriting Texture Analysis for Writer Identification Using the DWT-Lifting Scheme". In: Proceedings of the 9th International Conference on Document Analysis and Recognition (ICDAR 2007), Curitiba, Paraná, Brazil, pp. 1133 – 1137, 2007.
- [GIL 11] Gilliam. T., "Writer identification in medieval and modern handwriting", PHD Thesis, University of York, 2011.
- [GOO 45] Goodenough. F.L., "Sex differences in judging the sex of handwriting", In: Journal of Social Psychology, Vol. 22, pp. 61 - 68, 1945.
- [GOR 13] Gordo. A., Fornés. A., Valveny. E., "Writer identification in handwritten musical scores with bags of notes", In: Pattern Recognition Journal, Vol. 46, N°. 5, pp. 1337 – 1345, 2013.
- [GRO 08] Grosicki. E., Carré. M., Brodin. J., Geoffrois. E., "RIMES evaluation campaign for handwritten mail processing", Proceedings of the 11th International Conference on

frontiers on handwriting recognition (ICFHR 2008), Montreal, Québec, Canada, 2008.

- [GRO 09] Grosicki. E., El-Abed. H., "ICDAR 2009 Handwriting Recognition Competition", In: Proceedings of the 10th International Conference on Document Analysis and Recognition (ICDAR 2009). Barcelona, Spain, pp. 1383 – 1387, 2009.
- [GRO 11] Grosicki. E., El-Abed. H., "ICDAR 2011: French handwriting recognition competition", In: Proceedings of the 11th International Conference on Document Analysis and Recognition (ICDAR 2011), Beijing, China, pp. 1459 – 1463.
- [GUP 08] Gupta. S., "Automatic Person Identification and Verification using Online Handwriting", Master of Science, International Institute of Information Technology Hyderabad, India, 2008.
- [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.
- [HAN 12] Hanusiak. R.K., Oliveira. L.S., Justino. E., Sabourin. R., "Writer verification using texture-based features". In: International Journal on Document Analysis and Recognition (IJAR), Vol. 15, N°. 3, pp. 213 - 226, 2012.
- [HAS 11] Hassaine. A., Al-Maadeed. S., Alja'am. J.M., Jaoua. A., Bouridane. A., "The ICDAR2011 Arabic writer identification contest", In: Proceedings of the 11th International Conference on Document Analysis and Recognition, Beijing, China, pp. 1470 – 1474, 2011.
- [HAS 12] Hassaine. A., Al-Maadeed. S., "ICFHR2012 Competition on writer identification - Challenge 2: Arabic scripts. ", In: Proceedings of 13th International Conference on Frontiers in Handwriting Recognition (ICFHR 2012), Bari, Italy, pp. 831– 836, 2012.
- [HAS 13] Hassaine. 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.
- [HE 08] He. Z., You. X., Tang. Y., "Writer identification of Chinese handwriting documents using hidden Markov tree model", In: Pattern Recognition Journal, Vol. 41, N°. 4, pp. 1295 – 1307, 2008.
- [HEL 08] Helli. B., Moghadam. M.E., "Persian Writer Identification Using Extended Gabor Filter", Lecture Notes in Computer Science, Vol. 5112, pp: 579 – 586, 2008, In: Proceedings of the 5th International Conference on Image Analysis and Recognition (ICIAR 2008), Póvoa de Varzim, Portugal, 2008.
- [HIR 96] Hirobumi. Y., Nakano. Y., "Cursive handwritten word recognition using multiple segmentation determined by contour analysis", In: IEICE Transactions on Information and Systems, E79-D(5):464-470, 1996.
- [HUA 11] Huang. D., Shan. C., Ardabilian. M., Wang. Y., and Chen. L., "Local binary patterns and its application to facial image analysis: A survey", In: IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews, Vol. 41, N°. 6, pp. 765 - 781, 2011.
- [HUB 99] Huber. R., Headrick. A., "Handwriting Identification: Facts and Fundamentals", CRC Press, 1999.
- [JAI 11] Jain. R., Doermann. D., "Offline writer identification using K-adjacent segments". In:

- Proceedings of the 11th International Conference on Document Analysis and Recognition (ICDAR 2011), Beijing, China, pp. 769 – 773, 2011.
- [JAI 13] Jain. R., Doermann. D., "Writer Identification Using an Alphabet of Contour Gradient Descriptors", In: Proceedings of the 12th International Conference on Document Analysis and Recognition (ICDAR 2013), Washington, USA, pp: 550 – 554, 2013.
- [JAM 91] James. H., "Sex differences in handwriting: A comment on spear", In: British Educational Research Journal, Vol. 17, N^o. 2, pp. 141 - 145, 1991.
- [JOH 78] Johansson. S., Leech. G., Goodluck. H., "Manual of Information to accompany the Lancaster-Oslo/Bergen Corpus of British English, for use with digital Computers", Technical report, Department of English, University of Oslo, Norway, 1978.
- [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.
- [KAR 84] Karin. W., Per-Erik. D., "A fast sequential method for polygonal approximation of digitized curves", In: Computer Vision, Graphics, and Image Processing, Vol. 28, N^o. 3, pp. 220 - 227, 1984.
- [KEM 06] Kameya. H., Mori. S., Oka. R., "A segmentation-free biometric writer verification method based on continuous dynamic programming", In: Pattern Recognition Letters, Vol. 27, N^o. 6, pp. 567 – 577, 2006.
- [KHE 11] Kherallah. M., Tagougui. N.M., Alimi. A., El-Abed. H., Margner. V., "Online Arabic Handwriting Recognition Competition", In: Proceedings of the 11th International Conference on Document Analysis and Recognition (ICDAR 2011), Beijing, China, pp. 1454 - 1458, 2011.
- [KIM 97] Kimura. F., Kayahara. N., Miyake. Y., Shridhar. M., "Machine and human recognition of segmented characters from handwritten words", In: Proceedings of the 4th International Conference on Document Analysis and Recognition (ICDAR 1997), Ulm, Germany, pp. 866 - 869, 1997.
- [KLE 13] Kleber. F., Fiel. S., Diem. M., Sablatnig. R., "CVL-Database: An Off-line Database for Writer Retrieval, Writer Identification and Word Spotting", In: Proceedings of the 12th International Conference on Document Analysis and Recognition (ICDAR 2013), Washington, USA, pp. 560-564, 2013.
- [KOH 89] Kohonen. T., "Self-Organization and Associative Memory", 3rd Edition, Springer-Verlag New York Inc, NYC, USA. p. 312, 1989.
- [LIU 11] Liu. C.L., Yin. f, Wang. D., Wang. Q., "CASIA Online and Offline Chinese Handwriting Databases", In: Proceedings of the 11th International Conference on Document Analysis and Recognition (ICDAR 2011), Beijing, China, pp. 37 – 41, 2011.
- [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^o. 1, pp. 87 – 92, 2011.
- [LOP 09] Lopes. R., Betrouni. N., "Fractal and multifractal analysis: A review", In: Medical Image

Analysis, Vol. 13, N°. 4, pp. 634 - 649, 2009.

- [LOU 11] Louloudis. G., Stamatopoulos. N., Gatos. B., "ICDAR 2011 - Writer identification contest", In: Proceedings of the 11th International Conference on Document Analysis and Recognition (ICDAR 2011). Beijing, China, pp. 1475 – 1479, 2011.
- [LOU 12] Louloudis. G., Gatos. B., Stamatopoulos. N., "ICFHR2012 Competition on writer identification challenge 1: Latin/Greek documents", In: Proceedings of the 13th International Conference on Frontiers in Handwriting Recognition (ICFHR 2012). pp. 825 – 830, 2012.
- [LOU 13] Louloudis. G., Gatos. B., Stamatopoulos. N., Papandreou. A., "ICDAR 2013 Competition on Writer identification", In: Proceedings of the 12th International Conference on Document Analysis and Recognition (ICDAR 2013), Washington, USA, pp. 1429 – 1433, 2013.
- [LOW 04] Lowe. D., "Distinctive image features from scale invariant Key points", In: International Journal of Computer Vision", Vol. 60, N°. 2, pp. 91 - 110, 2004.
- [MAA 86] Maarse. F.J., Schomaker. L., Thomassen. A.J.W.M., "The influence of changes in the effector coordinate systems on handwriting movements", In: H.S.R. Kao, G.P. van Galen, & R. Hoosain (Eds.), Graphonomics: Contemporary research in handwriting (pp. 33-46). Amsterdam: North-Holland, 1986.
- [MAH 12] Mahmoud. S.A., Ahmad. A., Al-shayeb. M., Al-Khatib. W.G., Parvez. M.T., Fink. G.A., Margner. V., EL-Abed. H., "KHATT: Arabic Offline Handwritten Text Database," In: Proceedings of the 13th International Conference on Frontiers in Handwriting Recognition (ICFHR 2012), Bari, Italy, pp. 447 – 452, 2012.
- [MAL 13] Malik. M.I., Liwicki. M., Alewijnse. L., Ohyama. W., Blumenstein. M., Found. B., "ICDAR 2013 Competitions on Signature Verification and Writer Identification for On- and Offline Skilled Forgeries (SigWiComp2013)", In: Proceedings of the 12th International Conference on Document Analysis and Recognition (ICDAR 2013), Washington, USA, pp. 1108– 1114, 2013.
- [MAR 01] Marti., U.V., Messerli. R., Bunke. H., "Writer Identification Using Text Line Based Features", In: Proceedings of the 6th International Conference of Document Analysis and Recognition (ICDAR 2001), Seattle, WA, USA, pp. 101-105, 2001.
- [MAR 02] Marti. U.V., Bunke. H., "The IAM-database: An English Sentence Database for Off-line Handwriting Recognition". In: International Journal on Document Analysis and Recognition, Vol. 5, N°. 1, pp: 39 - 46, 2002.
- [MAR 05] Margner. V., Pechwitz. M., El-Abed. H., "ICDAR 2005 Arabic handwriting recognition competition", In: Proceedings of the 8th International Conference on Document Analysis and Recognition (ICDAR 2005), Seoul, Korea, pp. 70 – 74, 2005.
- [MAR 07] Marquis. R., "Etude de caractères manuscrits : de la caractérisation morphologique à l'individualisation du scripteur", Thèse de doctorat, Université de Lausanne, 2007.
- [MAR 99] Marti. U.V., Bunke. H., "A full English sentence database for off-line handwriting recognition", In: Proceedings of the 5th International Conference on Document Analysis and Recognition (ICDAR 1999), Bangalore, India, p. 705 – 708, 1999.
- [MEU 89] Meulenbroek. R., VanGalen. G.P., "The production of connecting strokes in cursive writing: Developing coarticulation in 8 to 12 year-old children". In R. Plamondon, C. Y. Suen, et M. L. Simner (Eds.), Computer recognition and human production of

handwriting (pp. 273-286). Singapore: World Scientific Publishers, 1989.

- [MHI 93] Mhidra. H., Brochard. J., Leard. M., "AR models and bidimensional discrete moments applied to texture modelling and recognition". In: Pattern Recognition Journal, Vol. 26, N°. 5, pp. 721 - 26, 1993.
- [MIT 97] Mitchell T., "Machine Learning", McGraw-Hill, 1997.
- [NEW 26] Newhall. S. M., "Sex differences in handwriting", In: Journal of Applied Psychology, Vol. 10, N°. 2, pp. 151 - 161, 1926.
- [NOS 04] Nosary. A, Heutte. L, Paquet. T., "Unsupervised writer adaptation applied to handwritten text recognition", In: Pattern Recognition Journal, Vol. 37, N°. 2, pp. 385 – 388, 2004.
- [NOS 99] Nosary. A., Heutte. L., Paquet. T., Lecourtier. Y., "Defining Writer Invariants to Adapt the Recognition Task", In: Proceedings of the 5th International Conference on Document Analysis and Recognition (ICDAR 1999), Bangalore, India, pp. 765-768, 1999.
- [OJA 02] Ojala. T., Pietikainen. M., Maenpaa. T., "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns", In: IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 24, N°. 7, pp. 971 – 987, 2002.
- [OJA 94] Ojala. T., Pietikainen. M., Harwood. D., "Performance evaluation of texture measures with classification based on kullback discrimination of distributions", In: Proceedings of the 12th International Conference on Pattern Recognition (ICPR 1994), Jerusalem, Palestine, pp. 582 - 585, 1994.
- [OJA 96] Ojala. T., Pietikainen. M., Harwood. D., "A comparative study of texture measures with classification based on feature distributions", In: Pattern Recognition Journal, Vol. 29, pp. 51 - 59, 1996.
- [OTS 79] Otsu. N., "A threshold selection method from gray-level histogram", In: IEEE Transactions on systems, man and cybernetics, Vol. 9, N°. 1, pp. 62-66, 1979.
- [OZA 06] Ozaki. M., Adachi. Y., Ishii. N., "Examination of effects of character size on accuracy of writer recognition by new local arc method", In: Proceedings of the 10th International Conference on Knowledge-Based & Intelligent Information & Engineering Systems (KES2006), South Coast of the United Kingdom, pp. 1170 – 1175, 2006.
- [PEC 02] Pechwitz. M., Snoussi Maddouri. S., Märgner. V., Ellouze. N., Amiri. H., "IFN/ENIT-Database of handwritten arabic words", Dans Actes du Colloque International Francophone sur l'Écrit et le Document (CIFED'2002), Hammamet, Tunisie, pp. 129-136, 2002.
- [PER 07] Pervouchine. V., Leedham. G., "Extraction and analysis of forensic document examiner features used for writer identification", In Pattern Recognition Journal, Vol. 40, pp. 1004 – 1013, 2007.
- [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.
- [PUR 10] Purkait. P., Kumar. R., Chanda. B., "Writer identification for handwritten Telugu documents using directional morphological features", In: Proceedings of the 12th International Conference on Frontiers in Handwriting Recognition (ICFHR 2010), Kolkata, India, pp. 658 – 663, 2010.

- [RIC 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.
- [SAB 00] Sabourin. R., Emptoz. H., Vincent. N., Bouletreau. V., "How to use fractal dimensions to qualify writings and writers", In *Fractals*, Vol. 08, N° 1, pp. 85 - 97, 2000.
- [SAI 00] Said. H., Tan. T., Baker. K., "Personal identification based on handwriting". In: *Pattern Recognition Journal* , Vol. 33, pp. 149-160, 2000.
- [SCH 00] Schomaker. L., Vuurpijl. L., "Forensic writer identification: A benchmark data set and a comparison of two systems", Internal report for the Netherlands Forensic Institute, 2000.
- [SCL 04a] Schlapbach. A., Bunke. H., "Using HMM Based Recognizers for Writer Identification and Verification", In: *Proceedings of the 9th International Workshop on Frontier Handwriting Recognition (IWFHR 2004)*, Kokubunji, Tokyo, Japan, pp: 167- 172, 2004.
- [SCL 04b] Schlapbach. A., Bunke. H., "Off-Line Handwriting Identification Using HMM Based Recognizers", In: *Proceedings of the 17th International Conference on Pattern Recognition (ICPR 2004)*, Cambridge, UK, pp: 654 – 658, 2004.
- [SCL 05] Schlapbach. A., Kilchherr. V., Bunke. H., "Improving writer identification by means of feature selection and extraction". In: *Proceedings of the 8th International Conference on Document Analysis and Recognition (ICDAR 2005)*, Seoul, Korea, pp: 131 – 135, 2005.
- [SCL 06] Schlapbach. A., Bunke. H., "Off-line Writer Identification Using Gaussian Mixture Models", In: *Proceedings of the 18th International Conference on Pattern Recognition (ICPR 2006)*, Hong Kong, China, pp: 992 – 995, 2006.
- [SCH 07] Schomaker. L., Franke. K., Bulacu. M., "Using codebooks of fragmented connected-component contours in forensic and historic writer identification", In: *Pattern Recognition Letters*, Vol. 28, N° 6, pp. 719–727, 2007.
- [SCL 07a] Schlapbach. A., "Writer Identification and Verification", PhD Thesis, University of Bern, 2007.
- [SCL 07b] Schlapbach. A., Bunke. H., "A Writer Identification and Verification System Using HMM Based Recognizers", In: *Pattern Analysis and Applications Journal*, Vol. 10, N° 1, pp: 33 – 43, 2007.
- [SCH 08] Schomaker. L., "Writer identification and verification", In N. Ratha and V. Govindaraju, editors, "Sensors, Systems and Algorithms, *Advances in Biometrics*", pp. 247 – 264, Springer Verlag, 2008.
- [SCH 98] Schomaker. L., "From handwriting analysis to pen-computer applications", In *IEEE*

Electronics Communication Engineering Journal, Vol. 10, pp. 93 – 102, 1998.

- [SCL 07] Schlapbach. A., Bunke. H., "A writer identification and verification system using HMM based recognizers", In: Pattern Analysis and Applications Journal, Vol. 10, pp. 33 – 43, 2007.
- [SER 03] Seropian. A., "Analyse de Document et Identification de Scripteurs", Thèse de doctorat, Université de Toulon et du Var, France, 2003.
- [SES 12] Sesa. N. E., "An online writer recognition system based on in-air and on-surface trajectories", PHDThesis, Université polytechnique de Catalogne, 2012.
- [SHA 06] Shahabi. N.F., Rahmati. M., "Comparison of Gabor-based features for writer identification of Farsi/Arabic handwriting", In: Proceedings of the International Workshop on Frontiers in Handwriting Recognition (IWFHR 2006), La Baule, France, 2006, pp. 545-550, 2006.
- [SHA 07] Shahabi. N.F., Rahmati. M., "A New Method for Writer Identification and Verification Based on Farsi/Arabic Handwritten Texts", In: Proceedings of the 9th International Conference on Document Analysis and Recognition (ICDAR 2007), Brazil, pp. 829-833, 2007.
- [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.
- [SHU 93] Shunichiro. O., "Texture segmentation method by using two-dimensional AR model and kullback information", In: Pattern Recognition Journal, Vol. 26, N°. 2, pp. 237 - 244, 1993.
- [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 09a] Siddiqi. I., "Classification of Handwritten Documents : Writer Recognition", Phd Thesis, University of Paris Descartes, France, 2009.
- [SID 09b] Siddiqi. I., Cloppet. F., Vincent. N., "Contour based features for the classification of ancient manuscripts", In: Proceedings of the International Graphonomics Society. Dijon, France, pp. 226 – 229 2009.
- [SID 09c] Siddiqi. I., Vincent. N., "A set of chain code based features for writer recognition", In: Proceedings of the 10th International Conference on Document Analysis and Recognition (ICDAR 2009), Barcelona, Spain, pp. 981 - 985, 2009.
- [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.
- [SID 14] Siddiqi I., Djeddi C., Raza A., Souici-Meslati L., "Automatic analysis of handwriting for gender classification", Soumis pour publication dans un journal international spécialisé, en phase de révision
-
- [SIT 01] http://www.pageshalal.fr/actualites/les_manuscrits_de_tombouctou_-fr-213.html
- [SIT 02] <http://www.tuniscope.com/index.php/article/26541/actualites/politique/bechir-nefzi-menace-mort-432316#.Ujlk2yVOLIU>

- [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.
- [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.
- [SRI 05] Srihari. S., Srinivasan. H., Babu. P., Bhole. C., "Handwritten Arabic Word Spotting using the CEDARABIC Document Analysis System", In: Proceedings of the Symposium on Document Image Understanding Technology (SDIUT 2005), pp: 123 – 132, 2005.
- [SRI 07] Srihari. S., Srinivasan. H., Desai. H ., "Questioned Document Examination using CEDAR-FOX", In: Journal of Forensic Document Examination, Vol. 18, N°. 2, pp: 1 – 20, 2007.
- [SRI 08] Srihari. S., Ball. G., "Writer Verification of Arabic Handwriting". In: Proceedings of the 8th International Workshop on Document Analysis Systems (DAS 2008), Nara, Japan, pp: 28 – 34, 2008.
- [STR 09] Strassel. S., "Linguistic resources for Arabic handwriting recognition", In: Proceedings of the 2nd International Conference on Arabic Language Resources and Tools, Cairo, Egypt, April 2009.
- [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.
- [SUN 06] Sung-hyuk. C., Sungsoo. Y., "Handwriting copybook style identification for questioned document examination", In: Journal of Forensic Document Examination, Vol. 17, 2006.
- [TAN 09] Tan. G.X., Viard-Gaudin. C., Kot. A.C., "Automatic writer identification framework for online handwritten documents using character prototypes", In: Pattern Recognition Journal, Vol. 42, N°. 12, pp. 3313 – 3323, 2009.
- [TEN 34] Tenwolde. H., "More on sex differences in handwriting". In: Journal of Applied Psychology, Vol. 18, pp. 705 - 710, 1934.
- [TUY 08] Tuytelaars. T., Mikolajczyk. K.. "Local invariant feature detectors - A survey", In: Journal of Foundations and Trends in Computer Graphics and Vision, Vol. 3, N°. 3, pp. 177 - 280, 2008.
- [VAN 03] Van Erp. M., Vuurpijl. L., Franke. K., Schomaker. L., "The WANDA Measurement tool for forensic document examination", In: Proceedings of the 11th Conference of the International Graphonomics Society (IGS 2003), Scottsdale, Arizona, USA, pp. 282 – 285, 2003.
- [VAN 05] Van-der-Maaten. L., Postma. E., "Improving automatic writer identification", In: Proceedings of the 17th Belgium-Netherlands Conference on Artificial Intelligence, Brussels, Belgium, pp: 260 - 266, 2005.
- [VIA 99] Viard-Gaudin. C., Lallican. P.M, Knerr. S., Binter. P., "The IRESTE On/Off (IRONOFF) Dual Handwriting Database", In: Proceedings of the 5th International Conference on Document Analysis and Recognition (ICDAR 1999), Bangalore, India, pp: 455 – 458, 1999.
- [VIC 90] Vicsek. T., "Mass multifractals". In: Physica A: Statistical Mechanics and its Applications, Vol. 168, N°. 1, pp. 490 - 497, 1990.
- [VIN 95] Vincent. N., Emptoz. H., Fractal Reviews in the Natural and Applied Sciences, chapter A

classification of writings based on fractals, pages 320 - 331. Chapman and Hall, 1995.

- [VIS 13] Visani. M., Kieu. V.C., Fornés. A., Journet. N., "The ICDAR 2013 Music Scores Competition: Staff Removal", In: Proceedings of the 12th International Conference on Document Analysis and Recognition (ICDAR 2013), Washington, USA, pp. 1407 – 1411, 2013.
- [WAN 03] Wang. X., Ding. X., Liu. H. L., "Writer Identification Using Directional Element Features and Linear Transform", in: Proceedings of the 7th International Conference on Document Analysis and Recognition (ICDAR 2003), Edinburgh, Scotland, UK, pp. 942 – 945, 2003.
- [WAN 04] Wang. X., Ding. X., "An Effective Writer Verification Algorithm Using Negative Samples". In Procs of the 9th International Workshop on Frontiers in Handwriting Recognition (IWFHR 2004), Kokubunji, Tokyo, Japan, pp: 509 – 514, 2004.
- [WIL 96] William N.H., "Identifying sex from handwriting". In: Perceptual and Motor Skills, Vol. 83, pp. 91 - 800, 1996.
- [WOO 10] Woodard. J., Lancaster. M., Kundu. A., Ruiz. D., Ryan. J., "Writer recognition of Arabic text by generative local features", In: Proceedings of the 4th IEEE Biometrics Theory Applications and Systems (BTAS 2010), pp: 1 – 7, 2010.
- [XIA 09] Xiaoyu. W., Tony X.H., Shuicheng. Y., "An hog-lbp human detector with partial occlusion handling", In: Proceedings of the 12th International Conference on Computer Vision, Kyoto, Japan, pp. 32 - 39, 2009.
- [XU 11] Xu. L., Ding. X., Peng. L., Li. X., "An Improved Method Based on Weighted Grid Micro-structure Feature for Text-Independent Writer Recognition", In: Proceedings of the 11th International Conference on Document Analysis and Recognition (ICDAR 2011), Beijing, China, pp. 638 – 642, 2011.
- [ZHU 00] Zhu. Y., Tan. T, Wang. Y., "Biometric personal identification based on handwriting", In: Proceedings of the 15th International Conference on Pattern Recognition (ICPR 2000), Barcelona, Spain, pp : 797 – 800, 2000.
- [ZOI 00] Zois. E., Anastassopoulos. V., "Morphological waveform coding for writer identification", In: Pattern Recognition Journal, Vol. 33, N°. 3, pp.385–398, 2000.

PARTICIPATIONS AUX COMPETITIONS ORGANISEES DANS LE CADRE DES CONFERENCES ICDAR ET ICFHR

L'intérêt porté à la reconnaissance du scripteur et aux méthodes d'évaluation de ces systèmes de reconnaissances a évolué massivement les dernières années. A cet effet, et en raison de l'importance du domaine, des chercheurs de différentes équipes spécialisées ont organisé plusieurs compétitions depuis 2011 et ce, dans le cadre des conférences ICDAR (International Conference on Document Analysis and Recognition) 2011 [LOU 11, HAS 11, FOR 11] et 2013 [LOU 13, MAL 13] et ICFHR (International Conference on Frontiers in Handwriting Recognition) 2012 [LOU 12, HAS 12].

Le but de ces compétitions est de fournir une plate-forme pour l'évaluation comparative des méthodes développées par des chercheurs appartenant à des institutions scientifiques, industrielles 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 et de contenus et scripts différents, les protocoles d'évaluation aussi sont complètement différents. Les bases de données utilisées pour l'évaluation des différents systèmes participants ont été rendues publiques juste après la fin des conférences ICDAR 2011, ICFHR 2012 ainsi que ICDAR 2013.

Cette annexe met l'accent sur ces compétitions, auxquelles nous avons participé avec plusieurs propositions, en présentant, une brève description de chacune d'elles, une brève description des bases de données utilisées ainsi que les protocoles d'évaluation adoptés, les résultats enregistrés par les différents systèmes participants seront aussi présentés.

A.1. The ICDAR2011 Arabic Writer Identification Contest

"The ICDAR2011 Arabic Writer Identification Contest" représente la première compétition sur l'identification de scripteurs à partir de documents arabes, elle a été organisée dans le cadre de la conférence ICDAR 2011 [HAS 11] afin de permettre aux chercheurs et aux industriels dans le domaine de l'identification du scripteur ou des domaines connexes de comparer les performances de leurs systèmes sur de nouvelles bases de données non publiées. Cette compétition a été organisée par une équipe de chercheurs de l'Université de Qatar. Elle a attiré l'attention d'une trentaine de participants, parmi ces derniers, sept participants uniquement ont acceptés de partager leurs identités ainsi que de brèves descriptions de leurs méthodes.

A.1.1. Description de la base de données utilisée

Dans cette compétition, 54 scripteurs ont été invités à écrire trois paragraphes différents en arabe. Les deux premiers paragraphes ont été utilisés pour l'entraînement (l'apprentissage) des systèmes et le troisième a été utilisé pour les tests. Pour certains scripteurs, les deux premiers paragraphes ont été retirés de l'ensemble d'apprentissage, afin de tester la capacité des systèmes à détecter les scripteurs inconnus (non appris par le système). Les images ont été fournies au format PNG couleur, niveau de gris et binaire. La binarisation des images a été réalisée en utilisant la méthode d'Otsu.

A.1.2. Evaluation et résultats

Durant cette compétition, les participants ont été invités à produire, pour chaque document D de l'ensemble de test, et pour chaque scripteur S de l'ensemble d'apprentissage, un score de probabilité $P(D, S)$ indiquant la probabilité que le document D a été écrit par le scripteur S . Il est important de noter que $P(D, S)$ prend la valeur 0 pour un scripteur inconnu.

Le tableau A.1 montre les résultats enregistrés par les systèmes soumis par les sept équipes de recherche qui ont accepté de partager leurs identités. La meilleure performance est obtenue par l'équipe de l'UCL du Royaume-Uni qui a réussi à obtenir un taux d'identification de l'ordre de 100%. Il est à noter que notre système nommé *Wifahd* a réalisé un taux d'identification de l'ordre de 75.68%, ce taux n'est pas assez bon et ceci peut être justifié par le fait que nous n'avons pas prévu la possibilité de rejeter le scripteur dont on teste les documents s'ils ne correspondent à aucune des identités répertoriées dans la base de référence. Ces scripteurs doivent être alors considérés comme inconnus du système et doivent être rejetés. Dans le tableau A.1, la méthode qui a la plus grande précision est marquée en gras, et le classement de chaque méthode est présenté entre parenthèses.

Système	Pays	TOP-1
UCL	Royaume-Uni	100 (1)
Team Shasta	Etats Unis d'Amérique	89,19 (2)
Wride	Pays-Bas.	81,08 (3)
Eu Jin Lok	Australie	78,38 (4)
Intelligentia	Autriche	78,38 (5)
Wifahd	Algérie	75,68 (6)
Robin	Royaume-Uni	5,45 (15)

Tableau A.1. Performances des systèmes participants à la compétition "The ICDAR2011 Arabic Writer Identification Contest" en %.

A.2. The ICDAR 2011 Music Scores Competition: Staff Removal and Writer Identification

La compétition "The ICDAR 2011 Writer Identification on Music Scores Competition" [FOR 11], est la deuxième compétition qui a été organisée en conjonction avec la conférence ICDAR 2011, elle a été proposée par une équipe de recherche du Groupe d'analyse de documents du centre de vision par ordinateur de l'université Autonoma de Barcelona, Barcelone, Espagne. Huit systèmes différents ont été soumis par des équipes de recherche de l'université de Tébessa, Algérie, de l'université Badji-Mokhtar d'Annaba, Algérie, de l'université de Northumbria, London, Royaume-Uni ainsi que de l'université de Qatar, Qatar. L'objectif essentiel de cette compétition

consiste à évaluer la capacité des systèmes soumis à reconnaître le scripteur à partir de partitions musicales.

A.2.1. Description de la base de données utilisée

Les organisateurs de cette compétition ont proposé l'utilisation de la base de données CVC-MUSCIMA [FOR 12, DAL 08] contenant 1000 images de notes musicales écrites par 50 musiciens différents (voir section 1.6.10), cette base de données est divisée en deux parties, la première se compose de 500 images (10 images pour chaque scripteur) et elle a été utilisée pour l'apprentissage, et la deuxième partie composée des 500 images restantes a été utilisée pour les tests. Les organisateurs ont fournis des images sans les lignes de portées (voir figure A.1) car elles sont particulièrement utiles pour l'identification du scripteur et afin faciliter la participation des chercheurs à cette compétition.



Figure A.1. Un échantillon de la base CVC.

A.2.2. Evaluation et résultats

Les résultats des systèmes soumis à cette compétition sont présentés au tableau A.2. On peut constater que la plupart des méthodes obtiennent des taux d'identification aux alentours de 65%. Les meilleurs systèmes sont "PRIP02-combinaison" et "TUA03-SVMOAA" (l'un de nos 5 systèmes soumis à cette compétition), qui en effet enregistrent des résultats très similaires (77% et 76,6% respectivement). Dans le tableau A.2, la méthode qui a la plus grande précision est marquée en gras, et le classement de chaque méthode est présenté entre parenthèses.

Système	Pays	TOP-1
PRIP02-edges	Qatar	65,4 (5)
PRIP02-grapheme	Qatar	63,8 (7)
PRIP02-combinaison	Qatar	77,0 (1)
TUA03-5NN	Algérie	53,4 (8)
TUA03-MLP	Algérie	64,8 (6)
TUA03-SVMOAA	Algérie	76,6 (2)
TUA03-SVMOAO	Algérie	66,6 (4)
TUA03-combinaison	Algérie	70,4 (3)

Tableau A.2. Performances des méthodes participantes à la compétition "The ICDAR 2011 Music Scores Competition: Staff Removal and Writer Identification" en %.

A.3. ICDAR 2011 Writer Identification Contest

"The ICDAR 2011 Writer Identification contest", est la troisième compétition organisée dans le cadre de la conférence ICDAR 2011 [LOU 11], elle a été organisée par une équipe de recherche de l'Institut d'informatique et de télécommunications au Centre National de la Recherche

Scientifique "Demokritos", en Grèce. Lors de cette compétition, sept équipes de recherche de différents pays ont soumis leurs systèmes. Un de ces groupes a soumis deux systèmes différents ce qui a fait un total de huit systèmes participants.

A.3.1. Description de la base de données utilisée

La base de données proposée par les organisateurs de cette compétition est constituée de 208 images de documents binarisés avec une résolution de 300 points par pouce. Elle a été construite avec l'aide de 26 scripteurs, où chacun d'eux a recopié 8 pages de textes différents en plusieurs langues (anglais, français, allemand et grec). Toutes ces pages ont été scannées et elles ne contiennent pas d'éléments non textuels (lignes, dessins, etc.) qui peuvent affecter les performances des systèmes participants. La figure A.2 présente quatre images de documents pour chacune des langues.

Ο Σωκράτης διδάσκει ότι η αρετή ταυτίζεται με την σοφία που από αυτήν απορρέουν όλες οι άλλες αρετές, γιατί αυτές είναι το υπέρτατο αγαθό και την αιτιολογεί ως αγαθό που φέρνει την οσιότητα στα οίκαια ευαγέδια, την ομορφιά, τον ήρωα, τη δύναμη, τη σωβριανή αρετή και τα κλέβη των ανθρώπων. Η καρδιά του Σωκράτη στο δικαστήριο μοιάζει πάρα πολύ με αυτή του Χριστού. Ο Σωκράτης στο δικαστήριο όμοια φηροσφιδός δεν ελεησάνθη, δεν έλασθη, δεν κατώργησέ βέ απορογίες ομοια ευσεβέσ απόλυτα διδασκαλία και πράξις. Ο Χριστός ήθελε να θυσιάσει και γ' αυτό στους δικαστές του δεν αστοχούσθησεν ώστε να θανατωθεί ιστορικώς κατόπιν να ανακηθεί αναδευκνόμενος την βελήν υποδοχή του Τέλειο ευσεβέσμεν η ψυχή και με την διδασκαλία του ώστε την επιγήν του θανάτου στον έσπαρό γμασι από τον πατέρα του να συζκαίρησ τους ανθρώπους διότι δεν συμπίθουν τι κένουν με το να τον έσπαρύντου.

(a)

Socrates was a Classical Greek philosopher. Credited as one of the founders of western philosophy, he is an enigmatic figure known only through the classical accounts of his students. Plato's dialogues are the most comprehensive accounts of Socrates to survive from antiquity. Forming an accurate picture of the historical Socrates and his philosophical viewpoints is problematic at best. This issue is known as the Socratic problem. The knowledge of the man, his life, and his philosophy is based on writings by his students and contemporaries. Foremost among them is Plato, however, works by Xenophon, Aristotle, and Aristophanes also provide important insights. The difficulty of finding the real Socrates arises because these works are often philosophical or dramatic texts rather than straightforward histories. Aside from Thucydides who makes no mention of Socrates or philosophers in general, there is in fact no such thing as a straightforward history contemporary with Socrates that dealt with his own time and place.

(b)

Socrate est un philosophe de la Grèce antique, considéré comme le père de la philosophie occidentale et l'un des inventeurs de la philosophie morale. Il n'a laissé aucune œuvre écrite; sa philosophie s'est transmise par l'intermédiaire de témoignages indirects. Socrate naquit en 470, à la fin des guerres médiques, sans doute au mois de mai, près d'Athènes, dans le deme d'Alcôpère, deme qui faisait partie de la tribu d'Antiochide. Son père, Sophronisque, était sculpteur ou tailleur de pierres, et sa mère, Phanarète, sage femme. Socrate avait un frère, Patrôclès, fils du premier mari de sa mère. Peu de choses de sa jeunesse sont connues. Qu'il fût esclave n'est qu'une hypothèse. Il reçut sans doute une éducation classique, que sa loi athénienne obligeait un père à donner à son fils: gymnastique, musique, art du chant, de la danse, apprentissage de la lyre et de la grammaire, ce qui implique l'étude d'Homère, d'Hésiode et d'autres poètes.

(c)

Sokrates war ein für das abendländische Denken grundlegender griechischer Philosoph, der in Athen lebte und wirkte. Seine herausragende Bedeutung zeigt sich u.a. darin, dass alle griechischen Denker vor ihm als Vorsokratiker bezeichnet werden. Sokrates entwickelte die philosophische Methode eines strukturierten Dialogs, die er *Alékein* nannte. Diese Kunst der Gesprächsführung und ihre philosophischen Inhalte sind nur indirekt überliefert worden, da Sokrates selbst nichts Schriftliches hinterlassen hat. Mehrere seiner Schüler, der berühmteste unter ihnen Platon, haben sokratische Dialog verfasst und unterschiedliche Züge seiner Lehre betont. Die unbeugsame Haltung des Sokrates in dem gegen ihn wegen angeblich verderblichen Einflusses auf die Jugend und wegen Missachtung der Griechischen Götter geführten Prozess hat zu seinem Nachruhm wesentlich beigetragen. Das Todesurteil nahm er als gütiges Fehdrittel getassen hin; bis zur Hinrichtung durch den Schierlingbecher beschäftigten ihn und die zu Besuch im Gefängnis weilenden Freunde und Schüler philosophische Fragen.

(d)

Figure A.2. Exemple d'un texte produit en quatre langues : (a) Grec, (b) Anglais, (c) Français, (d) Allemand, de la base de données de la compétition "The ICDAR 2011 Writer Identification contest".

A.3.2. Evaluation et résultats

Afin de mesurer la précision des systèmes soumis à cette compétition, les organisateurs ont proposé l'utilisation de deux critères différents, ces critères sont le "soft TOP-N" ainsi que le "hard TOP-N". Pour chaque image de document de la base de données, sa distance par rapport à toutes les autres images de documents de la base de données est calculée. Ensuite, les résultats sont triés en fonction de cette distance de l'image du document le plus proche au moins proche.

Pour le critère "soft TOP-N", on considère qu'on a une bonne classification si au moins une image de document du même scripteur est comprise dans les N images de documents les plus similaires. Concernant le critère "hard TOP-N", on considère qu'on a une bonne classification si tous les documents les plus similaires sont effectivement écrits par le même scripteur.

Pour toutes les images de documents présentes dans la base de données, on compte ensuite le nombre de bonnes classifications. Le quotient du nombre total de bonnes classifications sur le nombre total d'images dans la base correspond à la précision TOP-N. Les valeurs de N utilisées pour le critère "soft TOP-N" sont : 1, 2, 5, 10. Comme la base est composée de 8 images par scripteurs, la valeur 7 correspond à la valeur maximum de N pour le critère "hard TOP-N".

L'évaluation des systèmes participants est basée sur deux scénarios. Dans le premier scénario, l'image du document toute entière est utilisée. Les résultats d'évaluation de tous les systèmes soumis portant sur la base toute entière sont présentés dans les tableaux A.4 et A.5. Les résultats d'évaluation pour chaque langue prise indépendamment des autres sont présentés aux tableaux A.6, A.7, A.8 et A.9. Dans tous les tableaux, les méthodes qui ont la plus grande précision sont marquées en gras, et le classement de chaque méthode est présenté entre parenthèses. La méthode avec laquelle nous avons participé à cette compétition se nomme TEBESSA. Concernant le premier scénario, la méthode de TSINGHUA a donné les meilleurs résultats et notre méthode a été classée en troisième position.

Système	Pays
ECNU	Chine
QUQA-a	Qatar
QUQA-b	Qatar
TSINGHUA	Chine
GWU	Etats Unis d'Amérique
CS-UMD	Etats Unis d'Amérique
TEBESSA	Algérie
MCS-NUST	Pakistan

Tableau A.3. Systèmes et pays des participants à la compétition "The ICDAR 2011 Writer Identification contest".

Système	TOP-1	TOP-2	TOP-5	TOP-10
ECNU	84,6 (7)	86,5 (6)	88,0 (4)	88,9 (4)
QUQA-a	90,9 (6)	94,2 (5)	98,1 (3)	99,0 (3)
QUQA-b	98,1 (4)	98,6 (3)	99,5 (2)	100 (1)
TSINGHUA	99,5 (1)	99,5 (2)	100 (1)	100 (1)
GWU	93,8 (5)	96,2 (4)	98,1 (3)	99,0 (3)
CS-UMD	99,5 (1)	99,5 (2)	99,5 (2)	99,5 (2)
TEBESSA	98,6 (3)	100 (1)	100 (1)	100 (1)
MCS-NUST	99,0 (2)	99,5 (2)	99,5 (2)	99,5 (2)

Tableau A.4. Résultats d'identification avec le critère "soft TOP-N" en utilisant toute la base de données en %.

Système	TOP-2	TOP-5	TOP-7
ECNU	51,0 (8)	2,9 (8)	0,0 (6)
QUQA-a	76,4 (7)	42,3 (7)	20,2 (5)
QUQA-b	92,3 (4)	77,4 (5)	41,4 (2)
TSINGHUA	95,2 (2)	84,1 (1)	41,4 (2)
GWU	80,3 (6)	44,2 (6)	20,2 (5)
CS-UMD	91,8 (5)	77,9 (4)	22,1 (4)
TEBESSA	97,1 (1)	81,3 (2)	50,0 (1)
MCS-NUST	93,3 (3)	78,9 (3)	38,9 (3)

Tableau A.5. Résultats d'identification avec le critère "hard TOP-N" en utilisant toute la base de données en %.

Système	TOP-1	TOP-2	TOP-5	TOP-10
ECNU	19,2 (8)	19,2 (6)	19,2 (5)	21,2 (5)
QUQA-a	76,9 (7)	86,5 (5)	96,2 (2)	98,1 (2)
QUQA-b	90,4 (4)	90,4 (3)	92,3 (3)	94,2 (4)
TSINGHUA	92,3 (3)	94,2 (2)	98,1 (1)	100 (1)
GWU	80,8 (6)	86,5 (5)	90,4 (4)	94,2 (4)
CS-UMD	96,2 (1)	96,2 (1)	96,2 (2)	96,2 (3)
TEBESSA	84,6 (5)	88,5 (4)	90,4 (4)	94,2 (4)
MCS-NUST	94,2 (2)	94,2 (2)	96,2 (2)	96,2 (3)

Tableau A.6. Résultats d'identification avec le critère "hard TOP-N" en utilisant les documents grecs en %.

Système	TOP-1	TOP-2	TOP-5	TOP-10
ECNU	15,4 (6)	15,4 (5)	15,4 (4)	17,3 (4)
QUQA-a	78,9 (5)	84,6 (4)	96,2 (3)	96,2 (3)
QUQA-b	100 (1)	100 (1)	100 (1)	100 (1)
TSINGHUA	96,2 (3)	96,2 (2)	98,1 (2)	100 (1)
GWU	84,6 (4)	88,5 (3)	96,2 (3)	98,1 (2)
CS-UMD	98,1 (2)	100 (1)	100 (1)	100 (1)
TEBESSA	96,2 (3)	96,2 (2)	98,1 (2)	100 (1)
MCS-NUST	96,2 (3)	96,2 (2)	98,1 (2)	100,0 (1)

Tableau A.7. Résultats d'identification avec le critère "hard TOP-N" en utilisant les documents anglais en %.

Système	TOP-1	TOP-2	TOP-5	TOP-10
ECNU	23,1 (6)	23,1 (5)	23,1 (3)	26,9 (2)
QUQA-a	94,2 (4)	96,2 (3)	96,2 (2)	100 (1)
QUQA-b	98,1 (2)	98,1 (2)	100 (1)	100 (1)
TSINGHUA	96,2 (3)	98,1 (2)	100 (1)	100 (1)
GWU	96,2 (3)	96,2 (3)	100 (1)	100 (1)
CS-UMD	100 (1)	100 (1)	100 (1)	100 (1)
TEBESSA	92,3 (5)	94,2 (4)	100 (1)	100 (1)
MCS-NUST	100 (1)	100 (1)	100 (1)	100 (1)

Tableau A.8. Résultats d'identification avec le critère "hard TOP-N" en utilisant les documents français en %.

Système	TOP-1	TOP-2	TOP-5	TOP-10
ECNU	46,2 (5)	46,2 (5)	46,2 (3)	46,2 (2)
QUQA-a	86,5 (4)	90,4 (4)	98,1 (2)	100 (1)
QUQA-b	100 (1)	100 (1)	100 (1)	100 (1)
TSINGHUA	100 (1)	100 (1)	100 (1)	100 (1)
GWU	92,3 (3)	94,2 (3)	98,1 (2)	100 (1)
CS-UMD	100 (1)	100 (1)	100 (1)	100 (1)
TEBESSA	94,2 (2)	98,1 (2)	100 (1)	100 (1)
MCS-NUST	100 (1)	100 (1)	100 (1)	100 (1)

Tableau A.9. Résultats d'identification avec le critère "hard TOP-N" en utilisant les documents allemands en %.

Système	TOP-1	TOP-2	TOP-5	TOP-10
ECNU	65,9 (7)	71,6 (7)	81,7 (7)	86,5 (7)
QUQA-a	74,0 (4)	81,7 (4)	91,8 (4)	96,2 (3)
QUQA-b	67,3 (5)	79,8 (5)	91,8 (4)	94,7 (5)
TSINGHUA	90,9 (1)	93,8 (1)	98,6 (1)	99,5 (1)
GWU	74,0 (4)	81,7 (4)	91,4 (5)	95,2 (4)
CS-UMD	66,8 (6)	75,5 (6)	83,7 (6)	89,9 (6)
TEBESSA	87,5 (2)	92,8 (2)	97,6 (2)	99,5 (1)
MCS-NUST	82,2 (3)	91,8 (3)	96,6 (3)	97,6 (2)

Tableau A.10. Résultats d'identification avec le critère "soft TOP-N" en utilisant toute la base des images recadrées en %.

Système	TOP-2	TOP-5	TOP-7
ECNU	39,4 (8)	2,9 (8)	0,0 (6)
QUQA-a	52,4 (4)	15,9 (7)	3,4 (5)
QUQA-b	47,6 (7)	22,6 (4)	6,3 (4)
TSINGHUA	79,8 (1)	48,6 (1)	12,5 (2)
GWU	51,4 (6)	20,2 (6)	6,3 (4)
CS-UMD	51,9 (5)	22,1 (5)	3,4 (5)
TEBESSA	76,0 (2)	34,1 (3)	14,4 (1)
MCS-NUST	71,6 (3)	35,6 (2)	11,1 (3)

Tableau A.11. Résultats d'identification avec le critère "hard TOP-N" en utilisant toute la base des images recadrées en %.

Système	TOP-1	TOP-2	TOP-5	TOP-10
ECNU	11,5 (7)	15,4 (8)	19,2 (8)	23,1 (7)
QUQA-a	44,2 (3)	51,9 (5)	73,1 (5)	90,4 (4)
QUQA-b	34,6 (6)	55,8 (4)	76,9 (4)	80,8 (5)
TSINGHUA	51,9 (2)	71,2 (1)	98,1 (1)	98,1 (1)
GWU	42,3 (4)	46,2 (6)	65,4 (7)	76,9 (6)
CS-UMD	40,4 (5)	44,2 (7)	67,3 (6)	76,9 (6)
TEBESSA	42,3 (4)	63,5 (3)	80,8 (3)	92,3 (3)
MCS-NUST	55,8 (1)	69,2 (2)	84,6 (2)	94,2 (2)

Tableau A.12. Résultats d'identification avec le critère "hard TOP-N" en utilisant les documents grecs recadrés en %.

Dans le deuxième scénario, les images des documents sont recadrées en préservant seulement les deux premières lignes, afin de diminuer la quantité d'écriture disponible. Puis, les mêmes expérimentations que dans le premier scénario sont répétées mais cette fois avec les images

recadrées. Les résultats du second scénario sont présentés dans les tableaux A.10, A.11, A.12, A.13, A.14 et A.15.

Système	TOP-1	TOP-2	TOP-5	TOP-10
ECNU	13,5 (8)	15,4 (8)	15,4 (6)	19,2 (5)
QUQA-a	55,8 (5)	67,3 (5)	75,0 (4)	82,7 (4)
QUQA-b	63,5 (4)	69,2 (4)	90,4 (2)	96,2 (2)
TSINGHUA	76,9 (1)	90,4 (1)	96,2 (1)	100 (1)
GWU	50,0 (6)	57,7 (6)	69,2 (5)	82,7 (4)
CS-UMD	44,2 (7)	50,0 (7)	69,2 (5)	82,7 (4)
TEBESSA	69,2 (2)	84,6 (2)	88,5 (3)	100 (1)
MCS-NUST	67,3 (3)	80,8 (3)	88,5 (3)	92,3 (3)

Tableau A.13. Résultats d'identification avec le critère "hard TOP-N" en utilisant les documents anglais recadrés en %.

Système	TOP-1	TOP-2	TOP-5	TOP-10
ECNU	46,2 (8)	46,2 (6)	46,2 (7)	46,2 (6)
QUQA-a	51,9 (6)	67,3 (5)	88,5 (4)	92,3 (3)
QUQA-b	48,1 (7)	67,3 (5)	84,6 (6)	88,5 (5)
TSINGHUA	80,8 (1)	90,4 (1)	96,2 (1)	96,2 (1)
GWU	57,7 (5)	69,2 (4)	86,5 (5)	92,3 (3)
CS-UMD	59,6 (4)	67,3 (5)	84,6 (6)	90,4 (4)
TEBESSA	63,5 (3)	78,9 (3)	90,4 (3)	94,2 (2)
MCS-NUST	65,4 (2)	82,7 (2)	92,3 (2)	96,2 (1)

Tableau A.14. Résultats d'identification avec le critère "hard TOP-N" en utilisant les documents français recadrés en %.

Système	TOP-1	TOP-2	TOP-5	TOP-10
ECNU	21,2 (7)	21,2 (8)	23,1 (7)	26,9 (6)
QUQA-a	71,2 (4)	78,9 (4)	86,5 (5)	94,2 (4)
QUQA-b	44,2 (6)	63,5 (7)	84,6 (6)	92,3 (5)
TSINGHUA	84,6 (1)	90,4 (1)	96,2 (1)	100 (1)
GWU	69,2 (5)	76,9 (5)	88,5 (4)	92,3 (5)
CS-UMD	73,1 (3)	80,8 (3)	90,4 (3)	96,2 (3)
TEBESSA	71,2 (4)	75,0 (6)	88,5 (4)	98,1 (2)
MCS-NUST	78,9 (2)	88,5 (2)	94,2 (2)	98,1 (2)

Tableau A.15. Résultats d'identification avec le critère "hard TOP-N" en utilisant les documents allemands recadrés en %.

Il est à noter que notre système été classé au bilan final comme troisième juste derrière les systèmes TSINGHUA et NCS-NUST.

A.4. ICFHR2012 Competition on Writer Identification Challenge 1: Latin/Greek Documents

L'objectif essentiel de la compétition "ICFHR2012 Competition on Writer Identification - Challenge 1: Latin/Greek Documents" [LOU 12] est d'évaluer sur des bases de test communes, et avec des outils d'évaluation standardisés les performances des systèmes de reconnaissance de scripteurs proposés par les participants sur des échantillons d'écriture de petite taille

(contenant peu d'informations). La compétition a attiré quatre équipes de recherche. L'une de ces équipes a soumis trois méthodes différentes, tandis qu'une autre en a présenté deux.

A.4.1. Description de la base de données utilisée

La base de données utilisée dans le cadre de cette compétition a été créée avec l'aide de 126 scripteurs à qui il a été demandé de copier quatre parties de texte en deux langues différentes (anglais et grec). Ces parties de texte étaient les mêmes pour tous les scripteurs. Parmi tous les documents, seuls les documents grecs ont été écrits en langue maternelle des scripteurs. Il est important aussi de noter que le nombre de lignes de texte qui ont été produites par les scripteurs se situe entre deux et cinq et que toutes les images de documents ont été binarisées et ne contiennent pas d'éléments non textuels (lignes, dessins, etc.).

A.4.2. Evaluation et résultats

Afin de mesurer la précision des systèmes soumis, les organisateurs ont utilisés les critères "soft TOP-N" et "hard TOP-N" présentés dans la section A.3.2.

Pour toutes les images de documents présentes dans la base, on compte le nombre de bonnes classifications. Le quotient du nombre total de bonnes classifications sur le nombre total d'images dans la base correspond à la précision TOP-N. Les valeurs de N utilisées pour le critère "soft TOP-N" sont : 1, 2, 5, 10. Comme la base est composée de 4 images par scripteurs, la valeur 3 correspond à la valeur maximum de N pour le critère "hard TOP-N".

Les résultats d'évaluation de toutes les méthodes portant sur la base entière sont présentés dans les tableaux A.16 et A.17. Les résultats d'évaluation pour chaque langue prise indépendamment des autres sont présentés aux tableaux A.18 (Anglais) et A.19 (Grec). En ce qui concerne les expérimentations dépendantes de la langue, seul le critère "soft TOP-N" est acceptable puisque deux documents uniquement sont disponibles par scripteur et l'un d'eux est utilisé comme requête.

Il est important de noter que les 104 images de documents issues des 26 premiers scripteurs ont été utilisées par les participants afin d'ajuster les paramètres de leurs systèmes, alors que les 400 images de documents restantes et qui sont issues des 100 scripteurs restants ont été utilisées par les organisateurs de la compétition afin d'évaluer les systèmes participants.

Dans tous les tableaux, les systèmes qui ont la plus grande précision sont marqués en gras, et le classement de chaque système est présenté entre parenthèses. Il est à noter que lors de cette compétition, nous avons soumis trois systèmes différents, ces systèmes sont: *TEBESSA-a*, *TEBESSA-b* et *TEBESSA-c*. L'un de nos systèmes *TEBESSA-c* a enregistré les meilleurs taux d'identification.

Système	TOP-1	TOP-2	TOP-5	TOP-10
TEBESSA-a	92,3 (3)	96,5 (2)	98,8 (2)	99,0 (2)
TEBESSA-b	89,8 (4)	94,3 (4)	97,8 (3)	98,8 (3)
TEBESSA-c	94,5 (1)	97,3 (1)	99,3 (1)	99,3 (1)
QATAR-a	70,3 (7)	80,8 (7)	91,8 (6)	95,3 (7)
QATAR-b	80,0 (6)	87,3 (6)	95,0 (5)	98,0 (5)
TSINGHUA	92,8 (2)	95,8 (3)	97,8 (3)	98,3 (4)
HANNOVER	85,5 (5)	90,3 (5)	95,3 (4)	97,3 (6)

Tableau A.16. Résultats d'identification avec le critère "soft TOP-N" en utilisant toute la base de données en %.

Système	TOP-2	TOP-3
TEBESSA-a	57,5 (2)	38,0 (1)
TEBESSA-b	57,5 (2)	29,3 (3)
TEBESSA-c	65,0 (1)	37,8 (2)
QATAR-a	32,3 (6)	11,3 (7)
QATAR-b	34,0 (5)	15,3 (6)
TSINGHUA	51,5 (3)	27,3 (4)
HANNOVER	41,5 (4)	22,8 (5)

Tableau A.17. Résultats d'identification avec le critère "hard TOP-N" en utilisant toute la base de données en %.

Système	TOP-1	TOP-2	TOP-5	TOP-10
TEBESSA-a	89,5 (3)	96,0 (1)	97,0 (2)	98,5 (1)
TEBESSA-b	83,0 (4)	90,0 (4)	96,0 (3)	97,0 (3)
TEBESSA-c	91,5 (2)	95,5 (2)	97,5 (1)	98,0 (2)
QATAR-a	53,5 (7)	66,5 (7)	85,0 (7)	90,0 (6)
QATAR-b	72,5 (6)	82,5 (6)	92,5 (5)	96,5 (4)
TSINGHUA	94,0 (1)	94,5 (3)	95,5 (4)	98,0 (2)
HANNOVER	82,0 (5)	88,0 (5)	91,5 (6)	95,0 (5)

Tableau A.18. Résultats d'identification avec le critère "soft TOP-N" en utilisant les documents anglais en %.

Système	TOP-1	TOP-2	TOP-5	TOP-10
TEBESSA-a	92,0 (2)	95,0 (2)	98,5 (2)	99,0 (2)
TEBESSA-b	85,5 (5)	93,5 (4)	95,5 (4)	99,0 (2)
TEBESSA-c	93,5 (1)	97,0 (1)	99,5 (1)	99,5 (1)
QATAR-a	76,0 (6)	86,0 (7)	94,5 (5)	96,5 (4)
QATAR-b	85,5 (5)	90,0 (6)	96,0 (3)	98,5 (3)
TSINGHUA	90,0 (3)	94,0 (3)	98,5 (2)	99,0 (2)
HANNOVER	87,5 (4)	93,0 (5)	98,5 (2)	99,5 (1)

Tableau A.19. Résultats d'identification avec le critère "soft TOP-N" en utilisant les documents grecs en %.

A.5. ICFHR2012 Competition on Writer Identification - Challenge 2: Arabic Scripts

La compétition "ICFHR2012 Competition on Writer Identification - Challenge 2: Arabic Scripts" [HAS 12] a été organisée sur la plateforme Kaggle. Son but était de réunir les chercheurs afin de comparer les progrès récents dans le domaine de l'identification de scripteurs arabes. Elle a attiré quarante-trois équipes de recherches industrielles et universitaires.

A.5.1. Description de la base de données utilisée

Dans cette compétition, 206 individus ont été invités à écrire trois paragraphes différents en arabe. Les deux premiers paragraphes ont été utilisés pour l'entraînement et le troisième a été utilisé pour les tests. Pour certains scripteurs, les deux premiers paragraphes ont été retirés de l'ensemble d'apprentissage afin de tester la capacité des systèmes participants à détecter les scripteurs inconnus. Les images ont été acquises à l'aide d'un scanner "EPSON GT-S80", avec une résolution 600 points par pouce. Les images ont été fournies en format binaire PNG sachant que la binarisation a été effectuée en utilisant la méthode d'Otsu.

A.5.2. Evaluation et résultats

Les organisateurs de la compétition ont demandé aux participants de produire, pour chaque image de paragraphe de l'ensemble de test, le numéro (ID) du scripteur le plus probable (parmi les scripteurs de l'ensemble d'apprentissage), ou zéro quand il est plus probable que le scripteur soit inconnu. Les méthodes participantes ont été classées en fonction de leur taux d'identification. Le tableau A.20 présente les résultats enregistrés par certains systèmes parmi ceux qui ont été soumis à cette compétition. Le système qui a la plus grande précision est marqué en gras, et le classement de chaque système est présenté entre parenthèses. Il est à noter que lors de cette compétition, notre système appelé AWReS a été classé en septième position avec un taux d'identification de l'ordre de 91.28% et que nous n'avons pas considéré le rejet des scripteurs inconnus.

Système	Pays	TOP-1
Wayne Zhang	Chine	95,30 (1)
Newell and Griffin	Royaume-Uni	95,30 (2)
YT	Australie	93,29 (3)
BFS	Chine	91,95 (4)
Ihata	Turquie	91,95 (5)
William Cukierski	Etats-Unis d'Amérique	91,28 (6)
AWReS	Algérie	91,28 (7)
Cess Northumbria	Royaume-Uni	91,28 (8)
Marcos Sainz	Etats-Unis d'Amérique	89,93 (9)
Sashi Dareddy	Royaume-Uni	89,26 (10)
D33B	Inconnu	87,92 (16)
Han & Kilian	Etats-Unis d'Amérique	83,89 (23)
Foxtrot	Pologne	82,25 (25)
Ben Hamner	Etats-Unis d'Amérique	81,88 (27)
Steinke	Inconnu	81,21 (33)
Luciferase	Afrique du sud	77,18 (37)
EBDF 16	Qatar	62,42 (40)
EBDF 8	Qatar	59,73 (43)
EBDF 12	Qatar	57,72 (44)
EBDF 4	Qatar	27,52 (45)

Tableau A.20. Résultats d'identification de certains systèmes parmi ceux qui ont participé à la compétition "ICFHR2012 Competition on Writer Identification - Challenge 2: Arabic Scripts " en %.

A.6. ICDAR 2013 Competition on Writer Identification

La compétition "ICDAR 2013 Competition on Writer Identification" [LOU 13] est la troisième édition d'une série de compétitions organisées par une équipe de recherche de l'Institut d'informatique et de télécommunications au Centre National de la Recherche Scientifique "Demokritos", en Grèce. (voir sections A.3 et A.4). Elle a été organisée pour enregistrer les avancées dans le domaine de l'identification de scripteurs à partir de documents latins et grecs. Cette compétition a attiré l'attention de six équipes de recherches qui ont soumis douze méthodes différentes.

A.6.1. Description de la base de données utilisée

La base de données utilisée dans le cadre de cette compétition a été créée avec l'aide de 250 scripteurs à qui il a été demandé de copier quatre parties de texte en deux langues différentes

(anglais et grec). Ces parties de texte étaient les mêmes pour tous les scripteurs. Parmi tous les documents, seuls les documents grecs ont été écrits en langue maternelle des scripteurs.

A.6.2. Evaluation et résultats

Les organisateurs de la compétition ont utilisé les mêmes critères d'évaluation qui ont été employés lors de la compétition "ICFHR2012 Competition on Writer Identification - Challenge 1: Latin/Greek Documents" (voir section A.4.2).

Les résultats d'évaluation de toutes les méthodes portant sur la base toute entière sont présentés dans les tableaux A.21 et A.22. Les résultats d'évaluation pour chaque langue prise indépendamment des autres sont présentés aux tableaux A.23 (Grec) et A.24 (Anglais).

Système	TOP-1	TOP-2	TOP-5	TOP-10
CS-UMD-a	95,1 (1)	97,7 (1)	98,6 (1)	99,1 (2)
CS-UMD-b	95,0 (2)	97,2 (2)	98,6 (1)	99,2 (1)
CS-UMD-c	85,5 (10)	90,9 (10)	95,0 (7)	96,8 (8)
CVL-IPK	90,9 (6)	93,6 (7)	97,0 (4)	98,0 (5)
HANNOVER-a	86,9 (9)	91,9 (9)	95,4 (6)	97,0 (7)
HANNOVER-b	91,5 (5)	94,2 (6)	97,0 (4)	98,0 (5)
HIT-ICG	94,8 (3)	96,7 (3)	98,0 (2)	98,3 (4)
QATAR-a	54,8 (12)	67,3 (12)	80,8 (9)	88,3 (10)
QATAR-b	78,4 (11)	85,8 (11)	91,5 (8)	95,1 (9)
TEBESSA-a	90,3 (7)	94,4 (5)	96,7 (5)	98,3 (4)
TEBESSA-b	90,1 (8)	93,4 (8)	97,0 (4)	97,9 (6)
TEBESSA-c	93,4 (4)	96,1 (4)	97,8 (3)	98,5 (3)

Tableau A.21. Résultats d'identification avec le critère "soft TOP-N" en utilisant toute la base de données en %.

Système	TOP-2	TOP-3
CS-UMD-a	19,6 (11)	7,1 (9)
CS-UMD-b	20,2 (10)	8,4 (8)
CS-UMD-c	21,2 (9)	5,7 (10)
CVL-IPK	44,8 (7)	24,5 (6)
HANNOVER-a	50,0 (6)	26,1 (5)
HANNOVER-b	54,3 (5)	27,3 (4)
HIT-ICG	63,2 (1)	36,5 (1)
QATAR-a	11,8 (12)	3,9 (11)
QATAR-b	34,6 (8)	16,5 (7)
TEBESSA-a	58,2 (3)	33,2 (2)
TEBESSA-b	55,5 (4)	29,5 (3)
TEBESSA-c	62,6 (2)	36,5 (1)

Tableau A.22. Résultats d'identification avec le critère "hard TOP-N" en utilisant toute la base de données en %.

Il est important de noter que les 400 images de documents issues des 100 premiers scripteurs ont été utilisées par les participants afin d'ajuster les paramètres de leurs systèmes, alors que les 600 images de documents restantes et qui sont issues des 150 scripteurs restants ont été utilisées par les organisateurs de la compétition afin d'évaluer les systèmes participants.

Dans tous les tableaux, les systèmes qui ont la plus grande précision sont marqués en gras, et le classement de chaque système est présenté entre parenthèses. Il est à noter que lors de cette compétition, nous avons soumis trois systèmes différents, ces systèmes sont: *TEBESSA-a*, *TEBESSA-b* et *TEBESSA-c*.

Système	TOP-1	TOP-2	TOP-5	TOP-10
CS-UMD-a	95,6 (1)	98,2 (1)	98,6 (2)	99,2 (1)
CS-UMD-b	95,2 (2)	97,6 (2)	98,8 (1)	99,0 (2)
CS-UMD-c	86,0 (10)	90,6 (9)	94,6 (9)	96,4 (6)
CVL-IPK	88,4 (7)	92,0 (7)	96,8 (5)	97,8 (4)
HANNOVER-a	86,4 (9)	91,2 (8)	95,2 (8)	97,4 (5)
HANNOVER-b	90,2 (6)	92,8 (6)	95,6 (7)	97,4 (5)
HIT-ICG	93,8 (3)	96,4 (3)	97,2 (4)	97,8 (4)
QATAR-a	58,8 (12)	66,6 (11)	79,6 (11)	85,8 (8)
QATAR-b	78,8 (11)	84,6 (10)	91,2 (10)	94,4 (7)
TEBESSA-a	91,0 (5)	94,0 (5)	96,8 (5)	97,8 (4)
TEBESSA-b	87,2 (8)	92,0 (7)	96,4 (6)	97,8 (4)
TEBESSA-c	92,6 (4)	96,0 (4)	98,0 (3)	98,4 (3)

Tableau A.23. Résultats d'identification avec le critère "soft TOP-N" en utilisant les documents grecs en %.

Système	TOP-1	TOP-2	TOP-5	TOP-10
CS-UMD-a	94,6 (1)	97,0 (1)	98,4 (1)	98,8 (2)
CS-UMD-b	94,4 (2)	96,6 (2)	98,4 (1)	99,0 (1)
CS-UMD-c	86,4 (7)	90,4 (9)	93,2 (8)	96,0 (7)
CVL-IPK	91,4 (4)	94,2 (4)	95,8 (4)	97,2 (3)
HANNOVER-a	84,6 (10)	88,6 (10)	92,0 (9)	94,0 (9)
HANNOVER-b	85,6 (9)	90,6 (8)	93,6 (7)	95,6 (8)
HIT-ICG	92,2 (3)	94,6 (3)	96,4 (2)	96,8 (4)
QATAR-a	50,0 (12)	64,2 (12)	78,0 (11)	85,4 (11)
QATAR-b	75,8 (11)	84,6 (11)	90,4 (10)	93,6 (10)
TEBESSA-a	86,0 (8)	91,6 (6)	94,4 (6)	96,0 (7)
TEBESSA-b	88,2 (6)	90,8 (7)	94,6 (5)	96,2 (6)
TEBESSA-c	91,2 (5)	93,4 (5)	96,2 (3)	96,6 (5)

Tableau A.24. Résultats d'identification avec le critère "soft TOP-N" en utilisant les documents anglais en %.

A.7. ICDAR2013 Competitions on Signature Verification and Writer Identification for On- and Offline Skilled Forgeries

La compétition "ICDAR2013 Competitions on Signature Verification and Writer Identification for On- and Offline Skilled Forgeries (SigWiComp2013)" [MAL 13] a été organisée par des chercheurs appartenant à plusieurs institutions scientifiques internationales. Trois tâches différentes ont été considérées lors de cette compétition : la vérification hors ligne de signatures manuscrites, la vérification en ligne de signatures manuscrites ainsi que l'identification de scripteurs à partir des écritures de styles différents issues du même scripteur. Cette section ne sera réservée qu'à la description des résultats de la troisième tâche (identification de scripteurs). Lors de cette compétition, les organisateurs ont reçu 8 systèmes de trois institutions différentes pour la tâche d'identification de scripteurs.

A.7.1. Description de la base de données utilisée

Les 330 échantillons d'écriture utilisés lors de cette compétition ont été recueillis auprès 55 individus parmi les employés de l'IFN (Netherlands Forensic Institute) ainsi qu'auprès de certains de leurs proches. Chacun des scripteurs a écrit six différents fragments de texte dans

différents styles d'écriture (cursif, majuscule et minuscule). L'ensemble d'apprentissage est composé de 165 fragments de texte issus des 55 scripteurs et écrits dans les 3 styles d'écriture différents. L'ensemble de test aussi est composée de 165 fragments de texte issus des 55 scripteurs et écrits dans 3 styles différents (autre que ceux disponibles dans l'ensemble d'apprentissage).

A.7.2. Evaluation et résultats

Les résultats de l'évaluation de la tâche d'identification de scripteurs en reliant différents styles d'écriture du même scripteur sont présentés dans le tableau A.25. Dans cette tâche, les organisateurs de la compétition ont utilisé des mesures standards telles que la précision et le rappel pour évaluer les systèmes participants et le classement final est basé sur la valeur de F-mesure (en termes de pourcentages). Il est à noter que nous avons soumis cinq systèmes différents lors de cette compétition et l'un de nos systèmes a été classé en première position (marqué en gras).

Système	Pays	F-mesure (%)
14	Qatar	7,87
15	Qatar	21,81
16	Algérie	17,57
17	Algérie	27,87
18	Algérie	26,06
19	Algérie	28,48
20	Algérie	27,27
21	Suisse	19,39

Tableau A.25 - Performances des systèmes participants à la tâche "Writer Identification" de la compétition "ICDAR2013 Competitions on Signature Verification and Writer Identification for On- and Offline Skilled Forgeries".

A PROPOS DE L'AUTEUR

BIOGRAPHIE DE L'AUTEUR



Chawki DJEDDI, est né en 1977 à Tébessa (nord-est algérien), où il a poursuivi ses études jusqu'à l'obtention du baccalauréat (option sciences de la nature et de la vie) en 1995. Il rejoint l'université d'Annaba la même année pour suivre une formation d'ingénieur d'état en Informatique, option matériel et logiciel. Il obtient son ingéniorat en 2000, puis accède, à partir de 2006, aux études de post-graduation pour préparer un mémoire de Magister en informatique (sur l'identification de scripteurs arabes par une méthode locale) qu'il soutient en 2009, avec mention très bien. Depuis cette date, il devient maître-assistant B, au département de mathématiques et d'informatique de l'université de Tébessa, et maître-assistant A à partir de 2012. Il a participé à l'organisation de plusieurs conférences nationales et internationales, qui ont eu lieu à l'université de Tébessa. Il a collaboré, en qualité de membre, à deux projets de recherche CNEPRU: Classification et Apprentissage dans les Systèmes Complexes (CLASYCO-1) et Intégration de paradigmes pour le développement des systèmes complexes (COSYPI-1). Actuellement, il est membre du Laboratoire de Mathématiques, d'Informatique et des Systèmes (LAMIS) de l'université de Tébessa.

Dans le cadre de ses activités scientifiques Chawki DJEDDI a participé à plusieurs conférences et compétitions spécialisées dans le domaine de l'analyse de l'écriture manuscrite et de l'identification du scripteur où ses propositions ont été très bien classées. Il est membre de plusieurs associations scientifiques, telles que : Groupe de Recherche en Communication Ecrite, Association internationale de la reconnaissance de formes (IAPR), Association Française pour la Reconnaissance et l'Interprétation des Formes et l'association internationale de l'informatique et des technologies de l'information (IACSIT).

1. Publications dans des journaux internationaux

- 1 Djeddi C., Siddiqi I., Souici-Meslati L., Ennaji A., "Text-Independent Writer Recognition Using Multi-script Handwritten Texts ", **Pattern Recognition Letters**, Vol. 34, No. 10, July 2013.
-

Informations sur le journal PRL

ISSN: 0167-8655.

Fréquence de publication: 16 numéros par an.

Processus de reviewing: peer reviewed.

Publié par: North Holland, an imprint of Elsevier, on behalf of the International Association of Pattern Recognition

Site Web: <http://www.journals.elsevier.com/pattern-recognition-letters/>

Indexé par : ACM Computing Reviews, Cambridge Scientific Abstracts, CompuScience, Computer Abstracts, Current Contents/Engineering, Computing & Technology, Engineering Index, Geographical Abstracts, INSPEC Information Services, SCISEARCH, Science Citation Index, Scopus and Zentralblatt MATH.

2012 impact factor: 1,266

5-year impact factor: 1.529 (JCR 2012)

- 2 Siddiqi I., Djeddi C., Raza A., Souici-Meslati L., "Automatic analysis of handwriting for gender classification", **Soumis pour publication dans un journal international spécialisé, en phase de révision**
-

2. Articles parus dans des Lectures Notes

1. Djeddi C., Souici-Meslati L., Ennaji A., "Writer Recognition on Arabic Handwritten Documents", Proceedings of the 5th International Conference on Image and Signal Processing, **ICISP'2012**, Agadir, Morocco, published in Lecture Notes in Computer Science, **LNCS 7340**, pp. 493-501, Springer, June 2012, ISBN: 978-3-642-31254-0.

3. Articles parus dans des journaux nationaux

1. Djeddi C., Souici-Meslati L., "Identification de scripteurs pour l'écriture arabe par une approche locale", RIST, **Revue de l'Information Scientifique et Technique** (publication du Centre de Recherche sur l'Information Scientifique et Technique CERIST, Algérie), Vo. 18, No. 1, 2010, ISSN : 1111-0015.

4. Communications avec comité de lecture international

1. Djeddi C., Souici-Meslati L., "Identification de scripteurs pour l'écriture arabe par une approche locale", MCSEAI'08, In Proceeding of the Maghrebian Conference on Software Engineering and Artificial Intelligence, **MCSEAI'2008**, Oran, Algérie, Avril 2008, pp. 565-570.
2. Djeddi C., Souici-Meslati L., "Une approche locale en mode indépendant du texte pour l'identification de scripteurs : Application à l'écriture arabe". In Proceeding of the Colloque International Francophone sur l'Ecrit et le Document, **CIFED'2008**, Rouen, France, Octobre 2008, pp. 151-156.
3. Djeddi C., Souici-Meslati L., "Combinaison de classifieurs de mots arabes pour la vérification du scripteur", In Proceeding of the Workshop ECOI (Extraction de COnnaisances et Image), Conférence EGC (Extraction et Gestion de Connaissances), Strasbourg, France, Janvier 2009, pp. A2 15- A2 25.
4. Djeddi C., Souici-Meslati L., "Vérification de scripteur pour l'écriture arabe par fusion de décision", In Proceeding of the 5th IEEE International Conference: Sciences of Electronic, Technologies of Information and Telecommunications, **SETIT'2009**, , Tunisia, 2009.
5. Djeddi C., Souici-Meslati L., "Fusion de Décisions pour l'Identification de Scripteurs", In Proceeding of the International Conference on Information Systems and Technologies, **ICIST'2011**, Tébessa, Algérie, Avril 2011, pp : 146-151.
6. Djeddi C., Souici-Meslati L., "A texture based approach for Arabic Writer Identification and Verification", In Proceeding of the IEEE International Conference on Machine and Web Intelligence, **ICMWI'2010**, Algiers, Algeria, Octobre 2010, pp : 115 – 120.
7. Djeddi C., Souici-Meslati L., "Artificial Immune Recognition System for Arabic Writer Identification", In Proceeding of the 4th IEEE International Symposium on Innovation in Information & Communication Technology, **ISIICT'2011**, 2011 , Jordan , Amman, pp : 159-165.
8. Djeddi C., Siddiqi I., Souici-Meslati L., Ennaji A., "Multi Script Writer Identification Optimized With Retrieval Mechanism", In Proceeding of the International Conference on Frontiers in Handwriting Recognition, **ICFHR'2012**, Bari, Italy, September 2012, pp: 507 - 512.
9. Djeddi C., Siddiqi I., Souici-Meslati L., Ennaji A., "Codebook for Writer Characterization: A Vocabulary of Patterns or a Mere Representation Space?", In Proceeding of the 12th International Conference on Document Analysis and Recognition, **ICDAR'2013**, Washington, USA, August 2013, pp : 423 - 427.
10. Siddiqi I, Raza A and Djeddi C and Ennaji A., "Multilingual Artificial Text Detection Using a Cascade of Transforms", In Proceeding of the 12th International Conference on Document Analysis and Recognition, **ICDAR'2013**, Washington, USA, August 2013, pp : 309 - 313.
11. Djeddi C., Souici-Meslati L., Siddiqi I., Ennaji A., El Abed H., Gattal A., "Evaluation of Texture Features for Offline Arabic Writer Identification", To appear in the proceeding of the 11th IAPR International Workshop on Document Analysis Systems, **DAS'2014**, Tours, France, April 2014.

PARTICIPATION AUX COMPETITIONS SPECIALISEES

2011

1. Troisième place pour la compétition "The ICDAR 2011 - Writer Identification Contest".
2. Deuxième place pour la tâche identification de scripteur à la compétition "ICDAR/GREC 2011 Music Scores: Writer Identification and Staff Removal competition".
3. Septième place (sur plus de 30 systèmes) pour la compétition "The ICDAR 2011 - Arabic Writer Identification Contest".

2012

1. Sixième place (sur plus de 49 systèmes) pour la compétition "The ICFHR 2012 - Writer Identification Contest for Arabic Scripts".
2. Première place pour la compétition "The ICFHR 2012 - Writer Identification Contest Challenge 1: Latin/Greek Documents".

2013

1. Première place pour la tâche "Writer Identification" à la compétition "ICDAR 2013 Signature Verification and Writer Identification Competitions for On- and Offline Skilled Forgeries".
2. Quatrième place pour la tâche "Dutch Offline Signature Verification" à la compétition "ICDAR 2013 Signature Verification and Writer Identification Competitions for On- and Offline Skilled Forgeries".
3. Quatrième place pour la tâche "Japanese Offline Signature Verification Task" à la compétition "ICDAR 2013 Signature Verification and Writer Identification Competitions for On- and Offline Skilled Forgeries".
4. Quatrième place pour la compétition "The ICDAR 2013 Competition on Writer Identification".

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 des individus et de leur genre à partir de leur écriture manuscrite. Deux contributions se dégagent de cette étude, dans la première contribution, nous proposons une méthode indépendante du texte pour la reconnaissance du scripteur dans un environnement multi-scripts. L'objectif est de reconnaître l'auteur d'un texte manuscrit dans un script à partir d'échantillons du même auteur dans un autre script et donc valider l'hypothèse que le style d'écriture d'un individu reste constant à travers différents scripts. La méthode proposée est basée sur des distributions de longueurs de segments qui sont comparées avec les méthodes les plus connues et les plus performantes de l'état de l'art. La classification est réalisée en utilisant les K plus proches voisins (k-PPV) ainsi que les séparateurs à vaste marge (SVM). Les résultats expérimentaux obtenus sur une base de données de 126 scripteurs avec 4 échantillons par scripteur montrent que la méthode proposée permet d'obtenir des performances intéressantes. Notre deuxième contribution consiste à présenter une étude pour la détermination du sexe des individus à partir de leurs écritures manuscrites. La méthode proposée est basée sur l'extraction d'un ensemble de caractéristiques de l'écriture à partir d'échantillons de scripteurs de sexe masculin et féminin et l'entraînement d'un classifieur afin qu'il puisse distinguer entre les deux catégories. Des attributs de l'écriture comme l'orientation, la courbure, la texture et la lisibilité sont estimés en calculant des caractéristiques locales et globales. La classification est effectuée à l'aide des réseaux de neurones artificiels (ANN) ainsi que les séparateurs à vaste marge (SVM). La méthode proposée a été évaluée en utilisant deux bases de données sous un certain nombre de scénarios où des résultats intéressants ont été enregistrés.

Keywords : Identification de scripteurs, Détermination du sexe, Ecriture manuscrite multi-scripts.

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. Two contributions emerge from this study. In the first one, we propose a text-independent writer recognition method in a multi-script environment. The objective is to recognize the writer of a handwritten text in one script from the samples of the same writer in another script and hence validate the hypothesis that writing style of an individual remains constant across different scripts. The proposed method is based on a set of run-length features which are compared with the well-known state-of-the-art features. Classification is carried out using K-Nearest Neighbors (K-NN) and Support Vector Machines (SVM). The experimental results obtained on a database of 126 writers with 4 samples per writer show that the proposed scheme achieves interesting performances. Our second contribution consists to present 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. Writing attributes like slant, curvature, texture and legibility are estimated by computing local and global features. Classification is carried out using Artificial Neural Networks and Support Vector Machine. The proposed technique evaluated on two databases under a number of scenarios realized interesting results.

Keywords: Writer recognition, gender determination, multi-scripts handwriting.

ان العمل المقدم في هذا المخطوط يندرج في اطار مجال التحليل و التعرف على الوثائق، و على نحو أدق، تحديد هوية و جنس صاحب الخط في الوثائق المكتوبة بخط اليد. مساهمتان اثنتان تبتقان عن هذه الدراسة، في المساهمة الأولى، نقترح طريقة للتعرف على هوية صاحب الخط في بيئة متعددة الأبجديات. الهدف هو التعرف على هوية صاحب الخط في وثائق مكتوبة بأبجدية من خلال عينات لنفس الشخص مكتوبة بأبجدية أخرى. وبالتالي التحقق من صحة الفرضية القائلة بأن أسلوب كتابة أي شخص يبقى ثابتا عبر أبجديات مختلفة. ان الطريقة المقترحة مبنية على ميزات توزيع قياسات أطوال المسارات التي قورنت مع ميزات أخرى جد معروفة. تمت عملية التصنيف باستخدام مصنفات الجار الأقرب و أجهزة المتجهات الاعتمادية. النتائج التجريبية التي تم الحصول عليها باستخدام قاعدة بيانات تحتوي على وثائق ل 126 شخصا مع 4 عينات لكل شخص تبين أن المخطط المقترح يحقق نتائج مثيرة للاهتمام. أما فيما يخص مساهمتنا الثانية، فهي تتمثل في تقديم دراسة حول تحديد جنس الأشخاص من خلال وثائقهم المكتوبة بخط اليد. وتستند الطريقة المقترحة على استخراج مجموعة من الميزات من عينات لوثائق لذكور و إناث و تدريب المصنفات لتكون قادرة على التمييز بين الفئتين. يتم تقدير صفات الكتابة مثل الميل، الانحناء، الملمس و الوضوح من خلال حساب ميزات خاصة و عامة. تمت عملية التصنيف باستخدام مصنفات الشبكات العصبية الاصطناعية و أجهزة المتجهات الاعتمادية. تم تقييم الطريقة المقترحة باستخدام قاعدتي بيانات في إطار عدد من السيناريوهات المختلفة حيث حققت نتائج مثيرة للاهتمام.

الملخص

الكلمات المفتاحية : تحديد هوية صاحب الخط، تحديد جنس صاحب الخط، كتابة بخط اليد متعددة الأبجديات.