République Algérienne Démocratique et Populaire
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique
Université Larbi Tébessi –Tébessa
Faculté des Sciences Exactes et des Sciences
de la Nature et de la Vie
Département de Mathématiques et d'Informatique

Mémoire de master

Domaine : Mathématiques et Informatique
Filière : Informatique
Option : Systèmes et Multimédia

Thème :

Une contribution pour la reconnaissance de scripts à partir de documents manuscrits et imprimés

Présenté par :

Guelai Zaineb

Devant le jury :

Mr. Menassel Rafik	*Maitre de conférences A*	*Université de Tébessa*	*Président*
Mr. Aouine Mohamed	*Maitre-Assistant A*	*Université de Tébessa*	*Examinateur*
Mr. Chawki Djeddi	*Maitre de conférences A*	*Université de Tébessa*	*Rapporteur*

Année Universitaire : 2020-2021

Résumé

Dans ce manuscrit, un système automatique pour l'identification de scripts à partir des images de documents est introduit. Le système proposé comprend deux étapes principales: l'extraction de caractéristiques et la classification (identification de script). Dans la première étape, des caractéristiques texturales basées sur les OBIFs sont extraites à partir des images de documents manuscrits et imprimés. Dans la deuxième étape, nous avons utilisé les machines à vecteurs de support pour la classification. Les expérimentations sont menées sur un ensemble de données qui comprend 10400 images de documents manuscrits et imprimés. Des résultats encourageants ont été enregistrés.

Keywords : Identification de script, documents manuscrits, documents imprimés, machines à vecteurs de support, OBIFs.

Abstract

In this manuscript, a script identification system based on printed and handwritten documents is introduced. The proposed system consists of two main stages: feature extraction and classification (script identification). In the first step, textural features are extracted from printed and handwritten documents. In the second step, we have used Support Vector Machines as classifier. The experiments are conducted on a database that includes 10400 printed and handwritten documents. Encouraging results have been recorded.

Keywords: Script identification, handwritten documents, printed documents, support vector machines, OBIFs.

الملـخص

في هذا العمل، تم تقديم نظام للتعرف على نوع الكتابة باستعمال وثائق مطبوعة و مكتوبة بخط اليد. يتكون النظام المقترح من مرحلتين أساسيتين: استخلاص المميزات وتصنيفها (التعرف على نوع الكتابة). في الخطوة الأولى يتم استخلاص مميزات الملمس من الوثائق المطبوعة و المكتوبة بخط اليد. خلال المرحلة الثانية استخدمنا آلة المتجه الداعم للتصنيف. أجريت التجارب على قاعدة بيانات تحتوي على 10400 وثيقة مطبوعة و مكتوبة بحط اليد. تم تسجيل نتائج مشجعة..

الكلمات المفتاحية : التعرف على نوع الكتابة، وثائق مكتوبة بخط، وثائق مطبوعة، آلة المتجه الداعم، مميزات الملمس.

Dédicace

*Au nom d'**Allah**, le tout Miséricordieux, celui qui fait Miséricorde.*

*Louange à **Allah** de m'a donné la force et le courage pour réaliser ce travail.*

*Au terme de ce travail, je profite de cette occasion pour adresser mes vifs remerciements à mon encadreur **Mr. DjeddiChawki** pour son aide et sa disponibilité.*

*De même je dédie ce modeste travail aux plus belles perles : Ma Mère « **Ouarda** » et Mon Père « **Senani** » qui m'ont toujours encouragé pour réussir à mes études, me donnent la volonté pour affranchir tous les obstacles.*

*A mes sœur : **ELchifa**et **Amira** à qui je souhaite la réussite dans leurs vies.*

*A mes frère **Ali** et**AbdELhak** à qui je souhaite un meilleur avenir.*

A mes grandes parents et ma grande famille

*A Toutes mes Famille : «**Guelai »** et « **Guenez »***

A mes amies et à toute la promotion (2021)

A toutes les personnes qui ont une place spéciale dans ma vie.

GUELAI ZAINEB

Remerciements

Mes remerciements vont à tous ceux et celles qui m'ont aidé à réaliser ce travail et plus précisément à monsieur ***DjeddiChawki****, Maitre de conférences à l'université de Tébessa, mon encadreur, pour tous ses conseils, pour sa confiance totale et pour sa grande disponibilité.*

Mes vifs remerciements s'adressent aussi auxmembres du jury pour l'intérêt qu'ils ont porté à notre travail en acceptant a le juger

Nos remerciements vont à Mr. ***Gattal Abdeljalil****, Mr.* ***GahmousseAbdellatif***
et Mr . ***MerzougSoltane*** *qui n'ont jamais cessé de m'aider dans notre travail.*

Mesremerciements vont à ma famille qui n''a jamais cessé de nous manifester son soutien.
Merci à toutes les personnes qui ont contribué à la réussite de ce projet.

Liste des tableaux

Liste des figures

Liste des symbols

ANN	Artificial neural networks
BIF	Basic Image Features
HMM	Hidden Markov Models
IBM	International Business Machines
ICR	Intelligent character recognition
LBP	Local binary patterns
MLP	Multilayer perceptron
NLP	Natural language processing
NN	Neuron network
OBIF	Oriented Basic Image Features
OCR	Reconnaissance optique de caractères
RLSA	Run-Length Smoothening Algorithm
SVM	Support Vector Machine
TTP	Text to phone
UNIVAC	Universal Automatic Computer

Table des matières

CHAPITRE 1

Introduction Générale

INTRODUCTION GENERALE

1.1. Contexte du travail

La quantité de données multimédia capturées et stockées augmente rapidement avec les progrès de la technologie informatique. Ces données comprennent des documents multilingues. Par exemple, les musées stockent des images de tous les anciens documents fragiles ayant une valeur scientifique, historique ou artistique et écrits dans différents scripts qui sont stockés dans des bases de données généralement volumineuses. Les systèmes d'analyse de documents qui aident à traiter ces images stockées présentent un intérêt à la fois pour un archivage efficace et pour donner accès à divers chercheurs. L'identification de script est une étape clé qui se pose dans l'analyse d'image de document, en particulier lorsque l'environnement est multi-script et multilingue. Un schéma d'identification automatique des scripts est utile pour (i) trier les images de documents, (ii) aider à sélectionner les OCR spécifiques aux scripts appropriés et (iii) rechercher dans les archives en ligne des images de documents celles contenant un script particulier.

Les approches existantes de classification de scripts peuvent être classées en deux grandes catégories, à savoir les approches locales et globales. Les approches locales analysent une liste de composants connectés (comme une ligne, un mot et un caractère) dans les images du document pour identifier le script dans l'image du document. Cependant, ces composants ne sont disponibles qu'après la segmentation ligne, mot et caractère de l'image du document en question. En revanche, les approches globales utilisent l'analyse de régions comprenant au moins deux lignes et ne nécessitent donc pas de segmentation. Par conséquent, la tâche de classification des scripts est simplifiée et exécutée plus rapidement avec l'approche globale plutôt que locale. Ceci est caractéristique intéressante pour un système de recherche documentaire basés sur les scripts.

1.2. Objectifs du travail

Dans ce mémoire, nous présentons un système automatique pour l'identification de scripts à partir d'images de document imprimés et manuscrits. Ce système est basé sur l'extraction de caractéristiques compatibles avec la perception humaine. De telles caractéristiques sont extraites en utilisant les OBIFs conçues à une échelle optimale et à des multiples orientations. Toutes les caractéristiques sont extraites globalement d'une image de texte imprimé ou manuscrit qui ne nécessite aucune segmentation de l'image du document en lignes et en caractères. Le système proposé est efficace et peut être utilisé pour des nombreuses applications pratiques qui nécessitent le traitement de gros volumes de données. Le schéma a été testé sur 13 scripts et s'est avéré relativement insensible au changement de taille de police. Le système proposé atteint un taux d'identification global de l'ordre de 92% sur un grand ensemble de données.

1.3. Organisation du mémoire

Ce mémoire est structuré en deux parties. La première est consacrée à la présentation des principaux concepts, outils et travaux relatifs à l'étude entreprise. Dans la deuxième partie du mémoire, nous abordons de manière détaillée nos choix conceptuels, la mise en œuvre ainsi que les résultats obtenus par le système proposé pour l'identification de scripts à partir des images de documents.

Chapitre 2. ***Identification de scripts à partir d'images : concepts et outils***

Ce chapitre est consacré à la présentation des concepts liés directement avec le problème étudié. Il présente les travaux connexes dans le domaine de l'identification de scripts à partir des images des documents. Nous terminons ce chapitre la présentation des applications possibles des domaines de recherche étudiés.

Chapitre 3. Caractéristiques texturales pour l'identification de scripts à partir d'images de documents

Ce chapitre se détache des aspects théoriques abordés dans le deuxiéme chapitre et s'oriente vers la présentation de notre contribution qui consiste en une approche d'identification de scripts à partir des documents imprimés et manuscrits en se basant sur la caractérisation des différentes images des textes par les OBIFs. Nous décrivons la base de données utilisée avant de nous focaliser sur la présentation de la méthode d'extraction de caractéristiques proposée, les machines à vecteurs de support (SVMs) sont employées pour la classification. Les expérimentations effectuées seront aussi présentées. A la fin de ce chapitre, les résultats sont exposés et discutés.

A la fin de ce mémoire, nous émettons nos conclusions sur le travail que nous avons entrepris dans le domaine d'identification de scripts à partir de documents imprimés et manuscrits. Nous présentons aussi les perspectives d'extensions futures du travail que nous avons présenté dans ce document.

CHAPITRE 2

Identification du script : Concepts et outils

CHAPITRE 2

Identification du script : Concepts et outils

Ce chapitre est consacré à la présentation de la définition du script ainsi que la différence entre l'identification du script et l'identification de la langue, les six grands systèmes des écritures seront également décrits brièvement. Il présente les travaux connexes dans le domaine de l'identification de script à partir des images de textes imprimés, manuscrits et hybrides. Nous décrivons les caractéristiques ainsi que les techniques de classification utilisées dans la littérature. Nous terminons ce chapitre par la présentation des applications possibles du domaine de recherche considéré dans le cadre de ce mémoire.

2.1. Introduction

L'identification du script à partir d'une image d'un document manuscrit ou imprimé est la première étape pour un système OCR traitant des documents multi-scripts. Dans ce monde multilingue/multiscript, les systèmes de traitement des documents reposant sur l'OCR qui nécessitent une intervention humaine pour sélectionner le package OCR approprié sont définitivement indésirables et inefficaces. Le développement des méthodes robustes et efficaces pour l'identification automatique du script d'un document est un sujet d'importance majeur pour l'analyse et la reconnaissance automatique des documents dans un environnement multilingue/multiscript. Ainsi, l'objectif de base est de proposer des méthodes intuitives ayant une mise en œuvre simple sans compromettre l'efficacité. L'objectif de ce travail est d'évaluer les

techniques d'extraction et de classification de caractéristiques de pointe dans le domaine d'identification automatique des scripts à partir des documents imprimés et manuscrits et de proposer la meilleure combinaison pour ceux-ci.

2.2. Script

L'expression de l'esprit a été une tâche très difficile pour les êtres humains et ils ont essayés, depuis des siècles, de dépeindre leur moi intérieur au monde extérieur et on pense même qu'ils ont maîtrisés cette compétence. L'une des méthodes communes et principales de cette représentation est l'utilisation de l'écriture [3]. La technique de représentation du langage sous une forme visuelle, démonstrative et tangible s'appelle l'écriture. En enregistrant ou en inscrivant à l'aide de signes et de symboles, l'écriture est devenue un moyen important de conversation et de communication humaine qui représente les émotions et généralement la langue. Les systèmes d'écriture emploient une collection des symboles pour indiquer les sons de la parole, les chiffres et parfois la ponctuation [2]. L'enregistrement des événements et l'application des lois. L'écriture s'est avérée être la méthode fiable d'enregistrement et de présentation des affaires de manière durable et permanente lorsque le commerce et l'administration en Mésopotamie sont devenus complexes et ont dépassés le rappel humain vers le 4e millénaire avant notre ère [12]. Les techniques d'écriture ont évolués grâce au calcul et à l'enregistrement du temps sur des longues périodes appelées calendriers. Il peut également avoir évolué avec le besoin d'enregistrement politique, historique et environnemental dans l'Égypte ancienne et la Méso-Amérique.

Avec l'invention des systèmes informatiques, une réplication des systèmes d'écriture a eu lieu de manière automatisée. Les utilisateurs peuvent écrire sur des systèmes informatiques comme ils écrivaient sur papier. Les documents qui contiennent ces écrits sont préparés, stockés et manipulés dans ces systèmes informatiques dans une variété des langues [13]. Des centaines de langues, écrites à l'aide de divers scripts, sont prises en charge par les systèmes informatiques. Divers scripts sont utilisés pour écrire différentes langues. Le mot « Script » fait généralement référence à un manuscrit ou à un

document. Parfois, le texte d'un manuscrit/page/document particulier est également appelé script. Il fait également référence à la collection des lettres et des caractères utilisés pour l'écriture. Ce script s'est avéré être une caractéristique très importante et significative de tout document.

Les dernières décennies ont été témoins d'une énorme recherche et développement dans le domaine très potentiel de la gestion des documents dans le domaine de l'informatique, avec un accent majeur sur la tâche d'identification des scripts dans lesquels un document est écrit.

La première OCR pratique, à laquelle l'identification des scripts est un précurseur, est apparue dans les années 1950 aux États-Unis. L'ordinateur commercial préliminaire UNIVAC est également apparu dans la même décennie [14]. Depuis, la technologie OCR a connu des progrès au cours de chaque décennie suivante. Les premiers modèles des dispositifs optiques en ce qui concerne le traitement du langage, tels que les lecteurs optiques, y compris les premiers IBM 1418 et 1962 IBM 1428, ont été produits par IBM au début des années 1960.

2.3. Identification de script vs. Identification de la langue

L'identification de la langue est un domaine de recherche impératif en NLP. Il s'agit d'une procédure qui détermine la langue d'un élément à l'étude. Il est considéré comme un problème de catégorisation de texte qui est résolu à l'aide de différentes méthodes statistiques et par diverses approches informatiques. Dans un autre contexte, l'identification de la langue s'applique aux documents textes. Il est défini comme le processus qui détermine automatiquement, dans un document particulier, la langue d'un passage de texte écrit. La tâche qui identifie, dans un exemple de discours, la langue parlée par un locuteur anonyme est également appelée identification de la langue [4]. Une énorme quantité des documents textes et des documents Web sont disponibles en ligne [7] et sa récupération devient compliquée sans identifier avec précision la langue dans laquelle le document a été rédigé [17]. Ceci est important car le scénario du Web est multilingue en raison de l'augmentation exponentielle de l'échelle de stockage et de la vitesse d'accès [10]. Alors que l'identification de script s'applique à l'identification d'un script, à

travers lequel une langue est écrite, dans des images de document. Une quantité importante des travaux de recherche et de développement a été rapportée sur l'identification des langues. Plusieurs méthodes statistiques et d'apprentissage automatique ont été utilisés efficacement pour répondre à ce problème, mais la majorité des méthodes se concentrent sur la configuration hors ligne et peu de travaux ont été rapportés sur les solutions en ligne [19]. Des nombreuses techniques de classification ont été utilisées pour l'identification des langues, telles que la classification basée sur l'entropie relative, les machines à vecteurs de support, la classification bayésienne, la régression linéaire multiple, les modèles de Markov, les arbres de décision et les réseaux de neurones [19]. Certaines applications de l'identification de la langue incluent les systèmes de mappage texte-téléphone (TTP) [20]. Pour chaque langue, des sons de référence ont été utilisés dans les procédures qui effectuent l'identification acoustique de la langue ; caractéristiques acoustiques, caractéristiques prosodiques, caractéristiques de forme d'onde et caractéristiques phonétiques générales [4]. Une classe de techniques a été utilisée pour ces caractéristiques qui incluent les classifieurs quadratiques, les HMM, le clustering, les systèmes experts et les ANNs. Par conséquent, il est établi qu'il existe une énorme différence entre le concept et les techniques d'identification de script et d'identification de langue.

2.4. Systèmes d'écritures

Il existe dans le monde six grands systèmes d'écriture [39], [78]. Chaque système d'écriture comprend un ou plusieurs scripts et chaque script peut être utilisé dans une ou plusieurs langues.

2.4.1. Système logographique

Un système logographique représente généralement des mots complets. L'écriture Han est incluse dans ce système et utilisée dans les écrits chinois, japonais et coréens. Cette écriture se distingue des autres écritures occidentales et asiatiques par ses multiples traits courts, sa densité des caractères optiques et ses caractéristiques visuelles basées sur l'apparence.

2.4.2. Système syllabique

Dans ce système, chaque symbole représente une syllabe. Les scripts japonais utilisent un mélange de Kanji logographique et de Kanas syllabiques et font partie de ce système. Dans ces écritures, la présence des Kanas plus simples entre les logogrammes est moins dense que dans les écritures chinoises et c'est la caractéristique distinctive entre les écritures japonaises et han.

2.4.3. Système alphabétique

Les écritures les plus importantes du système alphabétique sont le grec, le latin, le cyrillique et l'arménien. Les Grecs ont été les premiers Européens à apprendre à écrire avec un alphabet et à partir de ce système, l'écriture alphabétique s'est répandue dans le reste de l'Europe, menant finalement à la création de tous alphabets européens modernes. L'écriture latine est utilisée dans de nombreuses langues à travers le monde telles que le latin, le cyrillique et l'arménien, ainsi que dans de nombreuses langues européennes comme l'anglais, l'italien, le français, l'allemand, le portugais, l'espagnol et l'austronésien, le malais moderne, le vietnamien, et l'indonésien. L'écriture cyrillique est utilisée dans certaines langues d'Europe de l'Est, asiatiques et slaves telles que le bulgare, le russe, le macédonien et l'ukrainien. Certains caractères de l'alphabet cyrillique sont empruntés au latin et au grec et modifiés avec des cédilles, des hachures ou des signes diacritiques.

2.4.4. Abjad

Dans ce système, chaque symbole représente un son consonantique. Il comprend des scripts arabes et hébreux. La caractéristique qui identifie clairement les scripts basés sur Abjad dans les systèmes informatiques à stylet est le sens d'écriture de droite à gauche.

2.4.5. Abugida

Ce système comprend la famille d'écritures brahmiques qui provient de l'ancienne écriture brahmi indienne et constitue la quasi-totalité des écritures de l'Inde et de l'Asie du Sud-Est. Le groupe nord des scripts brahmiques est utilisé dans les langues Devnagari, Bangla,

Manipuri, Gurumukhi, Gujrati et Oriya, langues balinaises. Une caractéristique importante de Devnagari, Bangla, Gurumukhi et Manipuri est que les mots sont écrits ensemble sans espaces. Le grand nombre des lignes horizontales dans les parties textuelles d'un document peut distinguer ces scripts des autres.

2.4.6. Système de fonctionnalités

Ce système comprend des fonctionnalités qui composent les phonèmes et comprend le script coréen Hangul. Dans ce système, un script est formé en mélangeant le Hanja logographique avec le Hangul fonctionnel. L'écriture coréenne est moins complexe et moins dense que les écritures chinoises et japonaises et elle contient plus de cercles et d'ellipses.

Les six systèmes d'écriture mentionnés ci-dessus incluent des nombreux scripts avec des caractères de forme similaire : les caractères de forme similaire sont la principale source de confusion dans l'identification des scripts. Habituellement, chaque script a plusieurs caractères spatiaux, des signes diacritiques, des graphiques multiples (y compris des digrammes) ou des ligatures qui diffèrent des autres scripts. dans un même système d'écriture. Ils sont importants pour identifier les scripts avec des caractères de forme similaire les uns aux autres. Par exemple, les 11 langues telles que l'afrikaans, le catalan, le néerlandais, l'anglais, le français, l'allemand, l'indonésien, le luxembourgeois, le malais, le portugais et les alphabets espagnols se composent de 26 caractères basés sur l'alphabet latin et se distinguent par des caractères spatiaux, des signes diacritiques, multi graphes et ligatures. En particulier, les 26 caractères latins sont inclus dans tous les alphabets linguistiques et seuls trois d'entre eux (anglais, indonésien et malais) sont sans signes diacritiques. Les 8 autres types d'alphabets linguistiques (afrikaans, catalan, néerlandais, français, allemand, luxembourgeois, portugais, espagnol) contiennent plusieurs signes diacritiques. Ces signes diacritiques peuvent être utilisés pour distinguer les 8 types de scripts dans l'identification des scripts au niveau des caractères. D'un autre côté, certains signes diacritiques sont communs à un certain nombre de scripts. Par exemple, le diacritique "é" n'est pas utile pour l'identification des scripts au niveau des caractères car il est courant pour l'afrikaans, le catalan, le néerlandais, le français, le

luxembourgeois, le portugais et l'espagnol. Dans ce cas, d'autres facteurs de ces langues doivent être pris en compte, tels que les multigraphes et les ligatures. Une comparaison des multigraphes et des ligatures dans les 11 écritures basées sur l'alphabet latin montre que les alphabets afrikaans, catalan, néerlandais et luxembourgeois n'ont pas de multigraphes ni de ligatures. Les alphabets anglais, français, allemand, indonésien, malais, portugais et espagnol contiennent plusieurs multigraphes. Les ligatures sont présentes dans les alphabets anglais, français et allemand. Par conséquent, ces multigraphes et ligatures peuvent être utilisés comme facteurs importants pour identifier ces types de scripts dans l'identification de scripts au niveau des caractères. Il est à noter que certains signes diacritiques ne sont pas utiles car ils sont communs à plusieurs scripts. Par exemple, le digramme ''ch'' est commun aux alphabets anglais, français, allemand, portugais et espagnol. Les caractères arabes, persans et ouïghours sont similaires : ils ont 18 caractères communs. De plus, les alphabets arabe et persan ont 8 caractères communs, l'arabe et l'ouïghour ont 2 caractères communs, le persan et l'ouïghour ont 6 caractères communs. Il y a 6 caractères ouïghours différents de l'arabe et du persan. Ainsi, l'identification de script au niveau caractère n'est pas efficace pour les 3 scripts et l'identification de script au niveau mot/composant connecté doit être envisagée.

2.5. Méthodes d'identification du script

La plupart des recherches dans le domaine de l'identification des scripts portent sur des documents imprimés ou manuscrits. Cependant, étant donné que plusieurs documents peuvent contenir des blocs de texte avec des scripts imprimés et manuscrits, certaines recherches portent désormais sur les documents hybrides. Par conséquent, sur la base du type de contenu, les documents peuvent être classés en trois catégories : imprimés, manuscrits et hybrides. De plus, l'acquisition de documents peut être effectuée non seulement à l'aide de scanners optiques, mais également via des appareils photo et des caméscopes. L'acquisition peut affecter la qualité de l'image du document et, par conséquent, des méthodes d'identification de script spécifiques ont récemment été proposées pour les acquisitions vidéo et par caméra. La recherche sur l'identification de scripts des documents imprimés, manuscrits et hybrides est discutée ci-après. Pour chaque type de document, différentes méthodes présentées

dans la littérature sont introduites en fonction du type de données qu'elles utilisent pour effectuer l'identification du script : Page/Paragraphe/Bloc de texte, Ligne de texte, Mot ou Caractère.

2.5.1. Identification du script à partir de documents imprimés

La plupart des recherches d'identification de script ont été effectuées sur des documents imprimés. Les principales sources des documents imprimés sont les livres, les magazines, les revues, les dictionnaires, etc. Certains chercheurs ont d'abord préparé des textes multiscripts à l'aide d'un logiciel de traduction automatique [36] ou d'autres logiciels [5], puis des documents multiscripts sont obtenus sous forme d'impressions informatiques. Étant donné la diversité des scripts et le manque de bases de données publiques disponibles, la plupart des chercheurs ont construit leurs propres bases de données / ensembles de données.

2.5.1.1. Identification du script à partir d'un texte

La plupart des recherches sur l'identification des scripts des documents imprimés ont été effectuées au niveau de la page. Hochberg et al. [46] ont utilisé des modèles basés sur des clusters pour discriminer 13 scripts différents. Spitz [119] a proposé un schéma d'identification des langues dans lequel les mots de 26 langues différentes ont d'abord été classés en écritures basées sur Han et sur latin. Successivement, les langues réelles ont été identifiées à l'aide de profils de projection de mots et de formes de caractères. Jie Ding et al. [34] ont présenté une méthode qui utilise une analyse combinée des plusieurs caractéristiques statistiques discriminantes pour catégoriser les écritures des langues européennes et orientales. Chaudhuri et Pal [21] ont développé un système pour identifier les écritures Bangla et Devnagari (Hindi) à l'aide d'un arbre de classification. Des recherches sur l'identification des scripts de documents imprimés ont également été menées au niveau des blocs de texte. Par exemple, Peake et Tan [85] ont proposé une méthode basée sur des filtres de Gabor multiples et des matrices de cooccurrence de niveaux de gris pour extraire les caractéristiques de texture de cinq scripts majeurs.

2.5.1.2. Identification du script à partir d'une ligne de texte

Dans l'identification de script au niveau de la ligne de texte, un bloc de texte est d'abord divisé en lignes. Pal et Chaudhuri [71] ont développé une technique automatique pour séparer les lignes de texte en utilisant des caractéristiques de script et des caractéristiques basées sur la forme. Ils ont également proposé un système d'identification des lignes de texte imprimées en romain, chinois, arabe, devnaguri et bangla à partir d'un seul document et une méthode d'identification des lignes de texte de différentes écritures indiennes à partir d'un document [73]. Une technique automatique pour l'identification des portions d'écriture japonaise et anglaise à partir d'une seule ligne d'un document imprimé a été proposée par Chanda et al. [13]. Padma et Vijaya [68] ont développé un modèle algorithmique monothétique pour identifier et séparer les lignes de texte en télougou, en hindi et en anglais d'un document multilingue imprimé. Une technique simple et efficace d'identification de script pour les lignes de texte en kannada, hindi et anglais a été présentée par Prakash et al. [90]. Ferrer et al. [36] ont proposé un système d'identification de script par ligne basé sur LBP pour identifier dix scripts différents.

2.5.1.3. Identification du script à partir d'un mot

Dhanya et Ramakrishnan [33] ont présenté une méthode efficace pour identifier l'écriture au niveau du mot dans un document bilingue contenant des écritures romaines et tamoules. Jaeger et al. [48] ont utilisé une analyse par filtre de Gabor des textures et un système de classifieurs multiples avec quatre classifieurs différents pour identifier les scripts arabes, chinois, hindi et coréens au niveau des mots. Dhandra et al. [28], [29] ont proposé une technique automatique d'identification de script au niveau du mot basée sur la reconstruction morphologique de deux scripts imprimés : Telugu et Devnagari. Une méthode basée sur la SVM a été proposée par Chanda et al. [14] pour l'identification des scripts anglais et thaï imprimés au niveau du mot à partir d'une seule ligne d'une page de document. Chande et al. [15] ont proposé une technique basée sur SVM pour l'identification au niveau des mots des scripts cinghalais, tamouls et anglais à partir d'une seule page de document, et un schéma basé sur SVM pour

l'identification des scripts imprimés au niveau des mots anglais, Devnagari et Bangla [16].

2.5.1.4. Identification du script à partir d'une lettre

Pal et Sarkar [74] ont utilisé une combinaison des caractéristiques topologiques, de contour et de concept de réservoir d'eau pour identifier l'écriture ourdou imprimée. Rani et al. [91] ont mené des expériences sur des caractères à polices multiples et à tailles multiples avec des caractéristiques Gabor et des caractéristiques Gradient pour identifier les écritures Gurumukhi et anglaises au niveau des caractères ou des chiffres.

2.5.2. Identification du script à partir de documents manuscrits

Les documents manuscrits sont un autre domaine d'application important pour les systèmes d'identification de scripts. Bien entendu, l'identification par script de documents manuscrits est plus difficile que l'identification par script de documents imprimés. En fait, il existe des différences pertinentes entre l'identification manuscrite et l'identification manuscrite. Par exemple, certains scripts se ressemblent beaucoup plus dans les documents manuscrits que dans les documents imprimés. De plus, les styles d'écriture peuvent être très variables. Les documents expérimentaux, rédigés par des individus différents à des moments différents, élargissent l'inventaire des formes possibles des caractères et de mots dans les documents manuscrits. De plus, les lignes directrices et la fragmentation des caractères sont courantes dans les documents manuscrits en raison de la variété des papiers et des instruments d'écriture utilisés. Toutes ces différences peuvent créer d'énormes défis pour l'identification des scripts dans les documents manuscrits.

2.5.2.1. Identification de script à partir d'un texte

La première étude menée sur l'identification des scripts manuscrits a été réalisée par Chaudhuri [22] et était similaire à celle proposée par Hochberg et al. [46] pour les documents imprimés. Cependant, la précision de la classification résultante était inférieure à celle des documents imprimés. Un système de reconnaissance des scripts manuscrits en ligne a été proposé par Namboodiri et Jain [64] pour

classer six scripts majeurs au niveau des mots. Onze caractéristiques différentes et six types de classifieurs ont été considérés. Une méthode basée sur les caractéristiques de texture pour l'identification de script dans une image de document manuscrit a été proposée par Hiremath et al. [44]. Ghosh et Shivaprasad [39] ont proposés une méthode d'identification des scripts manuscrits dans laquelle une approche « possibiliste » a été utilisée pour l'analyse de cluster.

2.5.2.2. Identification du script à partit d'une ligne de texte

Namboodiri et Jain [64] ont proposés une méthode pour classer les mots et les lignes dans l'une des six écritures principales : arabe, cyrillique, devnagari, han, hébreu ou romain. La classification est basée sur onze caractéristiques spatiales et temporelles différentes extraites des traits des mots.

2.5.2.3. Identification du script à partir d'un mot

Roy et al. [96] ont proposé une méthode d'identification des scripts manuscrits au niveau des mots pour l'automatisation postale indienne concernant l'identification des scripts en bengali et en anglais au niveau des mots. La méthode utilise principalement des caractéristiques basées sur le concept de réservoir d'eau, des caractéristiques basées sur des fractales et un classifieur de réseau neuronal. Roy et Majumder [97] ont également développé une technique de séparation des scripts des documents postaux manuscrits en alphabet bengali, romain et devanagri. L'algorithme de lissage de la longueur d'exécution (RLSA) a été utilisé pour segmenter les pages du document en lignes, puis en mots. Des caractéristiques fractales, de zone occupée et topologiques ont été utilisées avec un classifieur de réseau neuronal (NN) pour l'identification de script. Une technique de séparation des scripts romains et oriya pour l'automatisation postale indienne a été proposée par Zhou et al. [133]. Ils ont présenté une méthode d'identification de script basée sur des caractéristiques basées sur le concept de réservoir d'eau, des caractéristiques basées sur des dimensions fractales et des caractéristiques topologiques avec un classifieur NN. Sarkar et al. [103] ont présenté un système de séparation automatique pour l'identification des scripts au niveau des mots du bengali ou du devanagri mélangé avec des scripts romains. Dhandra et Hangarge

[30] ont utilisés une approche en deux étapes. Dans la première étape, certaines caractéristiques globales et locales ont été appliquées pour identifier les mots du texte. Dans la deuxième étape, le chiffre écrit dans différentes écritures a été identifié. Pour tester le système, des documents manuscrits écrits en Kanada, Devanagri et Roman ont été considérés. Un système d'identification d'écriture manuscrite au niveau des mots pour les documents bi-script écrits en écriture persane et romaine a été proposé par Roy et al. [99]. Le système utilisait un ensemble calculable simple et rapide de douze caractéristiques basées sur la dimension fractale, la position des petits composants et la topologie. Un schéma d'identification d'écriture manuscrite au niveau du document à partir de six documents d'écriture indienne populaires a été présenté par Roy et al. [98]. Dans le schéma proposé, un petit ensemble de caractéristiques basées également sur la dimension fractale sont calculés à l'aide d'un classifieur MLP. Obaidullah et al. [67] ont proposé un schéma pour identifier les six écritures populaires Bangla, Devnagari, Malayalam, Urdu, Oriya et Roman dans les documents indiens, et comparé les performances en utilisant différents classifieurs bien connus.

2.5.2.4. Identification du script à partir d'une lettre

Pal et al. [75] ont proposés un système basé sur un classifieur quadratique modifié pour la reconnaissance des chiffres manuscrits hors ligne de six écritures indiennes populaires : Devnagari, Bangla, Telugu, Oriya, Kannada et Tamil. Razzak et al. [93] ont présenté une approche basée sur des règles floues pour la reconnaissance des chiffres en ourdou et en arabe dans un environnement sans contrainte.

2.5.3. Identification du script à partir de documents hybrides

Les documents hybrides comprennent des textes imprimés et manuscrits. Une identification automatique multilingue de l'arabe et du latin en écriture manuscrite et imprimée a été proposée par Ben Moussa et al [6]. Une méthode de différenciation des blocs de texte arabe et latin pour les scripts imprimés et manuscrits est discutée par Kanoun et al. [50]. La méthode est basée sur une analyse morphologique de chaque script au niveau du bloc de texte et une analyse géométrique au niveau des lignes et des composants

connectés. Benjelil et al. [8] ont proposé un système précis basé sur une transformation pyramidale orientable pour l'identification des scripts arabes et latins au niveau des mots. En utilisant de nouvelles caractéristiques structurelles, une tentative réussie a été faite par Saidani et al. [101] pour identifier l'écriture arabe ou latine d'un document imprimé à la machine ou manuscrit au niveau du mot.

2.6. Caractéristiques pour l'identification de script :

L'extraction de caractéristiques est une partie vitale de tout système de reconnaissance pratique. Au cours des dernières années, différents types de caractéristiques ont été évalués pour l'identification de script en fonction des caractères de chaque script. Deux grandes catégories d'entités ont été établies dans le champ d'identification de script : l'entité locale et l'entité globale. Les caractéristiques locales sont extraites de petites composantes textuelles de l'image du document. Par conséquent, elles dépendent fortement de l'efficacité de la procédure de segmentation. Les caractéristiques statistiques, structurelles et basées sur des modèles sont des exemples de caractéristiques locales [26]. Les caractéristiques globales sont extraites des blocs de texte de l'image du document. Les caractéristiques basées sur la texture et la pyramide orientable sont des exemples de caractéristiques globales [26].

2.7. Techniques de classification pour l'identification du script

Bien que la classification soit une étape cruciale des systèmes d'identification de scripts, la littérature montre que seuls quelques classifieurs simples ont été utilisés dans les travaux antérieurs, comme le rapporte le tableau 5 [21],[31], [34], [46], [118], [119], [129]. Les classifieurs K-Nearest Neighbor (K-NN) ont été largement utilisés dans les systèmes d'identification de scripts basés sur le filtre de Gabor [81],[85], les moments cartésiens [1], les approches de modèle basées sur l'apparence [125], la cooccurrence des niveaux de gris caractéristiques matricielles [85], caractéristiques statistiques [20], densité des traits et caractéristiques basées sur la distribution [56], caractéristiques de texture [20], [42]. La machine à vecteurs de support (SVM) a également été appliquée à l'identification de scripts. Les systèmes basés sur SVM pour l'identification de script utilisent des caractéristiques structurelles, des caractéristiques topologiques et des

caractéristiques basées sur le principe du réservoir d'eau [14], [15], la caractéristique basée sur le moment de Zernike [18], [104], Gabor et les caractéristiques de gradient [104]. Autre classification ont été considérées pour l'identification des scripts telles que Neural Network [9], [96], les classifieurs quadratiques [43], [64], [75], [126], [134], les classifieurs basés sur des règles[1], [90], [92], [93], Classifieurs discriminants linéaires [43], [55], [83], Modèle de mélange gaussien [11], [48], [50], [99], etc...

2.8. Applications

Une quantité colossale de travail a été signalée dans le rapide progrès et vaste domaine de l'identification de script au cours des dernières décennies. Le nombre croissant des études, des recherches et du développement de l'identification des scripts sont guidés par la variété des scripts utilisés partout dans le monde et sont guidés par un certain nombre d'applications importantes dans la gestion de documents numériques systèmes. La technologie d'identification des scripts s'est avérée applicable dans l'ensemble du spectre des industries et a ainsi révolutionné le processus de gestion des documents. Ces dernières années, les efforts d'identification des scripts ont les intervenants ont permis de traiter et de réaliser des documents numérisés plus que de simples fichiers images leur permettant de les convertir dans des documents entièrement consultables qui possèdent un texte contenu reconnu par les ordinateurs. Cela a conduit à traitement de l'information correct, efficace et efficient. Il permet un traitement plus rapide en économisant un temps précieux. Scénario l'identification trouve son utilisation dans un certain nombre des domaines tels que :

Archivage: L'identification du script fournit solution d'archivage c'est-à-dire de transcription automatique des différents types des documents.

Saisie de données: L'identification de script permet des données automatiques saisie des documents commerciaux, par ex. chèques, factures, conversion de passeport, relevé bancaire, formulaires et reçus texte imprimé ou imprimé à la main à partir du papier document en données électroniques pouvant être utilisées dans un système informatique (à l'aide d'un logiciel OCR ou ICR).

Traitement automatisé des formulaires: Aide à l'identification des scripts pour identifier les informations importantes (nom, adresse, date de naissance, articles, prix, qualification, conditions de paiement, etc.), suivi de la correspondance les informations du document avec les données dans le système d'entreprise.

Indexation et étiquetage: Les identifications de script fournissent entrées pour les services d'annuaire, d'étiquetage et de catalogage pour les images de documents qui permet une meilleure et efficace stockage des données numérisées.

Recherche et tri: L'identification du script donne des entrées précieuses pour la recherche et le tri numérisés images de document en identifiant le contenu des images. Crée des images électroniques de documents imprimés facilement consultable.

Sélection de l'OCR: L'identification du script agit comme un précurseur pour sélectionner l'OCR approprié dans un environnement multilingue

Autres applications: L'identification de script s'avère fructueuse dans des secteurs tels que la banque, la santé, l'assurance, l'éducation, les finances et les agences gouvernementales pour gérer leurs Besoins. Il économise de l'espace et du temps. De plus, en passant au crible boîtes de dossiers papier est éliminé. Une autre application clé de l'identification automatique des scripts est l'archivage automatique et traitement concernant les documents multilingues et suivi d'une OCR multilingue.

2.9. Conclusion

Dans ce chapitre, Nous nous présentons un examen exhaustif de diverses techniques de l'identification des scripts dans des contextes internationaux et nationaux A été présenté. Nous avons souligné la différence entre l'identification de la langue et l'identification du script. Enfin, un ensemble de domaines d'application de l'identification de script ont été enrôlés.

Bibliographie

[1] V. Ablavsky and M. R. Stevens, ``Automatic feature selection with applications
to script identification of degraded documents," in *Proc. ICDAR*,Aug. 2003, pp. 750_754.
[2]Data entry: Script identification enables automatic data entry for business documents, e.g. cheques, invoices,passport, bank statement, forms and receipts converting printed or hand-printed text from the paper document into electronic data that can be used in a computer system (using OCR or ICR software).
[3] Archival: Script identification provides automatic solution for archiving i.e. automatic transcription of different kinds of documents.
[4] Muthusamy YK, Barnard E, Cole RA (1994) Reviewing automatic language identification. IEEE Signal Process Mag 11:33–41
[5] R. Bashir and S. Quadri, ``Identification of Kashmiri script in a bilingual document image," in *Proc. ICIIP*, Shimla, India, Dec. 2013, pp. 575_579.
[6] S. Ben Moussa, A. Zahour, A. Benabdelha_d, and A. M. Alimi, ``Fractalbased system for Arabic/Latin, printed/handwritten script identification,"in *Proc. ICPR*, Tampa, FL, USA, Dec. 2008, pp. 1_4.
[7] Selamat A, Ng CC (2011) Arabic script web page language identification using decision tree neural networks. Pattern Recognit 44:133–144
[8] M. Benjelil, R. Mullot, and A. Alimi, ``Language and script identification based on steerable pyramid features," in *Proc. ICFHR*, Bari, Italy,Sep. 2012, pp. 716_721.
[9] U. Bhattacharya and B. B. Chaudhuri, ``Handwritten numeral databases of Indian scripts and multistage recognition of mixed numerals," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 3, pp. 444_457, Mar. 2009.
[10] Mishra G, Nitharwal SL, Kaur S (2010) Language identification using fuzzy-SVM technique. In: Proceedings of IEEE 2nd international conference on computing, communication and networking technologies

[11] A. Busch, W. W. Boles, and S. Sridharan, ``Texture for script identification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 11,pp. 1720_1732, Nov. 2005.
[12]ROBINSON C (2003) Literacy – new meanings and dimensions.In: Paper presented at the UNESCO expert meeting: renewed vision of literacy and policy implications. Paris
[13] S. Chanda, U. Pal, and F. Kimura, ``Identification of Japanese and English
script from a single document page," in *Proc. IEEE-CIT*, Oct. 2007,pp. 656_661.
[14] S. Chanda, O. R. Terrades, and U. Pal, ``SVM based scheme for Thai and English script identification," in *Proc. ICDAR*, Paraná, Argentina,Sep. 2007, pp. 551_555.

[15] S. Chanda, S. Pal, and U. Pal, ``Word-wise Sinhala Tamil and English script identification using Gaussian kernel SVM," in *Proc. ICPR*, Tampa,FL, USA, Dec. 2008, pp. 1_4.
[16] S. Chanda, S. Pal, K. Franke, and U. Pal, ``Two-stage approach for wordwise script identification," in *Proc. ICDAR*, Jul. 2009, pp. 926_930.
[17]Nguyen TD, Zreik K (2004) Multilingual hyperdocumnet recognition: a document mining approach. In: IEEE
[18] S. Chanda, fK. Franke, and U. Pal, `` identification of Indic scripts on torndocuments," in *Proc. ICDAR*, Beijing, China, Sep. 2011, pp. 713_717.
[19] Tacki H, Gungor T (2012) A high performance centroid-based classification approach for language identification. Pattern Recognit Lett 33:2077–2084
[20] S. A. Chaudhari and R. M. Gulati, ``An OCR for separation and identification of mixed English_Gujarati digits using kNN classi_er," in *Proc.ISSP*, Gujarat, India, Mar. 2013, pp. 190_193.
[21] B. B. Chaudhuri and U. Pal, ``An OCR system to read two Indian language scripts: Bangla and Devnagari (Hindi)," in *Proc. ICDAR*, Ulm, Germnay,1997, pp. 1011_1015.
[22] B. B. Chaudhuri, ``On multi-script OCR system evaluation,"in *Proc. Int. Workshop Perform. Eval. Issues Multi-Lingual (OCR)*, 1999, p. 1. [Online]. Available: http://www.kanungo.com/ workshop/abstracts/chaudhuri.html
[23] Bilcu EB, Astola J (2007) Improved hybrid approach for language recognition from text. In: Proceedings of 5th international symposium on image and signal processing and analysis
[26] S. Dalal and L. Malik, ``A survey of methods and strategies for feature extraction in handwritten script identification," in *Proc. Int. Conf. Emerg.Trends Eng. Technol. (ICETET)*, Nagpur, India, Jul. 2008, pp. 1164_1169.
[28] B. V. Dhandra, H. Mallikarjun, R. Hegadi, and V. S. Malemath, ``Wordwise
script identification based on morphological reconstruction in printed bilingual documents," in *Proc. IET Int. Vis. Inf. Eng. (VIE)*, Bangalore, India, Sep. 2006, pp. 389_393.
[29] B. V. Dhandra, H. Mallikarjun, R. Hegadi, and V. S. Malemath, ``Wordwise script identi_cation from bilingual documents based on morphological reconstruction," in *Proc. Int. Conf. Dig. Inf. Manage.*, Bangalore,India, Dec. 2006, pp. 389_394.
[30] B. V. Dhandra and M. Hangarge, ``Global and local features based handwritten text words and numerals script identi_cation," in *Proc. ICCIMA*, Dec. 2007, pp. 471_475.
[31] B. V. Dhandra, M. Hangarge, R. Hegadi, and V. S. Malemath,`Word level script identi_cation in bilingual documents through discriminating features," in *Proc. ICSCN*, Chennai, India, Feb. 2007,pp. 630_635.
[33] D. Dhanya, A. G. Ramakrishnan, and P. B. Pati, ``Script identification in printed bilingual documents," *Sadhana*, vol. 27, no. 1, pp. 73_82,Feb. 2002.
[34] J. Ding, L. Lam, and C. Y. Suen, ``Classification of oriental and European scripts by using characteristic features," in *Proc. ICDAR*, Ulm, Germany,Aug. 1997, pp. 1023_1027.

[36] M. A. Ferrer, A. Morales, and U. Pal, ``LBP based line-wise script identification," in *Proc. ICDAR*, Washington, DC, USA, Aug. 2013,pp. 369_373.
[39] D. Ghosh and A. P. Shivaprasad, ``Handwritten script identification using possibilistic approach for cluster analysis," *J. Indian Inst. Sci.*, vol. 80,no. 3, p. 215, 2000.
[42] J. Gllavata and B. Freisleben, ``Script recognition in images with complex backgrounds," in *Proc. ISSPIT*, Athens, Greece, Dec. 2005,pp. 589_594.
[43] M. Hangarge, K. C. Santosh, and R. Pardeshi, ``Directional discrete cosine transform for handwritten script identification," in *Proc. ICDAR*,Washington, DC, USA, Aug. 2013, pp. 344_348.
[44] P. S. Hiremath, S. Shivashankar, J. D. Pujari, and V. Mouneswara, ``Script identification in a handwritten document image using texture features,"in *Proc. IACC*, Patiala, India, Feb. 2010, pp. 110_114.
[46] J. Hochberg, P. Kelly, T. Thomas, and L. Kerns, ``Automatic script identification from document images using cluster-based templates," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 2, pp. 176_181, Feb. 1997.
[48] S. Jaeger, H. Ma, and D. Doermann, ``Identifying script on wordlevel with informational con_dence," in *Proc. ICDAR*, Aug. 2005,pp. 416_420.
[50] S. Kanoun, A. Ennaji, Y. Lecourtier, and A. M. Alimi, ``Script and nature differentiation for Arabic and Latin text images," in *Proc. IWFHR*, Aug. 2002, pp. 309_313.
[55] X.-R. Lin, C.-Y. Guo, and F. Chang, ``Classifying textual components of bilingual documents with decision-tree support vector machines," in *Proc. ICDAR*, Beijing, China, Sep. 2011, pp. 498_502.
[64] A. M. Namboodiri and A. K. Jain, ``On-line Script Recognition,"*IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 1, pp. 124_130,Jan. 2004.
[67] S. M. Obaidullah, K. Roy, and N. Das, ``Comparison of different classifiers for script identification from handwritten document," in *Proc.ISPCC*, Sep. 2013, pp. 1_6.
[68] M. C. Padma and P. A. Vijaya, ``Monothetic separation of Telugu, Hindi and English text lines from a multi script document," in *Proc. SMC*,Oct. 2009, pp. 4870_4875.
[71] U. Pal and B. B. Chaudhuri, ``Script line separation from Indian multi-Script documents," in *Proc. ICDAR*, Bangalore, India, 1999,pp. 406_409.
[73] U. Pal, S. Sinha, and B. B. Chaudhuri, ``Multi-script line identification from Indian documents," in *Proc. ICDAR*, Bangalore, India, 2003,pp. 880_884.
[74] U. Pal and A. Sarkar, ``Recognition of printed urdu script," in *Proc.ICDAR*, Bangalore, India, 2003, pp. 1183_1187.
[75] U. Pal, N. Sharma, T. Wakabayashi, and F. Kimura, ``Handwritten numeral recognition of six popular Indian scripts," in *Proc. ICDAR*, Parana, Sep. 2007, pp. 749_753.
[78] U. Pal, ``Automatic script identification: A survay," *J. VIVEK*, Bombay, vol. 16, no. 3, pp. 26_35, 2006.
[81] P. B. Pati and A. G. Ramakrishnan, ``HVS inspired system for script identification in Indian multi-script documents," in *Document Analysis*

Systems VII. DAS, (Lecture Notes in Computer Science), vol. 3872. H.Bunke and A. L. Spitz, Eds. Berlin, Germany: Springer, 2006, pp. 380_389.
[83] P. B. Pati and A. G. Ramakrishnan, ``Word level multi-script identification,'' *Pattern Recognit. Lett.*, vol. 29, no. 9, pp. 1218_1229, 2008.

[85] G. S. Peake and T. N. Tan, ``Script and language identification fromdocument
images,'' in *Computer Vision-ACCV* (Lecture Notes in Computer Science), vol. 1352. R. Chin and T. C. Pong, Eds. Berlin, Germany:Springer, 1998,pp. 97_104.
[90] K. A. Prakash, G. Rajesh, U. A. Dinesh, M. Krisnamoorthi, and N. V. Subbareddy, Text line script identification for a trilingual document,''in *Proc. ICCCNT*, 2010, pp. 1_3.
[91] R. Rani, R. Dhir, and G. S. Lehal, ``Script identification of pre-segmented multi-font characters and digits,'' in *Proc. ICDAR*, Washington, DC,USA, Aug. 2013, pp. 1150_1154.
[92] R. Rani, R. Dhir, and G. S. Lehal, ``Performance analysis of feature extractors and classifiers for script recognition of English and Gurmukhi words,'' in *Proc. DAR*, 2012, pp. 30_36.
[93] M. I. Razzak, S. A. Hussain, and M. Sher, ``Numeral recognition for Urdu script in unconstrained environment,'' in *Proc. ICET*, Oct. 2009, pp. 44_47.
[96] K. Roy, U. Pal, and B. B. Chaudhuri, ``Neural network based word-wise handwritten script identification system for Indian postal automation,'' in *Proc. ICISIP*, Jan. 2005, pp. 240_245.
[97] K. Roy and K. Majumder, ``Trilingual script separation of handwritten postal document,'' in *Proc. ICVGIP*, Dec. 2008, pp. 693_700.
[98] K. Roy, S. K. Das, and S. M. Obaidullah, ``Script identification from handwritten document,'' in *Proc. NCVPRIPG*, Dec. 2011, pp. 66_69.
[99] K. Roy, A. Alaei, and U. Pal, ``Word-wise handwritten persian and roman script identification,'' in *Proc. ICFHR*, Kolkata, India, Nov. 2010,pp. 628_633.
[101] A. Saïdani, A. K. Echi, and A. Belaïd, `` identification of machine-printed and handwritten words in Arabic and Latin Scripts,'' in *Proc. ICDAR*,Washington, DC, USA, Aug. 2013, pp. 798_802.

[103] R. Sarkar, N. Das, S. Basu, M. Kundu, M. Nasipuri, and D. K. Basu,``Word level script identification from bangla and devanagri handwritten texts mixed with roman script,'' *J. Comput.*, vol. 2, no. 2, pp. 103_108,2010.
[104] N. Sharma, S. Chanda, U. Pal, and M. Blumenstein, ``Word-wise script identification from video frames,'' in *Proc. ICDAR*, Washington, DC,USA, Aug. 2013, pp. 867_871.
[118] A. Spitz, ``Script and language determination from document images,'' in *Proc. Symp. Document Anal. Inf. Retr.*, Las Vegas, NV, USA, Apr. 1994,pp. 229_235.
[119] A. L. Spitz, ``Determination of the script and language content of document images,'' *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 3,pp. 235_245, Mar. 1997.

[125] T. N. Vikram and D. S. Guru, ``Appearance based models in document script identification,'' in *Proc. ICDAR*, Parana, Sep. 2007, pp. 709_713.
[126] B. Waked, S. Bergler, C. Y. Suen, and S. Khoury, ``Skew detection, page segmentation, and script classification of printed document images,'' in *Proc. ICSMC*, San Diego, CA, USA, Oct. 1998, pp. 4470_4475.
[129] S. L. Wood, X. Yao, K. Krishnamurthi, and L. Dang, ``Language identification for printed text independent of segmentation,'' in *Proc. ICIP*,vol. 3. Washington, DC, USA, Oct. 1995, pp. 428_431.
[133] L. Zhou, Y. Lu, and C. L. Tan, ``Bangla/English script identification based on analysis of connected component pro_les,'' in Document Analysis Systems VII. DAS (Lecture Notes in Computer Science), vol. 3872.H. Bunke and A. L. Spitz, Eds. Berlin, Germany: Springer, 2006,pp. 243_254.
[134] L. Zhou, X. J. Ping, E. G. Zheng, and L. Guo, ``Script identification based on wavelet energy histogram moment features,'' in *Proc. ICSP*, Beijing,China, Oct. 2010, pp. 980_983.

CHAPITRE 3

Caractéristiques texturales pour l'identification de script à partir d'images de documents

CHAPITRE 3

Caractéristiques texturales pour l'identification de scripts à partir d'images de documents

Ce chapitre présente notre contribution principale qui consiste en une étude sur l'identification de scripts à partir de documents imprimés et manuscrits. La méthode proposée est basée sur l'extraction d'un ensemble des caractéristiques texturales à partir d'images de documents manuscrits et imprimés et l'entraînement d'un classifieur afin qu'il puisse identifier le script d'un document en question. Des attributs des textures sont estimés en calculant les OBIFs (Oriented Basic Image Features). La classification est effectuée à l'aide les séparateurs à vaste marge (SVM) avec la stratégie un contre tous. La méthode proposée a été évaluée en utilisant un ensemble de données contenant 10.400 documents imprimés et manuscrits où des résultats intéressants ont été enregistrés.

3.1. Introduction

Le présent chapitre vise à mettre au point un système d'identification des scripts entièrement automatique qui pourra être appliqué à différentes fins telles que la compréhension de scènes [01], la recherche d'images [02], la navigation sur téléphone portable, la reconnaissance de sous-titres vidéo [03] et la traduction automatique [04, 05]. Le système proposé est conçu pour fonctionner sur des documents imprimés et manuscrits.

Ce chapitre est organisé comme suit. Dans la section suivante, nous décrivons la base de données. Nous présentons ensuite la description de la technique de classification utilisée dans la section 3.3, suivies

par la description des caractéristiques proposées dans la section 3.4. Les résultats expérimentaux et leur analyse sont présentés dans la section 3.5, tandis que la dernière section conclut ce chapitre.

3.2. Base de données

La méthode proposée est évaluée sur l'ensemble de données fourni par les organisateurs de la compétition internationale sur l'identification de scripts à partir de documents historiques (1st competition on script identification in the wild SIW 2021) [05] qui s'est déroulée en conjonction de la conférence ICDAR 2021. Cet ensemble des données contient à la fois des documents imprimés et manuscrits obtenus à partir d'une grande variété d'écritures, telles que l'arabe, le bengali, le gujarati, le gurmukhi, le devanagari, le japonais, le kannada, le malayalam, l'oriya, le romain, le tamoul, le télougou et le thaï. L'ensemble des données se compose de 1137 documents numérisés à partir de journaux locaux, ainsi que de lettres et de notes manuscrites. De plus, ces documents sont segmentés en lignes et mots, comprenant un total de 13 983 et 86 675 lignes et mots, respectivement.

Figure 3.1 : Exemples d'images de textes écrits dans différents scripts.

3.3. Classification

Une fois que les images de documents à comparer sont représentées par leurs caractéristiques, nous procédons à l'utilisation de ces vecteurs de caractéristiques pour l'entrainement ou le test. L'entrainement et la classification sont effectuées en utilisant les séparateurs à vaste marge multi-classes (SVM). Une brève introduction à ce classifieur est donnée dans la sous section suivante.

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.

3.4. Extraction de caractéristiques

Le système d'identification de script proposé est basé sur un ensemble des caractéristiques texturales basées sur les OBIFs extraites à partir des images des documents manuscrits et imprimés. Ces caractéristiques sont décrites dans les sous sections suivantes.

L'ensemble des mesures de texture étudiées dans notre étude comprend les OBIFs. Ces caractéristiques ont été appliquées avec succès à des problèmes tels que la classification des textures [07], la reconnaissance des chiffres [08] et l'identification des scripteurs [06, 10] et la classification du genre à partir de l'écriture manuscrite [09]. Les OBIFs représentent une extension des BIFs (Basic Image Features) [11, 12]. L'idée clé est d'étiqueter chaque emplacement de l'image avec l'une des sept classes de symétrie locales. Les caractéristiques sont calculées en appliquant un banc des dérivées de filtres gaussiens contrôlés par un paramètre d'échelle. Un paramètre supplémentaire ε est utilisé pour classer un emplacement comme « plat ». Trois de ces types de symétrie sont accompagnés de *n* orientations possibles, la classe de pente a *2n* orientations possibles tandis que trois des classes n'ont aucune orientation. Cela donne un vecteur OBIF de *5n + 3*. Plus de détails sur les aspects calculatoires des oBIF peuvent être trouvés dans [09].

Dans le présent travail, nous quantifions l'orientation en *n=4* directions, produisant ainsi 23 entrées *(5×4+3)* dans le dictionnaire OBIFs. Semblable au descripteur LBP, les OBIFs à deux échelles différentes (en ignorant le type de symétrie plat) et l'histogramme de la colonne sont calculés. L'histogramme a un total de *(5n+2)2=484* entrées. Le paramètre d'échelle est choisi dans l'ensemble $\sigma \in \{1, 2, 4, 8, 16\}$, tandis que ε est choisi dans un ensemble de trois petites valeurs *{0,1, 0,01, 0,02}*. L'histogramme est finalement normalisé pour un traitement ultérieur.

Les histogrammes des oBIFs sont calculés pour décrire des images de textes manuscrits et imprimés. En faisant varier les paramètres dans le calcul des caractéristiques, différentes configurations d'OBIFs sont produites.

3.5. Résultats et discussion

Dans cette section, nous présentons et analysons les performances des caractéristiques proposées dans l'identification de scripts en utilisant des documents imprimés et manuscrits.

Lors de la première série d'expérimentations, trois images de chacun des scripts considérés sont utilisées pour l'apprentissage tandis qu'une autre est utilisée dans le test. Nous avons envisagé deux

scénarios d'évaluation différents. Dans le premier, nous avons choisi 300 images par scripts (13 scripts) pour l'apprentissage et 100 images par script (13 scripts) pour le test. Tandis que dans le deuxième scénario, nous avons choisi 67% des images de chaque script pour l'apprentissage et 33% des images restantes de chaque script pour le test. Le tableau 3.1 présente les taux globaux d'identification enregistrés en utilisant les séparateurs à vaste marge multi-classes (SVM) en se basant sur les résultats des deux scénarios considérés.

Type d'image	Scénario I	Scénario II
Imprimé	90.38%	91.64%
Manuscrit	92.00%	90.69%

Tableau 3.2. Taux d'identification réalisés pour les deux scénarios.

Les résultats obtenus pour l'identification de scripts démontrent clairement le potentiel des caractéristiques proposées pour la reconnaissance des scripts à partir des documents imprimés et manuscrits.

3.6. Conclusion

Ce travail avait pour objectif de présenter une méthode pour l'identification de scripts à partir de documents imprimés et manuscrits. Nous avons utilisé des caractéristiques texturales qui ont montré des résultats prometteurs sur une base de données de documents imprimés et manuscrits. Les évaluations ont été effectuées sur une base de données contenant des échantillons de textes imprimés et manuscrits contenant 10.400 images différentes. Les résultats obtenus pour l'identification de scripts sont encourageants. Ils reflètent l'efficacité des caractéristiques texturales sur des documents imprimés et manuscrits.

La contribution que nous avons proposée dans le cadre de ce mémoire nous ont permis d'aboutir à des résultats prometteurs, mais nous ont aussi ouvert plusieurs voies pouvant être exploitées dans le futur. Les études ultérieures sur ce sujet seront destinées à introduire

des caractéristiques supplémentaires et ensuite appliquer un mécanisme de sélection des caractéristiques pour savoir quelles sont les caractéristiques les plus discriminantes pour ce problème et pour des problèmes similaires. Il est nécessaire de rappeler que la performance du système proposé ne dépend pas seulement des techniques de classification utilisées, mais aussi des caractéristiques choisies.

Dans ce cadre, il serait très intéressant d'exploiter la combinaison des plusieurs caractéristiques avec celles de l'état de l'art afin d'améliorer les performances du système proposé. Pour les techniques de classification utilisées, nous pensons qu'il serait intéressant d'envisager l'utilisation d'autres techniques de classification que celle que nous avons adoptée dans le présent mémoire. Il serait aussi très intéressant aussi d'envisager et d'expérimenter des possibilités de combinaison des techniques de classification et de considérer des bases de données plus volumineuses que celle utilisée dans le cadre du présent travail.

Bibliographie :

[01] Z. Yuan, H. Wang, L. Wang, T. Lu, S. Palaiahnakote, C.L. Tan, Modeling spatial layout for scene image understanding via a novel multiscale sum-product network, Expert Systems with Applications, 63 (2016) pp.231-240.
[02] J. He, J. Feng, X. Liu, T. Cheng, T.H. Lin, H. Chung, S.F. Chang, Mobile product search with bag of hash bits and boundary reranking. In Computer Vision and Pattern Recognition (CVPR), IEEE Conference on IEEE, (2012) pp.3005-3012.
[03] V. Khare, P. Shivakumara, P. Raveendran, A new Histogram Oriented Moments descriptor for multi-oriented moving text detection in video, Expert Systems with Applications, 42(21) (2015) pp.7627-7640.
[04] V. Alabau, A. Sanchis, F. Casacuberta, Improving on-line handwritten recognition in interactive machine translation, Pattern Recognition, 47(3) (2014) pp.1217-1228.
[05] https://sites.google.com/view/ICDAR21-SIW2021/home

[06] Newell, A.J., Griffin, L.D.: Writer identification using oriented basic image features and the delta encoding. Pattern Recogn. 47(6), 2255– 2265 (2014)
[07] Newell, A.J., et al.: Texture-based estimation of physical characteristics of sand grains. In: 2010 International Conference on Digital Image Computing: Techniques and Applications, pp. 504– 509 (2010)
[08] Gattal, A., et al.: Isolated handwritten digit recognition using obifs and background features. In: 2016 12th IAPR Workshop on Document Analysis Systems (DAS), pp. 305– 310 (2016)
[09] Gattal, A., et al.: Gender classification from offline multi-script handwriting images using oriented basic image features (obifs). Expert Syst. Appl. 99, 155– 167 (2018)
[10] Abdeljalil, G., et al.: Writer identification on historical documents using oriented basic image features. In: 2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR), pp. 369– 373 (2018)
[11] Griffin, L.D., et al.: Basic image features (bifs) arising from approximate symmetry type. In: International Conference on Scale Space and Variational Methods in Computer Vision, pp. 343– 355 (2009)
[12] Griffin, L.D., Lillholm, M.: Symmetry sensitivities of derivative-of-Gaussian filters. IEEE Trans. Pattern Anal. Mach. Intell. 32(6), 1072– 1083 (2009)

CONCLUSION GENERALE

Ce travail a examiné la possibilité de n'utiliser que l'analyse globale des scripts pour les identifier. Nous avons présenté un ensemble des caractéristiques basées sur l'analyse de la texture des images de documents pour accomplir une classification de manière supervisée. Ces caractéristiques ont été utilisées pour développer un système d'identification de scripts. Le système est très simple et pratique et ne nécessite aucun prétraitement. Les résultats enregistrés du système proposé ont révélé qu'une bonne précision des performances (92%) peut être obtenue à l'aide d'une analyse globale, illustrant ainsi sa force et son utilité. Le système peut être étendu à plusieurs échelles pour gérer les scripts imprimés et manuscrits à une résolution différente. Le schéma proposé peut également être utilisé pour d'autres scripts et langues avec une modification minimale.