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 Multimédias

***Thème* :**

Datation des document manuscrits historiques par l'apprentissage profond

Présenté par :

Sakout Soufiene

Devant le jury :

Mr. Bendib Issam	*Université de Tébessa*	*Président*
Mr. Abdeljalil Gattal	*Université de Tébessa*	*Examinateur*
Mr. Chawki Djeddi	*Université de Tébessa*	*Encadrant*
Mr. Abdellatif Gahmousse	*Université de Tébessa*	*Co-Encadrant*

Année Universitaire : 2022-2023

Résumé

Le mémoire aborde la problématique de la datation des documents historiques, en proposant une approche basée sur l'extraction de caractéristiques locales et globales. Avec l'utilisation des méthodes SIFT et ORB pour extraire les caractéristiques locales des documents, et un modèle CNN pré-entraînés (VGG) pour extraire les caractéristiques globales. Les vecteurs issus de ces méthodes sont utilisés pour créer des vecteurs de représentation de document. Ces vecteurs sont ensuite utilisés par un modèle PyTorch pour évaluer les performances de la datation des documents historiques. L'étude expérimentale est réalisée sur l'ensemble de données KERTAS. L'objectif est d'améliorer la précision de la datation des documents historiques.

Mots clés : Document historiques, datation, apprentissage profond, extraction des caractéristiques, caractéristiques locales, caractéristiques globales, SIFT, ORB, CNN, VGG, modele tabulair, PyTorch, KERTAS.

Abstract

This work addresses the problem of dating historical documents by proposing an approach based on the extraction of local and global document features. We use SIFT and ORB methods for extracting local features from the images, and a pre-trained CNN model (VGG) for extracting global features. The vectors obtained from these methods are then used to create documents representation vectors. These vectors are used by a PyTorch model to evaluate the performance of historical document dating. The experimental study is conducted on the KERTAS dataset. The objective is to improve the accuracy of historical document dating.

Key words : historical documents, dating, deep learning, feature extraction, local features, global features, SIFT, ORB, CNN, VGG, tabular model, PyTorch, KERTAS .

ملخص

هذا العمل يتناول مشكلة تحديد تاريخ الوثائق و المستندات التاريخية من خلال اقتراح نظام يستند على استخراج الميزات الخاصة والعامة للمستند. يستخدم هذا النظام طريقة SIFT وORB لاستخراج الميزات الخاصة من المستند، ونموذج CNN مدرب مسبقًا (VGG) لاستخراج الميزات العامة. يتم العمل بالميزات المتحصل عليها لإنشاء تمثيل للوثيقة. يتم استخدام النتائج المدمجة بواسطة نموذج خاص PyTorch لتقييم أداء النظام المدرب في تحديد تاريخ الوثائق التاريخية. يتم إجراء الدراسة التجريبية على مجموعة البيانات KERTAS. الهدف هو تحسين دقة تحديد تواريخ الوثائق التاريخية.

الكلمات المفتاحية: الوثائق التاريخية، تحديد التاريخ، استخراج الميزات، الميزات الخاصة، الميزات العامة، SIFT، ORB، التعليم المعمق، النموذج الجدولي، PyTorch، KERTAS.

Sommaire

Abréviations et acronymes

DAR	Analyse et reconnaissance de documents (Document Analysis and Recognition)
HR	Reconnaissance de l'écriture manuscrite (Handwriting Recognition)
LS	Segmentation de ligne (Line Segmentation)
TD	Détection de tableau (Table Detection)
CR	Reconnaissance de caractères (Character Recognition)
LLA	Analyse logique de la mise en page (Logical Layout Analysis)
HMM	Modèle de Markov caché (Hidden Markov Model)
LBP	Modèle binaire local (Local Binary Pattern)
SVM	Soutenir la machine vectorielle (Support Vector Machine)
CNN	Réseau de neurones convolutifs (Convolutional Neural Network)
DNN	Réseau neuronal profond (Deep Neural Network)
BRNN	RNN bidirectionnel (Bi-directional RNN)
MDCC	Classification connexionniste multidimensionnelle (Multi-Dimensional Connectionist Classifica¬tion)
FOTS	Repérage de texte orienté rapide (Fast Oriented Text Spotting)
YOLO	Vous ne regardez qu'une seule fois (You Look Only Once)
DTW	Déformation temporelle dynamique (Dynamic Time Warping)
PSNR	Rapport signal/bruit maximal (Peak Signal-to-Noise Ratio)
Pr	Précision (Precision)
TP	Vrai positif (True Positive)
EDM	Métrique de détection d'entité (Entity Detection Metric)
PAW	Partie de mots arabes (Part-of-Arabic-words)
BLD	Détection de base (Baseline Detection)
GC	Classement par sexe (Gender Classification)
HTAR	Analyse et reconnaissance de texte manuscrit (Handwritten text analysis and recognition)
LD	Détection de ligne (Line Detection)
On	En ligne (Onine)
Binz	Binarisation (Binarization)
Rec	Reconnaissance (Recognition)
DU	Compréhension des documents (Document Understanding)
GLCM	Matrice de cooccurrence des niveaux de gris (Gray Level Co-occurance Matrix)
GSC	Concavité structurelle dégradée (Gradient structural concavity)
MLP	Perceptron multicouche (Multi-Layer Perceptron)
DNN	Réseaux de neurones profonds (Deep Neural Networks)
RNN	Réseau neuronal récurrent (Recurrent Neural Network)
LSTM	Longue mémoire à court terme (Long Short-Term Memory)
MSER	Régions extrêmes stables au maximum (Maximally Stable Extremal Regions)
ROI	Région d'intérêt (Region-of-Interest)
Seq2Seq	Séquence à séquence (Sequence-to-sequence)
NPC	Classificateur de prototype le plus proche (Nearest Prototype Classifier)
NRM	Mesure du taux négatif (Negative Rate Metric)
mAP	Précision moyenne moyenne (mean Average Precision)
TN	Vrai négatif (True Negative)
SM	Métrique de segmentation (Segmentation Metric)
DIA	Analyse d'images de documents (Document Image Analysis)
SI	Identification des scripts (Script Identification)
OCR	Reconnaissance optique de caractères (Optical Character Recognition)
PS	Segmentation des pages (Page Segmentation)
Off	Hors ligne (Offline)
IR	Récupération d'images (Image Retrieval)
Clsf	Classification (Classification)
CCA	Analyse des composants connectés (Connected Component Analysis)
HoG	Histogramme des dégradés (Histogram of Gradients)
k-NN	k Voisin le plus proche (k Nearest Neighbour)
NN	Réseau neuronal (Neural Network)
PCA	Analyse des composants principaux (Principal component analysis)

MDRNN	RNN multidimensionnel (Multi-dimensional RNN)
BLSTM	Mémoire à court terme bidirectionnelle (Bi-directional Long Short-Term Memory)
SOM	Cartes auto-organisées (Self-Organizing Maps)
RPN	Réseau de proposition de région (Region Proposal Network)
SDSW	Déformation spatiale dynamique statistique (Statistical Dynamic Space Warping)
DFE	extraction de caractéristiques discriminantes (discriminant feature extraction)
DRD	Métrique de distorsion réciproque de distance (Distance Reciprocal Distortion Metric)
GT	Vérité terrain (Ground Truth)
FP	Faux positif (False Positive)
CER	Taux d'erreur de caractère (Character Error Rate)
PMCOA	Sous-ensemble PubMed Central Open Access (PubMed Central Open Access Subset)
SR	Reconnaissance de scripts (Script Recognition)
DLA	Analyse de la mise en page du document (Document Layout Analysis)
WI	Identification de l'auteur (Writer Identification)
TNC	Classification texte/non-texte (Text/non-text Classification)
DR	Reconnaissance des chiffres (Digit Recognition)
WSeg	Segmentation des mots (Word Segmentation)
CJK	chinois japonais coréen (Chinese Japanese Korean)
DWT	Transformées en ondelettes discrètes (Discrete Wavelet transforms)
RBF	Fonction de polarisation radiale (Radial Bias Function)
ANN	Réseau neuronal artificiel (Artificial Neural Network)
FCN	Réseau entièrement convolutionnel (Fully Convolutional Network)
MSTDNN	Réseau de neurones à temporisation multi-états (Multi-State Time Delay Neural Network)
CTC	Classification temporelle connexionniste (Connectionist Temporal Classification)
RLSA	Algorithme de maculage de longueur d'exécution (Run length smearing algorithm)
SSD	Détecteur multi-boîtes à un seul coup (Single-shot Multi-box Detector)
MQDF	Fonction discriminante quadratique modifiée (Modified Quadratic Discriminant Function)
DLQDF	Fonction discriminante quadratique d'apprentissage discriminant (Discriminative Learning Quadratic Discrim¬inant Function)
MPM	Métrique de pénalité de mauvaise classification (Misclassification Penalty Metric)
B	Image binaire (Binary Image)
FN	Faux négatif (False Negative)
WER	Taux d'erreur de mot (Word Error Rate)
SV	Vérification de la signature (Signature Verification)
OD	Détection d'objet (Object Detection)

1- Liste des tableaux

2- Liste des figures

Introduction générale

Dans l'étude des manuscrits historiques, les chercheurs explorent couramment quatre questions importantes : quoi, par qui, quand et où. Les réponses à ces quatre questions aident à comprendre le contexte historique des manuscrits. Ce mémoire de fin d'etude se concentre sur la question du "quand", c'est-à-dire les dates des manuscrits. Estimer la date d'un manuscrit historique nécessite l'expertise des paléographes. Les paléographes s'appuient sur leurs connaissances et leurs expériences pour faire une estimation. Ce processus d'estimation prend en compte plusieurs aspects, notamment le style d'écriture, le contenu et même les matériaux d'écriture. Ce processus nécessite beaucoup de temps et d'efforts humains. De plus, en raison de la nature subjective de ces approches, il existe souvent des divergences d'opinions quant à l'estimation d'une date d'un manuscrit historique. Un système automatique basé sur des techniques modernes de reconnaissance serait un outil utile pour les paléographes, les aidant à évaluer des hypothèses ainsi qu'à en proposer de nouvelles.

L'objectif du travail

Présenter une contribution visant à améliorer la précision de la datation des documents historiques en combinant l'extraction des caractéristiques locales et globales du document. La méthode propose l'utilisation des techniques SIFT et ORB pour extraire les caractéristiques locales, et un modèles CNN pré-entraînés VGG pour extraire les caractéristiques globales. Les vecteurs descripteurs issus de ces différentes méthodes sont ensuite utilisés pour former des vecteurs de représentation des document. Ces vecteurs sont ensuite employés par un modèle CNN PyTorch pour évaluer les performances de cette approche de datation des documents historiques. L'ensemble de données KERTAS sera utilisé pour évaluer les performances de cette solution.

Organisation du mémoire

Ce mémoire est structuré en trois chapitres. Le premier chapitre se concentre sur la présentation des concepts et des outils clés liés à l'analyse et à la reconnaissance de documents (DAR). Dans le deuxième chapitre, nous examinons l'état de l'art de la datation des documents historiques, puis nous concluons avec le dernier chapitre qui présente notre contribution spécifique à la datation des documents historiques et les résultats obtenus par le système proposé.

Chapitre 01 : Concept et outils

1-1. Introduction

Le parcours de recherche pour l'Analyse et la Reconnaissance de Documents (DAR) a commencé avec le problème de la reconnaissance automatique de caractères. Aujourd'hui, il couvre une vaste gamme de tâches de reconnaissance telles que la reconnaissance de texte, l'identification de script, l'indexation de mots, la vérification de signatures, etc., dans des scripts tels que le latin, l'arabe, le chinois, le japonais, etc. Les avancées considérables dans les techniques d'apprentissage profond ont permis d'obtenir des résultats de pointe pour diverses tâches de (DAR). Dans ce chapitre, nous explorons les défis sous différents angles et passons en revue les techniques de (DAR) pour les documents en ligne / hors ligne et imprimés / manuscrits. Nous explorons les défis, les techniques, les ensembles de données, les mesures d'évaluation, les aspects liés au script, ainsi que les orientations futures possibles dans le domaine de l'analyse et de la reconnaissance de documents.

L'analyse et la reconnaissance de documents (DAR) désignent la capacité à localiser et transformer le contenu textuel d'images de documents en un format lisible par une machine. Cela concerne l'apprentissage des motifs dans les images de texte. La localisation du texte, l'extraction des caractéristiques et la classification des images de documents peuvent produire du texte transcrit pour une interprétation par machine. Le texte dans les images contient des informations sémantiques qui sont utiles pour diverses applications de reconnaissance de motifs et de vision par ordinateur, telles que la recherche d'images, l'inspection intelligente, l'automatisation industrielle, la navigation des robots, la traduction instantanée PS [1], la localisation des blocs d'adresses, la localisation des plaques d'immatriculation, l'indexation d'images/vidéos basée sur le contenu, etc. Au début même de la (DAR), des méthodologies artisanales étaient utilisées pour différentes tâches. Ces méthodes nécessitaient des opérations de prétraitement et de post-traitement spécifiques à l'application. De plus, les caractéristiques artisanales (handcrafted) étaient limitées par les motifs et les structures prédéterminés dans les données qu'elles pouvaient traiter. Ainsi, elles étaient incompétentes dans l'environnement convoluté et innommable.

La reconnaissance dépend des étapes préliminaires de localisation ou de segmentation du document en objets textuels tels que des lignes, des mots ou des caractères. Les images de documents peuvent être imprimées ou manuscrites, hors ligne ou en ligne. La perspective de la DAR pour les documents imprimés ne correspond pas à celle de la (DAR) pour les documents manuscrits. Les documents imprimés peuvent présenter un ensemble de caractères différent, des polices différentes ou des styles de mise en page différents, ainsi que des objets graphiques tels que des tableaux, des figures, des graphiques, des logos, etc. Le système pour les documents imprimés doit être capable de prendre en compte ces variations de documents. Cependant, la même approche peut être insuffisante pour la (DAR) des documents manuscrits en raison de la complexité supplémentaire liée aux défis de l'écriture à la main.

Différentes techniques sont nécessaires pour aborder les problèmes de la (DAR) pour les documents imprimés et manuscrits. L'analyse et la reconnaissance manuscrites en ligne constituent un problème totalement différent de la DAR hors ligne. Elle concerne la dynamique de l'écriture manuscrite. Elle capture les données séquentielles qui contiennent des signaux spatiaux et temporels du signal d'entrée. Cette analyse nous donne une meilleure compréhension des caractéristiques spécifiques de l'écriture manuscrite de l'auteur. Cependant, cet avantage est limité aux scripts romains et autres scripts occidentaux. Les scripts tels que le chinois comportent des milliers de caractères et ne suivent pas d'ordre d'écriture spécifique, ce qui rend la DAR en ligne pour de tels cas difficile. D'autres scripts, tels que les scripts asiatiques et du Moyen-Orient tels que l'arabe et le BRAHMI, ont leurs particularités respectives. Ces subtilités liées aux scripts jouent un rôle important dans la modélisation réussie des techniques pour diverses tâches de la DAR.

1.1.1. Systèmes d'écriture du monde

La documentation d'informations sous forme imprimée ou numérique est influencée par le système d'écriture de la manière suivante. On a les différents styles d'écriture associés à une ou plusieurs langues parlées à travers un système d'écriture. Dans le contexte actuel, il existe quatre systèmes d'écriture utilisés dans le monde entier [2], [3] :

1) Système logographique
2) Système syllabique
3) Système alphabétique (abjads, abugidas et alphabets purs)
4) Système à traits distinctifs

Les systèmes logographiques, tels que les scripts chinois, japonais et coréen (CJK), utilisent des symboles, des caractères et des mots pour représenter visuellement l'écriture. En revanche, dans un système syllabique, chaque syllabe est utilisée pour représenter un son phonétique. Les scripts japonais combinent à la fois des systèmes logographiques et syllabiques.D'autre part, les scripts comme le latin, le grec, le cyrillique, etc., utilisent le système alphabétique pour représenter les phonèmes d'une langue. Le script latin est couramment utilisé dans des langues telles que l'anglais, l'allemand, le français, l'espagnol, le portugais, etc. Quant au russe, à l'ukrainien, au bulgare, etc., ils utilisent le script cyrillique. Parmi ces scripts, le script latin est le plus répandu, couvrant l'Amérique du Nord et du Sud (à l'exception du Canada), l'Afrique de l'Ouest, centrale et australe, ainsi que l'Australie et certaines régions d'Europe et d'Asie du Sud-Est.

Le système abjad a un schéma d'écriture de droite à gauche, suivi par des langues comme l'arabe, le persan, l'ourdou, l'hébreu, etc. Ces scripts sont principalement utilisés dans la région du Moyen-Orient. Ce système ne représente que les consonnes et pas les voyelles. Les abugidas ont à la fois des représentations de consonnes et de voyelles, où les consonnes sont des représentations primaires et les voyelles des représentations secondaires. Les

systèmes d'écriture tels que la famille Brahmi en Asie du Sud, en Asie du Sud-Est et au Tibet avec des scripts comme les scripts indiens (devanagari, gourmoukhî, gujarati, bengali, manipuri, oriya, tamoul, télougou, kannada, malayalam) utilisent les abugidas. Dans les systèmes à traits distinctifs, les traits distinctifs des phonèmes sont les syllabes, par exemple le script coréen (hangul).

La classification des scripts ci-dessus explique l'ascendance d'un système d'écriture sur une langue ou un script. Cependant, les systèmes d'écriture du monde sont ambigus et dérivent de multiples systèmes d'écriture. Par exemple, l'anglais utilise des logogrammes tels que #, $, & etc., ce qui contraste avec les systèmes logographiques car ils ne correspondent pas à la langue parlée. Les subtilités liées aux scripts ont un impact énorme sur la (DAR).

La phase initiale de la DAR a vu l'utilisation de systèmes spécifiques aux scripts, avec des recherches approfondies sur les scripts chinois, japonais et arabes. La plupart des recherches portaient sur la reconnaissance des caractères manuscrits (HCR).

Les systèmes traditionnels de DAR sont conçus pour modéliser des facteurs tels que la mise en page des documents, l'espacement des lignes, la conception et la densité des informations, etc. Nous devons développer des systèmes génériques automatiques et semi-automatiques capables de s'adapter à de nouveaux scripts et à de nouveaux scénarios. Ces systèmes doivent apprendre les particularités des différents scripts pour modéliser des systèmes intelligents. Ils doivent prendre en compte les complexités des scripts, notamment:

1) les ensembles logographiques des scripts CJK sont très différents des alphabets et des syllabes des systèmes alphabétiques tels que les scripts romains, arabes et brahmi. Il est difficile de modéliser de telles variétés dans un système générique unique.

2) Les voyelles et les signes diacritiques d'un script peuvent introduire des motifs confus même pour un modèle conçu pour un script spécifique. Cependant, le caractère non superposé est une caractéristique commune à tous les scripts, un avantage pour les systèmes de reconnaissance.

Contrairement aux approches antérieures de la reconnaissance, récemment, des systèmes sont conçus pour le problème global de la (DAR) afin de répondre à la nécessité d'un système (DAR) générique et robuste.

1.2. Analyse et reconnaissance de documents (DAR)

L'analyse et la reconnaissance de documents (DAR) ont été initialement conçues comme un problème de reconnaissance automatique de caractères axé sur du texte imprimé avec une police spécifique. La collecte automatique de données, l'identification d'objets et La reconnaissance optique de caractères sans intervention humaine étaient appelées OCR [4], [5]. Un dispositif mécanique capable de reconnaître un seul caractère à la fois était utilisé

pour l'OCR. Un photodétecteur comparait l'entrée avec les modèles stockés, reconnaissant ainsi uniquement quelques petits ensembles de polices de caractères ou de documents manuscrits. En raison des contraintes de stockage, des algorithmes d'analyse de structure plus performants étaient proposés pour l'OCR. Cependant, ces approches étaient limitées par leur puissance de traitement et leurs options d'acquisition d'images. Avec l'évolution des technologies de l'information, des dispositifs d'acquisition plus précis tels que des scanners, des téléphones mobiles et des méthodes modernes d'extraction et de reconnaissance de caractéristiques, la capacité des OCR modernes sur les documents numérisés a dépassé les 99 %. Les OCR peuvent convertir des fichiers PDF, des documents scannés ou des images de documents capturées par appareil photo en formats textuels lisibles par machine. Les techniques des systèmes OCR modernes sont dérivées de la reconnaissance de motifs, de l'apprentissage automatique, de l'apprentissage profond et d'autres techniques [6]. Les résultats générés par l'OCR permettent la recherche et la modification du texte. Des OCR dédiés et câblés sont utilisés pour des problèmes de reconnaissance spécifiques tels que la lecture de chèques, la numérisation de codes-barres, etc.

Des moteurs OCR commerciaux prêts à l'emploi tels que Tesseract [7], EasyOCR, Ocropus, Ocular, Attention OCR, Doctr, etc., offrent des performances exceptionnelles. Tous les systèmes de pointe supposent des documents d'entrée de bonne qualité et sont conçus pour des applications génériques, ce qui les rend inadaptés aux problèmes de reconnaissance hétérogènes et non contraints (documents spécifiques à une entreprise, documents manuscrits et historiques). Les techniques de (DAR) pour les environnements non contraints sont supérieures aux systèmes OCR, bien que ces derniers aient progressé dans la reconnaissance de problèmes d'impression et de reconnaissance manuscrite (hors ligne). L'association des OCR avec les techniques modernes d'apprentissage automatique et d'apprentissage profond peut aboutir à un système de reconnaissance au potentiel immense.

Dans ce chapitre, nous avons classé les approches de la Reconnaissance Automatique de Documents (DAR) en méthodologies, celles-ci sont discutées en détail dans la section 4. Certaines méthodes de (DAR) effectuent un traitement étape par étape et nécessitent la pré-segmentation des documents en lignes, mots ou caractères pour une reconnaissance correspondante. Cela implique de diviser le document en primitives différentes, puis de reconnecter ces primitives en segments souhaités. Le problème de segmentation a débuté avec la détection et la reconnaissance des caractères. Le développement des systèmes de reconnaissance de caractères a intensifié le problème de segmentation des caractères autour des années 1970. Cela a suscité la nécessité de la segmentation de page (PS) pour extraire les lignes des pages et, par conséquent, les caractères de ces lignes. Avec le développement de structures de documents sophistiquées, ce problème a dépassé la reconnaissance des lignes et des mots, où les mises en page des documents peuvent être diverses (documents scientifiques, lettres commerciales, formulaires officiels, etc.). De plus, les documents peuvent avoir une mise en page complexe (texte imprimé et manuscrit). La tâche de segmentation dépend des mises en page et des types de documents qu'elle traitera. Les défis liés au contenu textuel comprennent les documents

bruyants, le texte incliné, le chevauchement/le contact entre les contenus et le texte courbé ou brisé. Ces défis ont été discutés dans la section III. Parmi les problèmes de segmentation, on trouve la segmentation de page (PS), la segmentation de ligne (LS), la segmentation de mot (WS) et la segmentation de caractère (CS). Le problème de la segmentation de ligne et de mot est relativement plus facile que les deux autres.

Les méthodologies de segmentation et de reconnaissance intégrées ont ouvert la voie à la reconnaissance automatique des documents texte. Les techniques de reconnaissance de bout en bout combinent la segmentation, l'extraction de caractéristiques et la classification en une seule chaîne de traitement et prennent en entrée des pages complètes ou des paragraphes. Récemment, l'attention s'est portée sur la formation d'un seul système de bout en bout pour préserver l'information tout au long du processus afin d'optimiser les différentes tâches intermédiaires. Cependant, l'optimisation conjointe de l'ensemble du processus de reconnaissance est un défi. Des structures de réseaux de neurones récurrents profonds telles que LSTM, BLSTM, des schémas basés sur l'attention, des architectures encodeur-décodeur, la détection d'objets (OD) et des méthodes sans segmentation ont été proposées pour résoudre les problèmes mentionnés ci-dessus. Ces méthodes sont discutées dans la section 4.

Des efforts antérieurs ont été déployés pour rendre les modèles séquentiels en combinant le processus d'extraction des caractéristiques avec des modèles ultérieurs tels que les modèles de séquence HMM (Modèles de Markov Cachés) et les modèles de représentation CTC (Connectionist Temporal Classification). Cependant, ces modèles étaient entraînés indépendamment et traitaient des entrées déjà segmentées (lignes ou mots).

1.3. Défis

l'évolution des techniques de traitement et de stockage, une plus grande diversité de problèmes de Reconnaissance Automatique de Documents (DAR) a été abordée. Chaque document présente ses propres complexités. Cependant, étant donné que tous les documents à traiter sont des images, l'acquisition d'images et les complications qui y sont liées viennent s'ajouter au problème de la DAR. Les techniques se sont spécialisées dans des tâches spécifiques et ont atteint des résultats de pointe. Cependant, cela devient un défi lorsque le problème de la DAR est considéré dans son ensemble. Il y a de nombreux obstacles [8] sur la voie de l'évolution de la DAR. Dans ce travail, nous examinons les défis de la DAR sous trois perspectives :

1. Défis de la DAR : couvre diverses complications liées à la reconnaissance des images de documents. Celles-ci peuvent être dues à des textes dans différentes langues, polices et styles, à la complexité du processus de reconnaissance et au manque de formats standard.
2. Défis de la reconnaissance en ligne : examine les difficultés liées aux dispositifs électroniques et aux technologies sous-jacentes telles que la traduction automatique, la ROC, etc.
3. Défis liés au texte dans les images de documents : aborde les problèmes d'acquisition d'image et de stockage liés aux différents types de textes en ligne ou hors ligne, manuscrits ou imprimés. La Figure 1 présente quelques exemples de défis classés en fonction de la variation de contenu dans les documents, de l'acquisition d'image et des défis liés à l'environnement. Ces défis sont discutés comme suit.

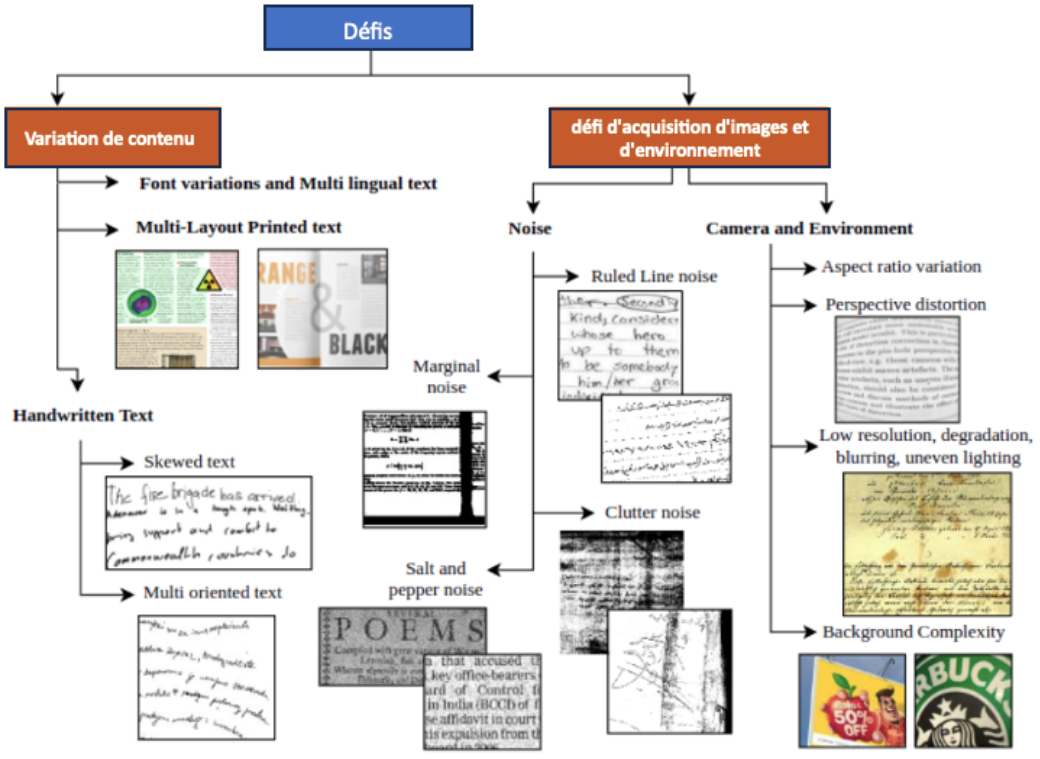


Figure 1 : Défis liés à l'analyse et à la reconnaissance de documents

1.3.1. Défis liés à la (DAR)

1) Environnements multilingues : Le texte peut être écrit dans différents systèmes d'écriture et langues. L'anglais a été la première et la plus réussie des langues choisies pour les OCR. Les langues telles que le chinois, le japonais, l'arabe et l'hindi posent de plus grands défis aux systèmes OCR [10]. Il existe des milliers de formes de caractères pour le chinois et le japonais. L'arabe est écrit en utilisant des composants connectés dont les formes varient en fonction de la position du caractère. La plupart des recherches se concentrent sur un modèle spécifique à une langue plutôt que sur le développement d'une solution multilingue. Ainsi, la création d'un OCR générique qui fonctionne bien pour tous est un défi.

2) Variations de polices : Il existe de nombreuses variétés de polices de caractères, chacune représentant une manière unique de représenter le même caractère. Il est difficile pour l'OCR de reconnaître avec précision toutes les variations de polices lorsque de nombreux caractères appartiennent à une même classe. Certains systèmes d'écriture, comme le système romain, présentent des caractères qui se chevauchent, ce qui rend la tâche de segmentation plus difficile [11], [12].

3) Texte manuscrit [13] : Chaque individu a un style et un modèle d'écriture uniques. De plus, l'écriture manuscrite d'une personne peut varier dans le temps. Il n'est pas facile de concevoir une technique pour détecter et reconnaître le texte pour de nombreuses raisons. La ligne de base du texte manuscrit est inclinée, ce qui nécessite une détection de la ligne de base (BLD) et une normalisation de l'inclinaison. Le texte penché augmente encore la complexité du système de détection et de reconnaissance du texte. Les mouvements de main erratiques lors de l'écriture introduisent davantage de variations dans l'écriture d'un individu.

4) Dépendance aux étapes intermédiaires : Les étapes préliminaires de la reconnaissance telles que le prétraitement, la segmentation et la classification ont un impact sur l'ensemble du processus de reconnaissance.

5) Manque d'outils linguistiques : La recherche en (DAR) est très spécifique aux systèmes d'écriture, et seuls certains systèmes, tels que le système romain, chinois et les systèmes d'Asie du Sud, ont connu des développements majeurs. Il est nécessaire de disposer de modèles linguistiques et d'outils linguistiques pour différents systèmes d'écriture afin qu'un modèle générique unique puisse prendre en charge davantage de systèmes d'écriture.

6) Difficulté des systèmes automatiques : Les systèmes génériques conviviaux et automatiques sont difficiles à réaliser. Avec des outils et des mécanismes puissants, on peut s'attendre à ce que les systèmes génériques pour des tâches de (DAR) triviales soient bientôt réalisables. Cependant, ces systèmes ne peuvent pas être entièrement automatiques et nécessiteraient un certain post-traitement pour produire les résultats attendus.

7) Besoin de normalisation : Le manque de formats standard de documents imprimés/manuscrits/en ligne rend le processus de (DAR) difficile. Il existe d'innombrables mises en page complexes de documents qui sont créées, et qui plus est, avec des sémantiques diverses. Les systèmes ne peuvent pas être formés pour gérer de telles variations considérables.

1-3.2. Défis liés à la reconnaissance en ligne [14]

1) Particularités de l'écriture manuscrite : L'écriture manuscrite est une caractéristique individuelle ; chaque individu a des traits d'écriture uniques. Cette singularité entraîne de grandes variations qui doivent être modélisées par le système de reconnaissance.

2) Facteurs comportementaux et personnels : Le comportement d'un individu affecte également les motifs d'écriture. Par exemple, le stress, l'excitation, la distraction, la paresse, etc., entraînent des changements significatifs dans la position et les traits de l'écriture.

3) Systèmes dépendants et indépendants de l'écrivain : La rareté des données d'entrée conduit à des systèmes qui ont appris certains styles de données spécifiques et deviennent ainsi dépendants de l'écrivain. Un système indépendant de l'écrivain nécessite de disposer de grandes quantités de données provenant de différents écrivains. Cela constitue un obstacle majeur dans le développement des systèmes de reconnaissance en ligne.

4) Idiosyncrasies des différents systèmes d'écriture : Les particularités caractéristiques des systèmes d'écriture constituent un défi supplémentaire pour la tâche de reconnaissance. Par exemple, des caractères comme ‘d’ et ‘l’ peuvent confondre le reconnaissant car ils ont des motifs de traits similaires.

5) Difficultés d'apprentissage pour les systèmes d'écriture avec de nombreux caractères : Pour les systèmes d'écriture comme le chinois, le japonais et le coréen, il est difficile de les entraîner et d'obtenir des précisions plus élevées.

1-.3.3. Défis liés au texte dans les images de documents

1) Bruit : Le bruit dans les images de texte correspond à la dégradation du contenu, qui peut être causée soit par une dégradation physique, soit par une dégradation due à la numérisation.
Les dispositifs de numérisation introduisent également du bruit dans l'image [15]. De nombreuses recherches ont classifié le bruit en différents types [16] tels que :

- Bruit de lignes régulières.
- Bruit marginal.
- Bruit de fouillis.

- Bruit de motif similaire à des traits.
- Bruit sel et poivre.

Le bruit de lignes régulières se produit en raison des lignes de règle présentes dans les documents. Il peut être épais/mince, cassé et fusionné avec les caractères comme L, Z. Le bruit marginal résulte de la numérisation du document et apparaît sur les côtés des images de document. Le contenu indésirable en avant-plan des images correspond au bruit de fouillis, généralement causé par des trous de perforation, du bruit sel et poivre, etc. Il rend la segmentation du texte difficile, entraîne une mauvaise connectivité du texte ou fait se chevaucher le texte. En 2011, le travail présenté dans [17] a introduit le bruit de motif similaire à des traits et a proposé une solution. Le bruit sel et poivre est causé par des saletés lors de la conversion des documents.

2) Défis liés à la caméra et à l'environnement : Comme les images sont capturées à partir de différentes sources telles que des appareils photo numériques, des webcams, des téléphones portables, des scanners, etc., elles souffrent d'une faible résolution, de flou, de distorsion de perspective, de complexité de l'arrière-plan, d'un éclairage inégal, de flou et de dégradation, ainsi que de rapports d'aspect. Par exemple, les images basées sur un appareil photo donnent une meilleure reconnaissance que les entrées basées sur un scanner [18]. En revanche, les appareils photo numériques sont très pratiques pour l'acquisition d'images, mais ils entraînent d'autres défis tels que la distorsion géométrique, la perte de mise au point et l'éclairage inégal du document. La faible résolution, le flou et l'éclairage inégal rendent même la détection de texte simple difficile. La distorsion de perspective se produit lorsque l'image n'est pas capturée sur un plan parallèle à l'image. Les images résultantes ont des caractères déformés. Les images capturées à partir d'une surface non plane entraînent des lignes courbées. Les images compressées posent des défis pour la reconnaissance de texte car elles ont tendance à perdre la netteté requise.

Un système générique d'analyse et de reconnaissance de texte pourrait résoudre les défis mentionnés ci-dessus. Les premières recherches utilisaient généralement des méthodes manuelles d'extraction de caractéristiques. Ces méthodes ne pouvaient couvrir qu'une partie du problème. L'accent est mis sur ces défis individuellement. Cependant, ces derniers temps ont connu des développements rapides dans la reconnaissance de formes et le traitement d'images. Les motivations derrière ce développement sont multiples.

- Systèmes informatiques haute performance.
- Augmentation des applications.
- Réseaux de reconnaissance à grande échelle.

Figure 2 : Caractères de différentes écritures avec leurs modificateurs de voyelle

1.3.4. Taches d'analyse et de reconnaissance de documents

Les DARs ont été abordées dans la littérature par différentes méthodes et systèmes. Cependant, toutes les méthodes peuvent être classées en trois catégories de tâches :

1. Étape de prétraitement élimine les éléments indésirables et améliore la qualité des images.
2. Les étapes de segmentation traitent les documents prétraités et extraient divers éléments de texte et de non-texte.
3. La reconnaissance de texte effectue l'extraction de caractéristiques et la classification.

Le résultat de la reconnaissance nécessite un post-traitement pour produire un texte UNICODE lisible par machine. La Figure 2 présente les différentes tâches du processus de (DAR) et leurs interrelations. La Figure 2 présente un organigramme pour représenter le flux des différentes tâches de (DAR) et leurs interrelations.

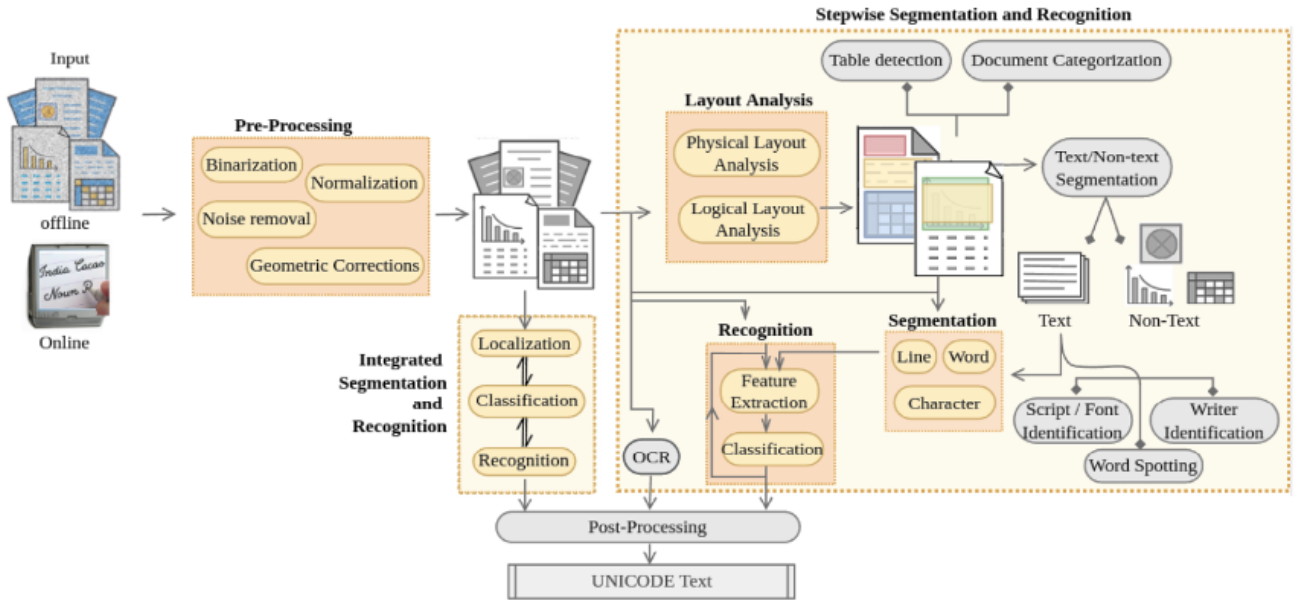


Figure 3 : Tâches d'analyse et de reconnaissance de documents avec des catégories principales[155]

1.4.1. Prétraitement

La première étape vers un système d'analyse et de reconnaissance de texte est le prétraitement des images [22]. Les techniques de prétraitement éliminent le bruit indésirable et d'autres problèmes dans l'image avant qu'ils ne soient alimentés dans le système d'analyse et de reconnaissance [22]. Il améliore les images et les convertit dans un format adapté [5]. Il gère les défis liés à l'acquisition d'images, comme discuté dans la section I. Le but principal d'un prétraitement est de produire une image prête pour l'extraction de caractéristiques de manière à améliorer la capacité de reconnaissance du système de reconnaissance de texte. Les images nécessitent une correction des artefacts qui s'écartent des normes et des méthodes d'amélioration visant à optimiser l'image pour une architecture spécifique. Différents chercheurs ont catégorisé différemment les sous-problèmes du prétraitement. Le prétraitement des documents manuscrits en ligne concerne la suppression du bruit dû au capteur ou à la résolution du dispositif numérique, aux traits abrupts [9] et aux points aléatoires.

1) Correction d'image

a) Réduction du bruit : Les méthodes de réduction du bruit [23] peuvent être envisagées de manière générale selon les points suivants :

- Les filtres.
- Les opérations morphologiques.
- Les transformations de Hough.
- Le profilage par projection.
- La méthode de réduction du bruit.

b) Normalisation : La normalisation consiste à corriger le texte pour éliminer toutes les variations présentes, principalement pour les textes manuscrits et numérisés. Les techniques de normalisation sont utilisées pour la correction de l'inclinaison, la normalisation de l'inclinaison, le lissage des contours et la normalisation de la taille. Les profils de projection, le regroupement des voisins les plus proches, la méthode de corrélation croisée entre les lignes et les transformations de Hough sont utilisés pour l'extraction de la ligne de base. Une fois la ligne de base détectée, des rotations sont appliquées. Cette solution est applicable dans différentes séquences pour la normalisation de l'inclinaison, de l'inclinaison et de la taille. Le lissage des contours réduit le nombre de points d'échantillonnage nécessaires pour représenter les données.

c) Correction de l'éclairage : Le problème des ombres, de la distribution instable et inégale de l'éclairage et des intensités de couleur inégales relève des défis liés à l'environnement. La décomposition en mode empirique basée sur les lignes [26], les modèles variationnels non locaux [27] et d'autres méthodes utilisant le filtrage sont utilisés pour la correction de l'éclairage [28].

d) Corrections géométriques : La rotation automatique de l'image, la correction de la perspective et les transformations d'image non linéaires (texte incurvé) font intervenir des méthodes de détection de la ligne de base (BLD) pour extraire la ligne de base, puis appliquer des rotations, etc., pour des corrections ultérieures [22].

2) Amélioration de l’image

a) Amélioration de l'éclairage, du flou et de la mise au point : Les images peuvent être améliorées en utilisant des méthodes d'éclairage global comme la remappage simple LUT, les opérations de points de pixels, l'égalisation d'histogramme et le remappage des pixels, ou des méthodes d'éclairage local comme les filtres de gradient, l'égalisation d'histogramme local et les filtres de rang.

b) Estimation de l'arrière-plan : L'estimation de l'arrière-plan est le plus souvent accompagnée de la binarisation des images. Les techniques classiques de binarisation comme Niblack [29] et Sauvola [30] utilisent une binarisation initiale à rappel élevé suivie du remplissage des régions avant-plan.

c) Conversion en niveaux de gris : Une image en niveaux de gris est généralement une condition préalable à la plupart des méthodes de binarisation. Une conversion de base de RVB en niveaux de gris consiste à calculer la luminosité par g = 0,21r + 0,72v + 0,07b. Cependant, cela n'est pas fructueux lorsque plusieurs couleurs sont présentes à l'avant-plan.

3) Compression de l’image

Les techniques de compression permettent de transformer une image du domaine spatial vers le domaine spatial afin de conserver les informations de forme. L'image compressée a une taille réduite permet de réduire la complexité de calcul du seuillage du réseau d'extraction et de reconnaissance des caractéristiques et d'affiner les méthodes de compression.

Le prétraitement implique principalement des techniques de traitement d'image pour diverses corrections et opérations d'amélioration. Ces techniques d'image dépendent largement des méthodes d'extraction de caractéristiques et de l'application des images, car chaque modèle d'extraction a un objectif différent en fonction de son application. Certaines méthodes de prétraitement basées sur l'application sont : les descripteurs de caractéristiques binaires locaux (comparaison d'intensité des pixels), les descripteurs spectraux, par exemple SIFT/SURF, pour les méthodes utilisant des pyramides d'images, les descripteurs de base, par exemple les méthodes de Fourier, les ondelettes, la transformée de Slant, la transformée de Walsh-Hadamard, la transformée KLT. Ces méthodes transforment les données dans un autre domaine pour l'analyse. Les descripteurs de forme polygonale sont utilisés pour extraire les formes des images. Cependant, ils ne sont pas utiles pour les images de documents texte.

4) Binarisation

La binarisation des documents consiste à diviser une image de document en deux groupes : les pixels de premier plan et les pixels d'arrière-plan. Les pixels de premier plan sont transformés en noir et le reste en blanc. Le processus de binarisation facilite l'analyse des documents en se concentrant sur les parties de données significatives. Cette étape est cruciale pour les documents fortement dégradés tels que les documents historiques. La dégradation peut être due à des arrière-plans complexes, des couleurs et des tailles de police multiples, des taches et des plis, etc. La binarisation peut être réalisée globalement ou localement. Une seule valeur est calculée pour la binarisation globale, tandis que les méthodes locales utilisent la binarisation adaptative. L'un des algorithmes les plus basiques est celui d'Otsu [31].

1.4.2. Segmentation de page

La segmentation d'un document entraîne des composants physiques et logiques. Un problème crucial lié à la segmentation de page est l'analyse de la mise en page du document (Document Layout Analysis - DLA). Avant la segmentation du texte, il est nécessaire de comprendre la mise en page du document et d'extraire le composant textuel pour une segmentation ultérieure [8]. Il n'existe pas de procédure d'analyse de mise en page générale, car les types de documents sont variés. Sur la base de ces informations, il existe deux niveaux de segmentation. Tout d'abord, un document est segmenté en régions de texte et de non-

texte (graphiques, tableaux, etc.). Ensuite, la région de texte est segmentée en paragraphes, lignes, mots et caractères.

Les facteurs affectant l'analyse de la mise en page peuvent être résumés comme suit :

- Propre vs Dégradé.
- Variation des lignes de texte.

- DLA, Pour commencer l'analyse de la mise en page (DLA), il est essentiel de faire certaines hypothèses sur les informations de base. Ces informations peuvent être classées en cinq catégories :
    - Document analysé : Imprimé ou manuscrit.
    - Objet analysé : Avant-plan ou arrière-plan.
    - Stratégie d'analyse : Ascendante (bottom-up) ou descendante (top-down).
    - Mise en page de l'analyse : Superposée ou non superposée
    - Primitifs d'analyse : Pixels, CC, etc.

La Figure 4 montre la catégorisation des différentes approches d'analyse de la mise en page physique et logique telles que discutées ci-dessus. Elle étend également l'analyse de la mise en page physique à l'analyse de la mise en page logique (LLA), qui concerne la classification des objets et la détection des relations.

Les méthodes d'apprentissage automatique telles que SOM [36] et SVM [37] ont été utilisées pour l'analyse de la mise en page avec une attention particulière portée aux ANN [38]. L'approche de segmentation, du niveau de la page aux niveaux inférieurs tels que les paragraphes, les figures, les tableaux, etc., est appelée top-down. Certaines méthodes utilisant cette approche sont : les méthodes basées sur la projection, l'analyse des espaces blancs [39], le RLSA [40], etc.

Avec toutes les différentes méthodes de segmentation, aucune méthode générique n'est applicable dans tous les cas. Avant de choisir une stratégie, une hypothèse sur le contenu est nécessaire pour de meilleurs résultats. Par exemple, les documents avec une mise en page rectangulaire simple peuvent être segmentés avec une grande précision par des algorithmes tels que le diagramme de Voronoi [35], Docstrum [41], etc. Les algorithmes coûteux en termes de calcul, tels que ceux basés sur l'estompage [40] et ceux basés sur la projection, peuvent être appliqués à une version échantillonnée de documents haute résolution. Les méthodes moins coûteuses en termes de calcul, telles que celles basées sur les composantes connexes: MST [42], [43] et Voronoi [35], sont préférables lorsque les primitives sont plus grandes dans le document.

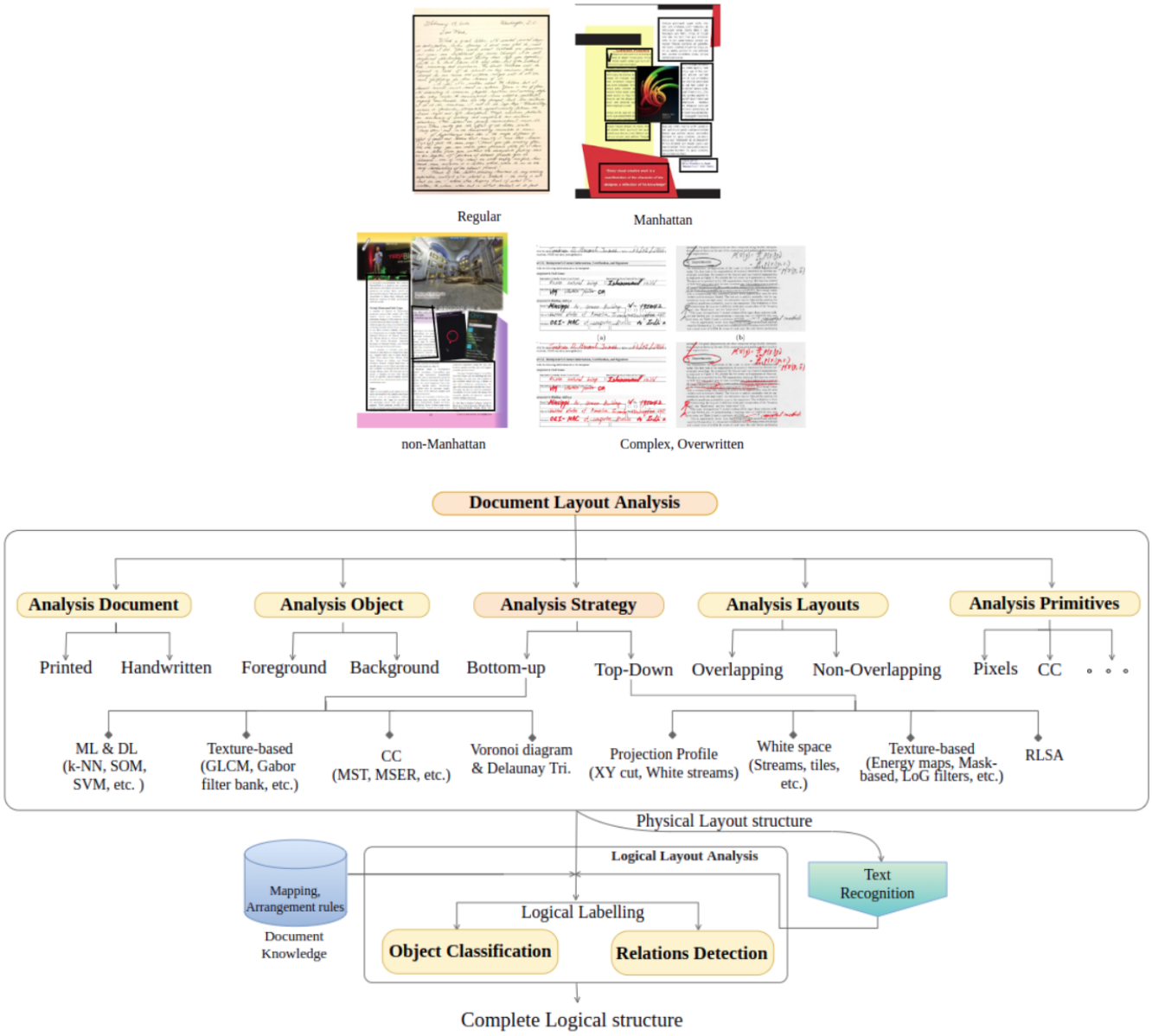


Figure 4 : la catégorisation des différentes approches d'analyse de la mise en page[155]

- DU, également appelée LLA, est une analyse sémantique des documents [8]. Elle extrait les composants logiques d'un document ou des informations relatives à un contexte donné, telles qu'un nom, des dates, etc. Il est essentiel de comprendre le contenu des documents pour permettre aux systèmes de les traiter automatiquement. Les méthodes de pointe ont obtenu des performances impressionnantes sur de nombreux ensembles de données disponibles à cet effet. Cependant, la recherche est encore limitée à des domaines restreints. Il existe d'ample possibilités de développer des systèmes génériques et d'améliorer la précision.

- Segmentation de lignes, de mots et de caractères : La segmentation du texte implique la segmentation en lignes, en mots ou en caractères. Elle peut être réalisée directement sur les images de documents prétraitées ou sur les régions de texte identifiées par l'analyse de la mise en page. Dans les deux cas, les techniques d'analyse de la mise en page sont également

utiles pour la segmentation des lignes, des mots ou des caractères. La segmentation des lignes consiste à localiser les lignes de texte dans une image de document. Elle peut être utilisée ensuite pour la segmentation des mots, l'identification, la reconnaissance, etc. Les lignes d'un document peuvent se chevaucher/se toucher/être brisées, manquer de ligne de base appropriée, être courbées, etc. De plus, le nombre de lignes n'est pas connu au départ.

Method	Description	Segm.
Horizontal Projection	Histogram of foreground pixels for horizontal lines: Pri, hw, [47]	Line Segmentation
Probability Density	Discrete estimate of probability density; computationally expensive: Pri [52]	
Dynamic programming Hough transform Energy maps	global minimization of segmentation cost function: On [56] helps to find the skew angle: Pri [56], Hw [58], [59] Energy map of pixels helps to determine the text and non-text seams: HW [49], Hist [48], [50]	
Watershed	Flooding the morphological surface and segmenting the input into catchment and basin: Pri, handwritten [61]	
Region-growth	Sub-groups of similar neighbouring pixels; bottom-up buildup procedure; computa- tionally expensive: Hist [51]	
Smearing based	smearing of consecutive black pixels along the horizontal direction and for a distance less than the threshold, white spaces are filled with black ones: Pri [53], Hist [54], handwritten [55]	
Grouping Graph-based Fringe maps	aggregating primitives in bottom-up approach: Pri, Hw [57] MST, other graph-based solutions: Pri, handwritten Fringe numbers are assigned based on the distance to the nearest black pixel: Pri/handwritten [60]	
Hybrid	Combination of above approaches: Hist	
Distance metric	Different metrics to compute the distance between CCs (Euclidean distance, Convex, hull, bounding box, average run length): Hw	Word Segmentation
Recognition	Feedback from recognition system; Finds word boundaries based on classification algorithms (Scale space, k-means clustering, Hough transform): Pri, Hw	
Stochastic	Probabilistic algorithms (HMM) used to find non-linear paths between overlapping text lines: HW [62]	Character Segmentation
Touched characters	Segmentation-based (Explicit)	
Level set method	separate the text regions from the background using partial derivatives and an external vector field in an iterative process: Pri, Hw, both	
Touched characters	Recognition-based (implicit)	

Tableau 1 : Divers travaux de littérature sur la segmentation des lignes, des mots et des caractères

La tâche devient plus compliquée avec des mises en page difficiles comme celles des documents complexes, superposés, manuscrits et historiques. Les premières méthodes de segmentation utilisaient des techniques de base telles que l'analyse des CC, MSER, le diagramme de Voronoi [35], [45] ; l'analyse de projection [46]–[47] ; les fonctions d'énergie [48]–[50], etc., pour déterminer les groupes de texte. La figure 4 montre comment la méthode du profil de projection peut être appliquée à divers types d'écriture pour segmenter les lignes.

Le tableau 1 présente divers travaux de littérature sur la segmentation des lignes, des mots et des caractères, ainsi que leurs applications. La segmentation du texte en lignes est un problème subjectif car les documents comportant du texte imprimé sont assez faciles à segmenter avec des techniques de base telles que les profils de projection, comme illustré dans la figure.

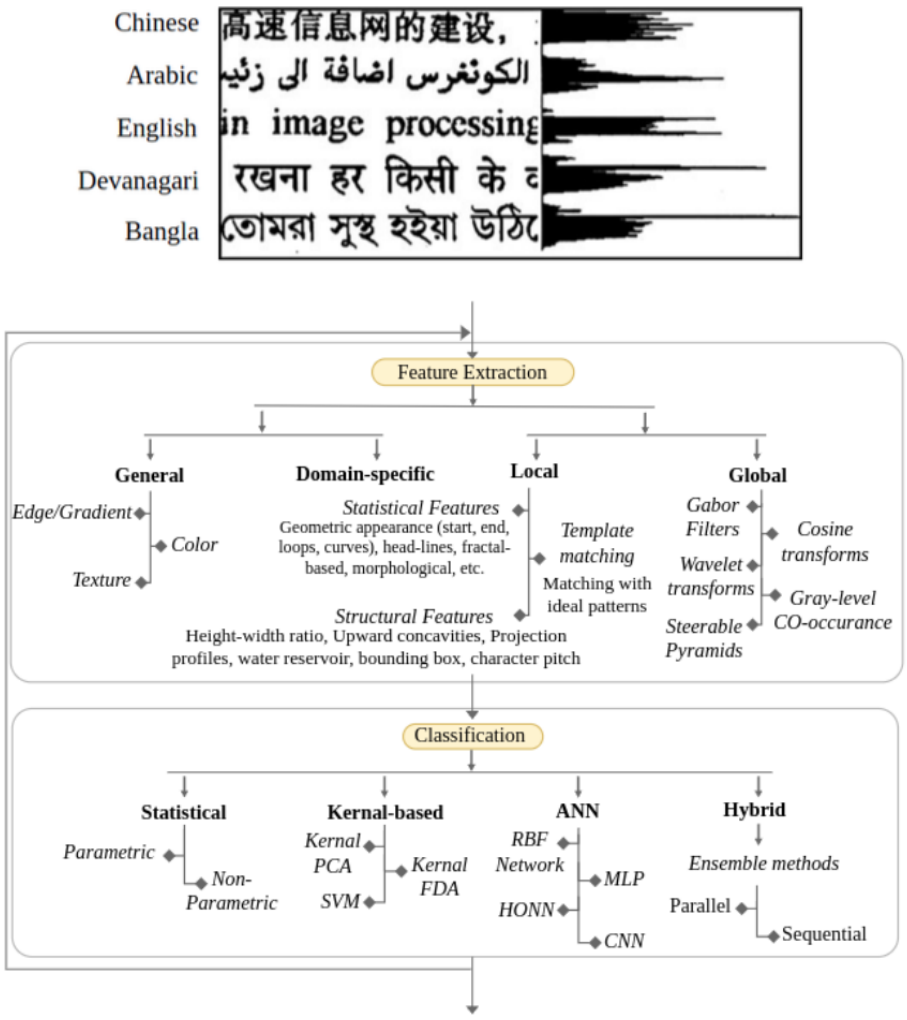


Figure 5 : Reconnaissance de texte (extraction de caractéristiques et classification approches)[155]

La segmentation des documents manuscrits, en particulier des documents historiques, est difficile à gérer. Plusieurs approches ont été proposées pour ces cas complexes, telles que les techniques de croissance de région pour les documents historiques [51]; la densité de probabilité pour les textes imprimés [52] et manuscrits; les méthodes basées sur le flou pour les documents imprimés [53], historiques [54] et manuscrits [55]; la programmation dynamique pour les documents en ligne [56]; les méthodes basées sur le regroupement pour les documents imprimés, manuscrits [57]; la transformée de Hough pour les documents imprimés [56], manuscrits [58], [59]; les graphes et les MST pour les documents imprimés et manuscrits; les cartes d'énergie pour les documents manuscrits [49] et historiques [48], [50]; les cartes de franges pour les documents imprimés et manuscrits [60]; les méthodes de watershed et de flooding pour les documents imprimés et manuscrits [61].

La segmentation des mots est un problème spécifique à chaque écriture. Dans certaines écritures comme le CJK, la séparation des mots est moins distinctive que dans d'autres écritures comme le romain, le brahmi et l'arabe. Par conséquent, les écritures CJK considèrent

principalement la segmentation des caractères. L'approche de la reconnaissance diffère également en fonction de la technique de reconnaissance.

En réponse au problème de la reconnaissance de caractères dans la phase initiale de l'analyse et de la reconnaissance de documents (DAR), la recherche s'est orientée vers la segmentation des caractères après l'extraction des lignes de texte. Depuis lors, le problème a été abordé de manière traditionnelle et moderne. Les lacunes de recherche persistantes sont les suivantes [8] :

- Caractères en contact dans les documents imprimés et manuscrits.
- Caractères brisés dans les documents imprimés.
- Absence de ligne de base dans les documents manuscrits.
- Variété de polices et de styles dans les documents imprimés.
- Différentes orientations du texte, etc.

Ces lacunes ont été abordées par les techniques suivantes. Des algorithmes probabilistes tels que les HMM (modèles de Markov cachés) trouvent des chemins non linéaires entre les lignes de texte qui se chevauchent et les documents manuscrits dans [62].

1-4.3. Reconnaissance de texte

La reconnaissance de texte à partir d'images de documents consiste à observer les motifs caractéristiques des éléments textuels et à utiliser les caractéristiques de ces éléments à des fins de reconnaissance. Les classifieurs utilisent ces caractéristiques sous forme de vecteurs de caractéristiques, de graphes, de chaînes de codes ou de séquences de symboles. Les performances d'un système de reconnaissance dépendent de la qualité des caractéristiques extraites, qui est elle-même basée sur les étapes de prétraitement et de segmentation. Un système de reconnaissance de caractères classe les segments de caractères dans une classe associée, suivi de modèles de décodage des mots. Des connaissances supplémentaires telles que les modèles de langage et un système de reconnaissance peuvent être utiles pour obtenir des résultats de reconnaissance significatifs.

1-4.4. Extraction de caractéristiques

Les caractéristiques des motifs ont été largement étudiées pour résoudre le problème de la reconnaissance automatique de documents (DAR). L'extraction de caractéristiques consiste à transformer les données d'entrée en caractéristiques informatives pouvant être traitées par des algorithmes d'apprentissage automatique ou d'apprentissage profond pour résoudre des problèmes de classification. Elle préserve les informations contenues dans l'ensemble de données d'entrée d'origine et cartographie ces caractéristiques dans une dimension

inférieure. Les caractéristiques peuvent être générales, telles que la couleur, la forme, les contours et la texture, ou spécifiques à un domaine, comme indiqué à la figure 5.

1.4.4.1. Caractéristiques générales

a) Caractéristiques basées sur la couleur : Elles sont principalement utiles pour des applications telles que les publicités, les couvertures de livres, les magazines, les affiches, les images de texte manuscrit, etc. Elles sont sensibles aux défis liés à l'acquisition d'images tels que l'éclairage, la luminosité, etc. [64].

b) Basé sur les contours (gradient) : L'approche basée sur les contours/gradient repose sur le gradient du texte par rapport à son arrière-plan. Les changements de motif brusques entraînent des changements de gradient importants, qui sont utilisés comme caractéristique pour extraire ces régions. Certains travaux dans cette direction sont [65], [66] [67].

c) Méthode basée sur la texture : Ces méthodes sont principalement appliquées aux documents denses, principalement couplées à des méthodes de classification basées sur les régions, par exemple des caractéristiques statistiques (moyenne, médiane, valeurs modales) pour différentes caractéristiques d'une image. GLCM est une matrice représentant le nombre d'occurrences des intensités de pixels dans une image. Elle est ensuite utilisée pour analyser les textures dans une image.

1-4.4.2. Caractéristiques spécifiques du domaine :

Différents systèmes d'écriture possèdent des caractéristiques uniques qui doivent être prises en compte pour développer un système de reconnaissance efficace. Il s'agit de caractéristiques spécifiques au domaine qui nécessitent des techniques d'analyse et d'extraction particulières. Les éléments de l'écriture, tels que les traits (horizontaux, verticaux ou diagonaux), les boucles, les ascendantes et les descendantes, l'inclinaison, etc., sont des caractéristiques descriptives des systèmes d'écriture. De tels cas nécessitent une connaissance a priori des caractéristiques uniques pour les traiter de manière explicite. Une analyse des composants est réalisée pour extraire les composants spécifiques au système d'écriture.

Caractéristiques locales :

Les caractéristiques locales concernent les détails spécifiques aux caractères tels que les caractéristiques statistiques, structurelles, morphologiques et basées sur le contour. Les caractéristiques statistiques examinent les éléments mathématiques tels que la circularité du composant, la surface du contour, l'écart-type, le rapport d'aspect, etc. Les techniques utilisées comprennent les concavités ascendantes [69], les profils de projection [46]–[47], les boîtes englobantes, le pas du caractère, etc. Les caractéristiques de l'histogramme ont été utilisées sous différentes formes telles que les caractéristiques GSC, les caractéristiques

d'éléments directionnels [70], les caractéristiques de percentile [71], etc. Une moyenne pondérée ou un "moment" des pixels du caractère est également utile pour décrire les caractéristiques locales telles que l'invariance d'échelle et de rotation. Une autre technique pour déterminer les caractéristiques locales est la segmentation par zones.

Caractéristiques globales

Les méthodes d'extraction de caractéristiques globales sont la transformée en ondelettes discrètes (DWT), les caractéristiques Gabor et les pyramides de caractéristiques. Elles sont principalement utilisées pour la détection d'objets (OD) contrairement aux caractéristiques locales qui sont utiles pour la reconnaissance. Les histogrammes de projection étaient la première solution d'extraction de caractéristiques et étaient principalement utilisés pour la segmentation du texte imprimé. La figure 4 montre les profils horizontaux de différentes langues.

1-4.5. Classification :

L'extraction de caractéristiques et la classification vont en pair. La classification des caractéristiques extraites en classes prédéfinies était le système de reconnaissance utilisé dans les débuts de la reconnaissance automatique de documents (DAR). Un problème de reconnaissance de caractères pour la langue anglaise comporte dix classes de chiffres, 52 classes de caractères majuscules/minuscules et 62 classes alphanumériques. Les méthodes de classification peuvent être supervisées ou non supervisées, en fonction de l'objectif du problème. La similarité des caractéristiques de certaines classes, par exemple 'o', "0", "O", pose un défi pour les classifieurs. Le choix de la technique d'extraction de caractéristiques est crucial pour les performances du classifieur.

1.4.5.1. Techniques de reconnaissance de texte

a) Reconnaissance hors ligne

La tâche de reconnaissance de texte a été abordée à l'aide de différentes techniques. Celles-ci peuvent être catégorisées en fonction de la manière dont elles modélisent les dépendances (visuelles ou séquentielles). Dans cette perspective, nous avons regroupé les techniques en trois catégories :

- Modélisation des dépendances visuelles (images)
- Modélisation des dépendances séquence à séquence (séquences de cartes de caractéristiques)
- Modélisation de bout en bout

1) Modélisation des dépendances visuelles (Images)

Nous cherchons à extraire le contenu (texte) des images de documents dans la (DAR). Cela comprend l'extraction de caractéristiques spécifiques correspondant aux étiquettes cibles. Diverses techniques ont été proposées pour modéliser la reconnaissance des motifs visuels dans les données. Certaines de ces techniques sont discutées ci-dessous :

a) k-NN : k-NN est une technique non paramétrique qui catégorise les objets en fonction d'une mesure de distance. Un objet non vu est étiqueté avec la catégorie de son voisin le plus proche.

b) SVM : Les SVM ont été largement utilisées pour des problèmes de classification tels que la catégorisation de texte, la reconnaissance de caractères, etc. L'objectif des SVM est de trouver un hyperplan optimal ou un ensemble d'hyperplans (équation 1) qui classifie entre deux ou plusieurs catégories.

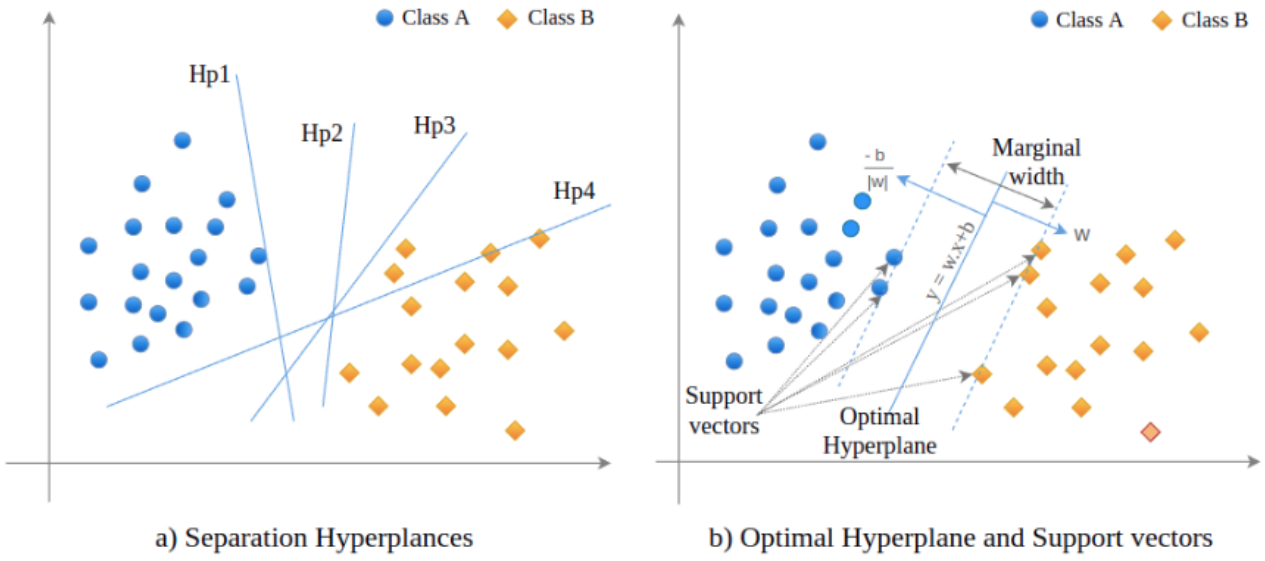


Figure 6 : Support Vector Machines (SVM)

Les applications des SVM avec (DAR) incluent la binarisation [33], la classification de caractères / symboles / chiffres individuels [37], la reconnaissance de caractères [68], [73], la vérification de mots (vérification de signature) [74], la classification texte / non-texte. Les SVM ont été largement utilisés comme la meilleure méthode pour travailler efficacement sur des ensembles de données de petite taille.

c) Architectures de réseaux neuronaux : Le développement d'architectures basées sur des neurones artificiels Perceptron et MLP, a été la première étape vers les architectures de réseaux neuronaux (NN). Avec l'avancée des techniques de stockage et de traitement de documents, il y a eu une augmentation de l'utilisation des NN pour un problème crucial de (DAR) : l'extraction de caractéristiques. D'autres applications des NN dans le DAR sont l'analyse de mise en page, la segmentation et la classification [38], [53], [75]. Comparés à

d'autres techniques comme l'approche des composantes connectées, les NN étaient plus robustes aux motifs bruyants dans les données.

Les CNN ont évolué avec LeNet-5. Il s'est inspiré du Neocognitron, qui pouvait reconnaître des motifs visuels dans les données. Une structure de CNN correspond à la perception visuelle grâce à des couches de convolution qui extraient des caractéristiques à l'aide de structures de convolution et de regroupement (pooling). Les CNN sont supérieurs aux ANN en ce qui concerne les capacités de calcul, telles que :

- Partage des poids et connexions locales.
- Réduction d'échantillonnage.

Les CNN ont été utilisés avec succès pour des tâches de classification telles que la reconnaissance de caractères et de chiffres [63], [72]. Les CNN sont invariants pour la reconnaissance de motifs avec des changements d'échelle et de distorsions. Cependant, certains inconvénients d'un CNN classique sont les suivants : l'exigence d'une taille d'entrée fixe et la dépendance vis-à-vis du champ récepteur, la couche dense permet une prédiction au niveau des pixels mais est coûteuse en termes de calcul, des contraintes sur la taille de sortie en raison des couches denses de taille fixe et l'impossibilité de réutiliser ou de partager les cartes des caractéristiques les rendent inefficaces pour la segmentation sémantique.

d) Méthodes d'ensemble de classifieurs : Les vecteurs de caractéristiques de haute dimension posent le problème du surapprentissage et de la malédiction de la dimensionnalité. Les méthodes d'ensemble [76] telles que AdaBoost et Random Forest peuvent combiner les résultats de différents classifieurs. Ces méthodes de sélection de caractéristiques peuvent résoudre ces problèmes en déterminant les caractéristiques sélectives à l'aide de facteurs de pondération pour les modèles de classifieurs [34]. Les poids des apprenants faibles sont accentués pour améliorer les performances de classification. Les méthodes d'ensemble sont utilisées pour des tâches de DAR telles que la binarisation [32] et la reconnaissance d'écriture manuscrite (HR) [76].

2) Modélisation des dépendances séquence à séquence :
La modélisation des dépendances dans une séquence implique l'extraction des caractéristiques à partir d'observations séquentielles et la mise en relation de ces caractéristiques pour produire une séquence de sortie. Au début de l'ère du DAR, on utilisait principalement les HMM (modèles de Markov cachés) pour modéliser les dépendances séquence à séquence.
a) HMM : Parmi les toutes premières techniques de reconnaissance de caractères se trouve le HMM, qui modélise avec succès les caractéristiques structurelles et statistiques du texte même aujourd'hui. Il existe trois catégories qui constituent un modèle de reconnaissance avec HMM:

- Évaluation (algorithme de Forward).
- Décodage (algorithme de Viterbi).

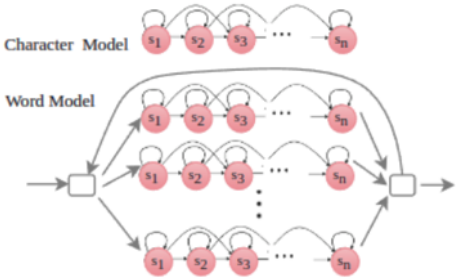
- Entraînement (méthode de Baum-Welch, algorithme EM).

Le succès du HMM peut être attribué à plusieurs facteurs :

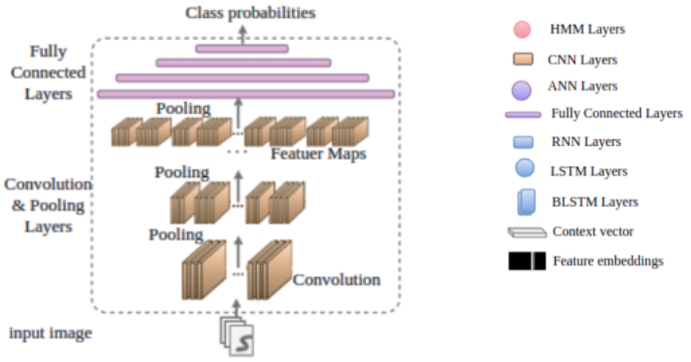
- la puissance de l'optimisation.
- le potentiel de réaliser la segmentation.
- la représentation des sources de connaissances

Il existe quelques limitations du HMM. Ils ne sont pas capables de modéliser les dépendances à long terme. De plus, étant donné que les HMM sont générateurs, ils ne peuvent pas traiter des tâches discriminatives telles que l'étiquetage de séquences [8]. Les réseaux neuronaux récurrents (RNN) sont des options alternatives pour résoudre ces limitations.

b) Réseaux neuronaux récurrents (RNN) : Les RNN sont des réseaux conçus pour modéliser les dépendances séquentielles en prenant les caractéristiques en entrée et en les convertissant en une séquence de sortie. Les caractéristiques distinctives des RNN sont :
1) Les éléments de mémoire : qui permettent aux RNN d'avoir des dépendances à long terme.
2) Partage des poids entre les couches.
Les RNN produisent une sortie pour chaque caractéristique d'entrée ; cela nécessite une pré-segmentation des entrées. Pour éviter cette pré-segmentation, le CTC convertit directement l'entrée en étiquettes cibles. Il existe de nombreuses variations des RNN unidirectionnels à une dimension traditionnelle, comme les RNN bidirectionnels (BRNN) et les RNN multidimensionnels (MDRNN).



a) HMM word and character model



b) CNN for character recognition

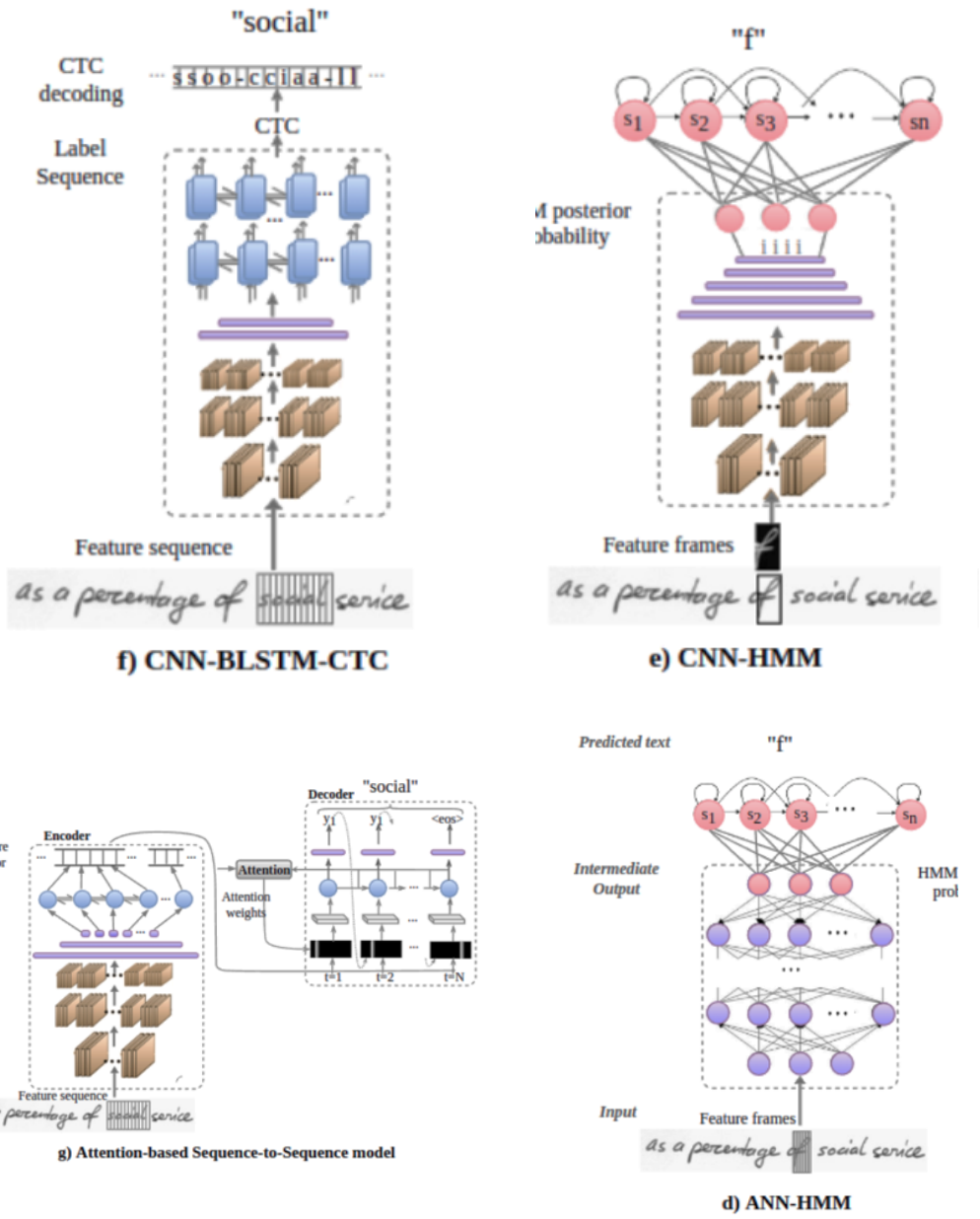


Figure 7 : Reconnaissance de texte avec divers modèles visuels et séquentiels

c) Paradigme séquence à séquence : Les architectures encodeur-décodeur ont été proposées pour modéliser des séquences d'entrée et de sortie de longueurs variables. Elles sont basées sur des architectures récurrentes. L'encodeur extrait les caractéristiques de la séquence d'entrée pour gérer les entrées de longueurs variables. Un vecteur de contexte de taille fixe représente les caractéristiques extraites. En conséquence, le décodeur prédit la sortie à partir du vecteur de contexte. Cependant, le vecteur de contexte de taille fixe ne peut pas représenter de longues séquences d'entrée, ce qui entraîne de mauvaises prédictions.

d) Classification Temporelle Connexionniste (CTC) : Les RNN sont entraînés pour modéliser des données temporelles car ils sont robustes au bruit spatial des séquences. Cependant, les sorties des RNN nécessitent un étiquetage de séquences pour la sortie finale. À cette fin, les

RNN sont combinés à un module CTC pour l'étiquetage de séquences indépendantes. Un modèle CTC utilise la programmation dynamique, qui calcule la somme du logarithme négatif de la probabilité de la séquence de sortie par le RNN.

3) Modélisation de bout en bout :

Les techniques de pointe pour la reconnaissance de bout en bout sont Fast Oriented Text Spotting (FOTS), TextSpotter, Mask TextSpotter, TextDragon, etc. D'autres techniques de systèmes de reconnaissance de bout en bout sont:

a) Modélisation hybride : Les premières méthodes utilisaient principalement des HMM pour la reconnaissance de bout en bout, mais les HMM étaient limités dans la modélisation des dépendances à long terme [77].

b) Méthodologies de détection d'objets : Les méthodologies de détection d'objets ont ouvert la voie à de nouvelles techniques pour la détection de texte. Ces méthodologies visent à trouver les catégories d'objets cibles dans une image d'entrée donnée en localisant les régions d'intérêt (ROI) avec un réseau de proposition de région (RPN) [78]–[79].

c) Reconnaissance sans segmentation : Certains travaux ont récemment exploré la reconnaissance de texte multi-ligne ou de paragraphes en continu, sans aucune étape de segmentation. Auparavant, il y avait des approches sans segmentation pour la reconnaissance de caractères ou de mots. Elles utilisaient une attention récurrente ou des procédures en une étape pour exploiter les attributs de reconnaissance bidimensionnels.

B) Reconnaissance en ligne

La reconnaissance en ligne de texte manuscrit se concentre sur les caractéristiques spatio-temporelles. Le travail [21] catégorise le processus de reconnaissance en quatre classes de procédures utilisées. Il s'agit des méthodes statistiques, structurales, syntaxiques et basées sur les réseaux neuronaux (NN). Les méthodes statistiques sont probabilistes et peuvent être paramétriques ou non paramétriques. Les méthodes paramétriques (par exemple, HMM) utilisent des distributions de probabilité pour modéliser les paramètres des variables (échantillons d'écriture manuscrite). Les paramètres sont sélectionnés lors de la procédure d'entraînement. Les méthodes non paramétriques (par exemple, k-NN) utilisent les données d'entrée (échantillons d'écriture manuscrite) pour estimer les paramètres inconnus. Ces méthodes nécessitent davantage de calculs à mesure que les données d'entraînement augmentent. Les composants structurels tels que les graphes, l'élasticité des chaînes, etc.,

Les méthodes d'extraction de caractéristiques sont basées sur l'extraction des structures et des informations basées sur les traits. Les travaux ont identifié des caractéristiques d'écriture telles que la position verticale des points, la direction d'écriture, la courbure, le lever/baisser

du stylo, le rapport d'aspect, la courbure, la pente, l'ascendant/descendant et la carte de contexte. Des caractéristiques spécifiques à la forme, telles que les ascendantes, les accents, les fermetures, etc., sont identifiées pour améliorer le processus de reconnaissance. Les techniques utilisées sont similaires à la reconnaissance hors ligne, telles que HMM, k-NN, SVM, NN, etc. Cependant, il existe des techniques spécifiquement proposées pour le processus en ligne.

1.5. Jeux de données (Datasets)

Les principaux systèmes d'écriture qui dominent la recherche sur l'analyse et la reconnaissance des documents sont les systèmes d'écriture romain, chinois-japonais et coréen, semblables à l'arabe, asiatiques et indiens [2], [21]. Ces systèmes d'écriture proviennent de types très différents de systèmes d'écriture. Comprendre ces systèmes d'écriture peut être fructueux pour concevoir un système exceptionnel de (RAD). Le développement d'OCR multilingues, multi-systèmes d'écriture et génériques en est encore à ses débuts en raison du besoin de jeux de données standard disponibles. La normalisation des ensembles de données peut aider la communauté de recherche à mieux évaluer les performances des systèmes de (RAD) [81]. Une plateforme standardisée de jeux de données (peut-être une possibilité à l'avenir) serait utile à la communauté de recherche sur la (RAD) pour le développement et l'évaluation.

1.5.1 Datasets de documents historiques

Les documents historiques sont les vestiges historiques d'écrits sur des supports tels que des feuilles de palmier, des pierres, des papyrus, etc. Certains des manuscrits les plus anciens qui subsistent sur ces supports sont : le shaivisme sanskrit sur des feuilles de palmier du IXe siècle, des scripts sur des rouleaux de bambou du siècle, des scripts sur des papyrus datant du quatrième millénaire avant notre ère. En raison du vieillissement de ces documents, leur structure physique est affectée par la dégradation causée pendant le stockage ou la numérisation. L'analyse de ces documents nécessite une certaine connaissance préalable de l'âge et de la période du document, car les caractères, les mots et le vocabulaire varient considérablement au fil des périodes. Leur mise en page et leurs styles de formatage divers posent un défi supplémentaire pour l'analyse de ces documents. Par conséquent, il est improbable de développer un seul système pour analyser et reconnaître les documents historiques. Le travail réalisé dans [82] a classé plus de 60 documents historiques dans l'ensemble de données de classification, de structure et d'analyse.

Ensemble de données	Type de document	Tâche	Période
Esposalles	Livrets de licence de mariage (Hw) (espagnol)	HR	
READ-BAD	Archives européennes (romaines) (Pr)	LD	1470-1930
DIVA-HisDB	3 scripts médiévaux romains (Hw)	DLA	11.14th cen
VML-HD	Livres arabes (Hw)	Rec	1088-1451
Pinkas	Écritures hébraïques	PS	1500-1800

Kuzushiji-MNIST	Caractères Kuzushiji (Pr)	CR	
GRK-Papyri	Écritures papyrus	WI	
Lontar Sunda	Scripts de palmier soudanais (Hw)	Binz, HR, LD	15th cen
Sunda AMADI LontarSet	Scripts de palmier balinais (Hw)	Binz, HR	
Muscima++	Musique RH		
IAM-HistDB (St-Gall)	Écritures romaines (Hw)	DLA, Rec	9th cen
IAM-HistDB (Parzival)	Écritures allemandes (Hw)	DLA, Rec	13th cen
IAM-HistDB (Washington)	Scripts anglais (Hw)	DLA, Rec	18th cen
HJ Dataset	Scans biographiques japonais (Hw, Pr)	IR	
ARDIS	Chiffre suédois (Hw)	DR	18-19th cen
Hugin-Munin	norvégien	HR, WI	1820- 1950
POPP [237]	Recensement de Paris	Rec, WI	1926
HTR Benchmarks (ICFHAR-2014, ICDAR2015, ICFHAR-2016, ICDAR-2017)	Collection Bentham et Ratsprotokolle (Hw)	Rec	
HIP2013 - HNLA2013	Journal historique (Pr)	DLA	17-20th cen
ICDAR2015 -ANDAR-TL- 1K	Documents ancestraux (Hw)	Rec, LD	18-19th cen
ICFHR2016 - CLaMM	Scripts médiévaux romains (Hw)	Clsf	
IC2017 - CLaMM	Scripts médiévaux romains (Hw)	Clsf	
IC2017 - DIVA-His-DB	DLA, LD	Medieval scripts	
ICDAR2017 - REID2017	Livres imprimés en bengali (Pr)	Rec, LD	1785-1909
ICDAR2017 - Historical- WI [12]	Documents manuscrits (allemand, français, arabe) (Hw)	WI	13.20th cen
ICFHR2018 - RASM2018	Manuscrits scientifiques (arabe) (Hw)	PS, LD, Rec	8-9th cen CE
ICFHR2018 - Asian Palm leaf	Manuscrits (balinais, khumer, soudanais, romain) (Hw)	Binz, HR, LD	
ICDAR2019 - DMAS2019	Images numérisées Magazines (Pr)	HR, LD	1800 - 1938
ICDAR2019 - DIBCO2019	images manuscrites et imprimées à la machine, images papyri (Hw, Pr)	Binz	19th cen
ICDAR2019 - cTDaR19 [86]	Documents comptables (Hw)	TD, Rec	
ICDAR2019 - HDRC- Chinese	Registres de famille chinois (Hw)	DLA, LD, Rec	
ICDAR2019 - REID2019	Livres(Bengali) (Pr)	LD, Rec	1713.1914
ICDAR2019 - RASM2019	Manuscrits scientifiques (arabe) (Hw)	LD, Rec	9-19th CE
ICDAR2019 - HDRC-IR	Pages de documents (Hw)	IR, WI	
ICFHAR-2020 - HisFra-gIR20	Livres européens du moyen âge (Hw)	IR	9-15th cenCE
ICDAR-2021 - HDC	Images (Romain) (Hw, Pr)	Clsf	

Tableau 2 : Ensembles de données historiques

Ensemble de données	Type de document	Images : catégories de classe	Tâche
Marmot [87]	Documents de conférence (anglais et chinois)	2 000 : Tableaux, figure	OD, TD
TableBank [20]	document en latex d'arXiv	4,17,000 : Tableaux	OD, TD and Rec

DeepFigures [84]	Documents scientifiques	14,00,000 : Tableaux, figure	OD, TD
UNLV	lettres commerciales, magazines, rapports, journaux, etc.	10 000 : 427 images avec tableaux	OD, TD and Rec
FinTabNet	Rapports financiers de l'org.	90 000 : 112 887 Tableaux	TD
NTable-ori [19]	Images originales de l'appareil photo (textuelles, électroniques, sauvages)	2 100+ : Tableaux	TD
NTable-cam [19]	Images de caméra augmentées (textuelles, électroniques, sauvages)	17 000+ : Tableaux	TD
NTable-gen [19]	Jeu de données synthétique	17 000+ : Tableaux	TD
DocBank [88]	document en latex d'arXiv	5,00,000 : Tableaux, figures, équations, listes, paragraphes, etc.	DLA
IIIT-AR-13k	document type entreprise	13 000 : Tableau, figure, image naturelle, logo, signature	Page OD, TD
UW-III	Images de documents	1 600 : Tableaux	TD
ICDAR-13 [89]	(EU & US) Gov. Images de fichier PDF	238 : 150 Tableaux	OD, TD and Rec
ICDAR-17 POD	Pages scientifiques (anglais)	2 417 : 2939 figures, 1069 tableaux, 4707 formules	OD, TD
ICDAR-19 cTDaR	Formulaires, documents financiers et articles scientifiques	3 600 : Formules, tableaux, figures, graphiques	OD, TD and Rec
MediTables [24]	rapports de pathologie, de diagnostic et d'hospitalisation	200 : 330 tableaux	TD
PubLayNet [90]	plus d'un million d'articles PDF PubMed Central	3,60,000 : Tableau, Figure, Titre, Texte, Liste	Page OD, TD
PubTabNet	articles scientifiques dans PM-COA	5,68,000 : Tableaux	TD and Rec
DocLayNet [85]	pages PDF	80 863 : Légende, note de bas de page, formule, élément de liste, pied de page, en-tête de page, image, en-tête de section, tableau, texte, titre	DLA
Laser-Printed Characters [91]		8 000 : 11 552 caractères	Forgery Detection
NCERT5K-IITRPR [92]	Livres scolaires NCERT	Plus de 5 000 : tableaux, graphiques, figures, images, équations, schémas de circuit, âges, équations, logos	Non-text component analysis

Tableau 3 : Ensembles de données imprimés pour diverses tâches de (DAR)

Cette partie donne une brève introduction à tous les ensembles de données. Elle fournit des détails sur le type de documents, l'utilisation des ensembles de données avec leur origine et leurs mesures de performance ainsi que d'autres métriques associées. Le tableau 3 donnent un aperçu de certains détails des principaux ensembles de données de documents historiques. Les ensembles de données contiennent une quantité substantielle de scripts romains, suivis de scripts asiatiques couvrant des documents du monde entier. La littérature récente utilise des méthodologies d'apprentissage profond comprenant des CNN. Ces architectures abordent diverses tâches telles que la binarisation, la reconnaissance manuscrite (caractère, mot, ligne), la détection de lignes, DLA, PS, WI, IR, etc. La plupart des ensembles de données sont utilisés pour des tâches de reconnaissance (mot, ligne, caractère). Récemment, les documents historiques sont également utilisés pour des tâches telles que la datation des documents et la localisation géographique.

1.5.2. Ensembles de données de documents imprimés

Un document imprimé peut être n'importe quoi, comme des livres, des documents scientifiques, des journaux, des lettres commerciales, des magazines, des documents officiels tels que des lettres commerciales ou des PDF gouvernementaux, ou des documents médicaux tels que des rapports de diagnostic, des analyses pathologiques, etc. Ces documents sont hétérogènes en ce qui concerne leur représentation, leur langue, leurs formats (numérisés, programmatiques ou les deux) et leur mise en page. Ces documents contiennent du contenu textuel et graphique comme des figures, des tableaux, des logos, des équations, des signatures, etc. Chaque composant d'un document, qu'il s'agisse de texte ou de graphiques, transmet des informations.

Auparavant, les ensembles de données de documents avaient généralement un nombre limité de catégories d'étiquettes. Par exemple, de nombreuses recherches étaient axées sur la détection de tableaux avec de nombreux ensembles de données disponibles, tels que l'ensemble de données de la compétition de tableaux ICDAR-2013 [89], cTDaR ICDAR 2019 [86], UNLV, reconnaissance de tableaux Marmot [87], TableBank [20], FinTanNet, NTable [19]. De plus, les ensembles de données étaient limités en taille.

Dataset	Language	samples	lines	Statistics words	characters	digits	Mode	Task
CEDAR [93]				10570	27835	21179	Off	Cursive DR, Rec
NIST [94]		3600			810000		Off	Hw DR, CR
MNIST [95]						70000	Off	DR
Firemaker		1000					Off	WI, Rec [32]
IAM [96]	Roman (English)	1539	13353	115320			Off	WS [27], ,WI [32], Hw LS , off-HR
								[76]
IAM-OnDB		1700	13049	86272			Off	On-HR [190], On-WI, GC
IAM on-Do		1000					Off	Content type detection, WS, TNC [20].
OnHW-chars	Roman (English)				31,275 (U/L)		On	CR
IBM-UB-1	Roman (English)	6500 On,			31,275 (U/L)		On	WI, WS, indexing, DAR
		6000 Off					Off	
GNHK		687	9,363	39,026			Off	
IRONOFF	Roman (French)			50,000	32,000		On	Rec, On-WI
RIMES [97], [98]	Roman (French)	12,723					Off	DLA, mail Clsf, Rec [74], and WI
RODRIGO	Roman (Spanish)	1853					Off	Rec

OHASD				3825	19,467		On	LD, Rec
AltecOnDB				152,680	644,530		On	Rec
Al-Isra			500	37,000		10,000	Off	HR, WI
IFN/ENIT		2265		26,449			Off	preprocessing, WR WI [30],
Checks DB		7000		29,498		15,000	Off	check HTAR
AHDB		105					Off	HR, WI
ARABASE		400					On /Off	On/Off HR, SV
CENPARMI-A [99]	Arabic			11,375	21,426	13,439	Off	CR, DR, WS [25]
LMCA				500	100,000	30,000	On	WR, DR
ADAB				20,000 +			Off	Seg, Rec, On-WI
KHATT		1000					Off	pre-processing, Seg, WI
QUWI		4068					Off	WI, writer demographic classification
AHTID-MW			3710				Off	Seg, WI
IAUT/PHCN	Arabic (Farsi)	1140		34,200			Off	pre-processing, WR
IFN Fars	Arabic (Farsi)			7271			Off	DR, WR
FHT	Arabic (Farsi)	1000	8050	106,600			Off	Seg, Rec, BLD, content discrimination, WI, DLA
CENPARMI-F	Arabic (Farsi)	432,357					Off	DR, HR,
HaFT	Arabic (Farsi)	1800					Off	Seg, Rec, WI
CENPARMI-U	Arabic (Urdu)				18,000		Off	HR, WS
UHSD]	Arabic (Urdu)	400					Off	Seg, Rec, WI

Tableau 4 : Ensemble de données manuscrits classés sur différents scripts

Récemment, la recherche s'est orientée vers des domaines uniques et spécifiques de l'analyse de documents plutôt que les domaines traditionnels. De plus en plus d'ensembles de données sont maintenant développés avec diverses étiquettes et des sources de collecte de données complètes. Des efforts ont également été déployés pour normaliser le format des données de référence. Les formats de données courants facilitent les procédures d'entraînement des méthodes de détection d'objets sur les ensembles de données. Certains ensembles de données de grande taille sont comme TableBank [20], DocBank [88], DeepFigures [84], PubTabNet, PubLayNet [90], etc. Parmi les plus récents, on trouve NCERT5k IITRPR [92] pour l'analyse des composants texte/non-texte, Laser Printed Characters Dataset [91] pour les applications de médecine légale documentaire, DocLayNet [85] pour une DLA à usage général. Le tableau 4 répertorie les ensembles de données disponibles soutenant la classification, l'analyse et la compréhension des documents, et résume les principales contributions aux ensembles de données imprimées.

1.5.3. Ensembles de données manuscrites

L'écriture manuscrite a été un moyen de communication et de stockage de l'information pendant longtemps. Avec l'avancement des technologies numériques pour la lecture et

l'écriture, les documents imprimés sont devenus plus pratiques et faciles à stocker. Cependant, cela ne peut jamais remplacer la commodité d'un stylo et d'un papier. Le besoin de systèmes d'analyse et de reconnaissance de texte manuscrit (HTAR) est inévitable. Les ensembles de données manuscrites standard ont été développés dès les années 1990 [81]. Les premières phases de (DAR) étaient généralement axées sur la reconnaissance de caractères, de mots et de chiffres. L'évolution des techniques de HR a ouvert la voie à des méthodologies de reconnaissance non contraintes. Un autre aspect concernant (DAR) est le développement de techniques visant à reproduire l'efficacité de reconnaissance humaine. Cela n'est possible que lorsque les ensembles de données reproduisent également des scénarios réels. L'environnement de collecte de données non contraint et flexible est essentiel pour créer de tels ensembles de données.

Les principaux ensembles de données manuscrites comprennent IAM [96], NIST [94], MNIST [95], CEDAR [93], RIMES [97], [98], UNIPEN, CENPARMI-Arabic [99], PE92 [102], etc. Les ensembles de données développés sont principalement en langues comme l'anglais avec IAM, CEDAR, NIST, MNIST, IAM-OnDB, etc., l'arabe avec AHDB, ARABASE, CENPARMI-A, LMCA, KHATT, CENPARMI-F, etc., le chinois avec HCL2000, CASIA, SCUT-COUCH, etc., et les langues indiennes comme le bengali avec BN-HTRd, Numerals DB, Devanagari DB, Multiscript Indian DB (bengali, devanagari, tamoul, télougou) [80], Multiscript DB 11 scripts (romain, devanagari, ourdou, kannada, oriya, gujarati, bengali, gurumukhi, tamoul, télougou, malayalam) [100].

Les tâches traditionnelles de DAR (reconnaissance et analyse de documents manuscrits), soutenues par la plupart des ensembles de données, sont le prétraitement, la segmentation et la reconnaissance. D'autres tâches telles que l'analyse de la mise en page, la recherche de mots et l'analyse de documents judiciaires (WI et vérification) ont très peu d'ensembles de données les concernant. Le tableau 4 présente différents ensembles de données manuscrites en ligne/hors ligne avec les langues respectives et les tâches prises en charge.

En examinant l'ensemble des travaux sur (DAR), l'accent est principalement mis sur les scripts anglais et arabes. Une des raisons en est la disponibilité d'échantillons de données gratuits et étiquetés dans ce domaine. Les travaux récents sur (DAR) ont vu un éventail plus large de scripts inclus dans une seule recherche. La création de solutions pour l'analyse et la reconnaissance multiscripts implique de découvrir les points communs des différents systèmes d'écriture et des individus.

1.6. Métriques d’évaluation

Les métriques d'évaluation analysent les points forts et les points faibles des techniques et des processus. Elles aident à contrôler la qualité et à vérifier les performances des systèmes. Le schéma d'évaluation de base compare la sortie prédite avec les étiquettes de vérification correspondantes des données. L'évolution des systèmes et des algorithmes (DAR) a rendu difficile leur évaluation sur un terrain commun. Les schémas d'évaluation ont

été classés en approches théoriques et expérimentales [8]. Les approches théoriques sont utilisées pour les algorithmes de bas niveau, principalement pour les tâches de reconnaissance de motifs visuels. Les méthodes expérimentales sont ensuite classées en deux catégories : avec données de vérification et sans données de vérification. Lorsque les comparaisons sont effectuées sans données de vérification, les mesures sont utilisées pour évaluer des caractéristiques spécifiques des algorithmes afin de tester leur qualité.

1.6.1. Métriques pour les méthodes de prétraitement

Les métriques de qualité d'image sont utilisées pour évaluer les méthodes de prétraitement telles que la binarisation. Les méthodes de binarisation sont principalement évaluées selon les mesures fournies par les compétitions de binarisation. Même après la standardisation des méthodes d'évaluation, il est toujours nécessaire d'avoir une méthode standard qui puisse être appliquée à la plupart des documents. Les mesures standard pour évaluer les méthodes de binarisation, telles que celles fournies dans différentes compétitions de binarisation [103]–[104], sont la mesure F, le PSNR, le NRM, le DRD et le MPM. Les métriques de précision et de rappel sont les plus significatives pour toutes les tâches de (DAR). Le rappel fait référence à l'exhaustivité et la précision fait référence à la pureté de la tâche de recherche d'informations.

1.6.2. Métriques pour les méthodes de segmentation

Les méthodes de segmentation comprennent des méthodes d'analyse de la mise en page et des méthodes de segmentation de page / ligne / mot / caractère. Les métriques discutées dans cette section sont tirées de la compétition PS [105], qui est principalement basée sur un MatchScore(i, j). Le nombre de correspondances détectées par rapport à la vérification est défini comme le nombre de résultats correspondants de la région GT j et de la région de résultat i. Le taux de détection et la précision de reconnaissance sont définis en fonction du MatchScore. Une autre métrique, EDM, mesure la performance de détection.

1.6.3. Métriques pour les méthodes de reconnaissance

Les métriques de reconnaissance mesurent les taux d'erreur, que ce soit pour les caractères ou les mots. Les métriques sont basées sur la distance de Levenshtein qui utilise la programmation dynamique. Le taux d'erreur de caractère (CER) calcule la distance de Levenshtein entre le caractère prédit et le caractère de vérification. Le taux d'erreur de mot (WER) calcule la distance de Levenshtein entre les mots prédits et les mots de vérification. Le travail réalisé dans [8] définit la précision (Rappel) et la précision pour les systèmes de reconnaissance comme le rapport des mots correctement reconnus sur le nombre total de caractères dans la vérité terrain ou la sortie OCR, respectivement.

1.7. Discussion

Les réseaux de neurones (NN) ont été la méthode de choix pour les problèmes d'extraction de caractéristiques et de classification. Ils sont utilisés pour une grande variété de problèmes de (DAR), tels que la reconnaissance en ligne ou hors ligne, la reconnaissance imprimée ou manuscrite de mots, de caractères et de chiffres. Cependant, pour une séquence complète de texte telle que des lignes, des paragraphes, etc., les NN sont combinés par des modèles capables de traiter les probabilités en séquences de caractères. Ces modèles incluent le CTC [101], le HMM, etc. Un NN prend une image en entrée et produit des vecteurs de caractéristiques correspondants. Une tâche de (DAR) peut nécessiter le traitement d'informations contextuelles ou temporelles en fonction de l'application. Les tâches simples comme la reconnaissance de caractères ou de chiffres peuvent être accomplies en ne tenant compte que de l'information contextuelle. Dans ces cas, les réseaux de neurones artificiels (ANN) ou les réseaux de neurones convolutifs (CNN) sont des choix efficaces. Cependant, le traitement de séquences de texte nécessite des informations temporelles et un stockage supplémentaire. Les réseaux de neurones récurrents (RNN), LSTM et Bi-LSTM, etc., sont utilisés. Les Bi-LSTM peuvent traiter les séquences passées et futures car ils traitent à la fois dans les directions de droite à gauche et de gauche à droite.

1.8. Défis spécifiques aux scripts

1) Prétraitement spécifique au script : La plupart des méthodes de prétraitement mentionnées ci-dessus sont indépendantes des scripts. Cependant, certains documents nécessitent un traitement spécifique aux caractéristiques du script avec des méthodes tenant compte des particularités propres au script.

2) Segmentation spécifique au script : Les représentations des mots et des caractères sont diverses dans tous les principaux systèmes d'écriture du monde. Dans les scripts romains, les mots ont généralement plus d'espaces entre les mots qu'à l'intérieur des mots. En revanche, les systèmes logographiques (CJK) ont un système syllabique où les mots n'indiquent pas d'espaces entre les mots. Avec ces scripts, il n'y a que la segmentation des lignes et des caractères. La segmentation des lignes dans les scripts CJK est plus facile et similaire à celle des scripts romains. La segmentation des caractères est plus facile dans les documents imprimés avec uniquement des scripts CJK, car les syllabes peuvent être facilement segmentées avec des méthodes comme les profils de projection. Cependant, la tâche est compliquée pour les documents manuscrits.

Les systèmes Abjad comprennent les scripts arabes, hébreux, syriaques et thaana. Ce sont les plus complexes, avec seulement des alphabets et un ensemble de caractères très large (environ 4 000 utilisés aujourd'hui). Certaines complications incluent l'espacement non homogène, l'écriture cursive, le chevauchement des mots, le style d'écriture de droite à gauche et les signes diacritiques autour des caractères.

Le système d'écriture des abugidas est basé sur le script Brahmi, y compris les scripts indic (Devanagari, Gurumukhi, Gujarati, Bengali, Manipuri, Oriya, Tamoul, Telugu, Kannada, Malayalam, etc.), les scripts indonésiens (balinais, bugis, etc.). Tous ces scripts présentent d'importantes disparités, qui sont dues à l'influence régionale au fil des années. Le système d'écriture des abugidas comprend un petit ensemble de consonnes et de voyelles formant un caractère. La combinaison de ces consonnes et voyelles varie d'un script à l'autre et en grand nombre, ce qui pose des difficultés en termes de segmentation et de reconnaissance.

3) Reconnaissance spécifique aux scripts : Les caractéristiques de haut niveau des scripts nécessitent des techniques spécifiques pour déterminer les principales caractéristiques de leurs composants. Cependant, les caractéristiques de bas niveau sont principalement indépendantes et peuvent être extraites à l'aide de méthodes communes d'extraction de caractéristiques (pour tous les scripts). Les scripts logographiques ont des radicaux et des composants comme éléments de base de leurs caractères. Une méthode d'extraction de caractéristiques de base consiste à squelettiser les radicaux et à calculer les composantes principales par PCA.

Category	Method	Dataset	Metric	Result
Projection-based methods	Projection Profile [46]	Multi-script	Skew correction Error	0.12
	Recursive XY cut [101]	private	Subjective	NA
Smearing-based methods	RLSA [40], [415]	private	FM	84.80%
CC-based methods	Docstrum [41]	private	Subjective	NA
	Delaunay triangulation	Multi-script	FM	100%
Background analysis methods	Voronoi diagram [45]	Multi-script	success rate	99.05%

Tableau 5 : Performance des méthodes DLA sur diverses métriques d'évaluation

Category	Method	Dataset	FM	PSNR
Global threshold	Otsu [31]	DIBCO-11 Santgall	82.22 80.71	16.94 17.09
Local threshold	Niblack [29]	DIBCO-11	68.52	12.76
	Sauvola [30]	DIBCO-11 Santgall	82.54 88.68	15.78 19.86

Edge Detection	Su [55]	DIBCO-11	87.8	17.56
Image transforms	Sehad [58]	DIBCO-11	88.90	17.51
Mixture models	FAIR [60]	DIBCO-11	92.36	19.32
Conditional Random Fields	Howe [62]	DIBCO-11	88.74	17.84
Game theory	GiB [65]	DIBCO-11	89.85	18.86
Deep Learning	Pastor [70]	DIBCO-13 Santgall	87.74 97.02	18.91 27.22

Tableau 6 : Performances de différentes méthodes de binarisation par des métriques d'évaluation (FM-PSNR

1.9. Conclusion

En conclusion, ce chapitre met en évidence les avancées et les défis de la reconnaissance automatique de documents (DAR) dans différents contextes, tels que l'identification de caractères imprimés ou manuscrits, en ligne ou hors ligne, et avec des contraintes spécifiques ou non. Les tâches fondamentales de la DAR, telles que le prétraitement, la segmentation et la reconnaissance, sont essentielles pour obtenir de bonnes performances des systèmes. Cependant, on observe une transition vers des systèmes de reconnaissance de bout en bout qui évitent les étapes intermédiaires et améliorent la transcription des textes contenus dans les documents.

Il convient également de noter que la DAR présente des défis particuliers lors de la reconnaissance de documents historiques. Les documents historiques peuvent être sujets à la détérioration et à l'altération au fil du temps, ce qui rend la reconnaissance encore plus complexe. De plus, ces documents peuvent contenir des styles d'écriture et des conventions spécifiques à une époque donnée, ce qui nécessite des approches de reconnaissance adaptées.

Le chapitre suivant examine les avancées récentes dans le domaine de la datation des documents historiques, en mettant en évidence les différentes approches et techniques utilisées dans le domaine de l'analyse et de la reconnaissance de documents. Nous explorerons les méthodes traditionnelles telles que l'analyse paléographique, qui consiste à étudier les caractéristiques de l'écriture et de la calligraphie pour évaluer l'âge des documents. Nous aborderons également les techniques scientifiques modernes, telles que les avancées de l'intelligence artificielle et de l'apprentissage automatique dans ce domaine.

Chapitre 02 : L’etat de l’art

2.1. Introduction

De nos jours, les méthodes d'apprentissage profond sont utilisées dans un large éventail de domaines de recherche. L'analyse et la reconnaissance de documents historiques, comme nous le présentons dans ce chapitre, ne font pas exception. Ce chapitre analyse les articles publiés ces dernières années sur ce sujet sous différents angles : nous commençons par donner une définition pragmatique des documents historiques du point de vue de la recherche dans ce domaine, puis nous examinons les différentes sous-tâches abordées dans ce domaine de recherche. Guidés par ces tâches, nous passons en revue les différentes relations Input/output attendues des approches d'apprentissage profond utilisées, et décrivons en conséquence les modèles les plus utilisés. Nous discutons également des datasets de recherche publiés dans le domaine et de leurs applications. Cette analyse montre que les dernières recherches représentent un bond en avant, car il ne s'agit plus simplement de l'utilisation des algorithmes récemment proposés pour résoudre des problèmes antérieurs, mais de nouvelles tâches et de nouvelles applications de méthodes de pointe sont maintenant prises en compte. Au lieu de simplement présenter une image conclusive de la recherche actuelle sur le sujet, nous suggérons enfin quelques tendances futures potentielles qui peuvent stimuler des directions de recherche novatrices.

2.2. Documents Historiques

L'histoire est définie à travers des documents et se distingue de la préhistoire grâce à l'existence de documents écrits qui décrivent sous une forme permanente différents aspects de la vie passée. Alors que cette considération facilite l'établissement d'un point de départ pour la définition des documents historiques ("un document ancien"), il est plus difficile, voire impossible, de fixer une limite de temps supérieure ("à partir de quel âge un document est-il considéré comme historique ?"). De ce point de vue, un document est historique s'il peut être étudié par des historiens pour analyser une période donnée. Pour mieux caractériser document historique, nous adopterons une approche plus pragmatique et considérerons comme documents historiques ceux qui sont réalisés sur des supports ou selon des techniques qui ne sont plus utilisés ou qui sont plus difficiles à analyser que les documents contemporains. Pour mieux définir la limite conceptuelle des documents historiques considérés dans notre analyse, nous fournissons, dans cette section, un résumé des éléments qui sont souvent considérés comme historiques par la communauté de recherche (DAR).

2.2.1. Documents anciens

La plupart des documents anciens sont écrits sur des supports différents du papier : feuilles de palmier, pierres, rouleaux de bambou, papyrus. Les auteurs de manuscrits sur feuilles de palmier utilisaient un support fait à partir de feuilles de palmier séchées, utilisées comme matériau d'écriture dans le sous-continent indien et en Asie du Sud-Est, datant du Ve siècle av. J.-C. et peut-être même plus tôt. L'un des plus anciens manuscrits sur feuilles de palmier conservés, contenant un traité complet en sanskrit sur le shivaïsme, remonte au IXe siècle. Il a été découvert au Népal et est actuellement conservé à la bibliothèque de l'Université de Cambridge [7-9]. Les premiers exemples de tablettes en bois ou de rouleaux de bambou remontent au Ve siècle av. J.-C., pendant la période des États en guerre. Au IVe siècle de notre ère, le bambou avait été largement abandonné comme support d'écriture [147]. Le papyrus a été fabriqué pour la première fois en Égypte dès le quatrième millénaire av. J.-C. et utilisé pour des documents dans le Journal de Merer, daté de 2560-2550 av. J.-C. [148].

2.2.2. Manuscrits et incunables

Les manuscrits sont un type particulier de document, provenant généralement d'Europe ou du "monde occidental", de la période classique aux premiers siècles de l'ère chrétienne. Certains de ces manuscrits étaient également écrits sans espaces entre les mots (scriptio continua) ou avec une écriture très difficile à comprendre, souvent ornée de décorations ou de coupures qui les rendent particulièrement difficiles à lire pour les non-initiés. Les exemplaires existants de ces premiers manuscrits écrits en grec ou en latin et datant généralement du IVe au VIIIe siècle de notre ère sont classés en fonction de leur utilisation de lettres majuscules ou minuscules. Plusieurs chercheurs se sont intéressés à l'analyse et à la reconnaissance de ces documents.

2.2.3. Autres documents

En plus des documents dont le contenu principal est textuel (décrits jusqu'à présent), d'autres éléments importants contiennent des informations graphiques qui nécessitent des techniques différentes et dont le but principal n'est pas de "lire" et de comprendre le texte.
Certains de ces documents sont conservés dans des archives qui contiennent souvent des documents manuscrits. En général, les documents d'archives sont plus diversifiés que les documents des bibliothèques et, dans la plupart des cas, les éléments d'archives individuels sont composés de quelques pages. Les documents d'archives décrivent souvent des informations financières dont la compréhension nécessite des techniques différentes de celles utilisées, par exemple, pour lire un roman ou un manuscrit. Des exemples de documents

d'archives écrits entre 1470 et 1930 provenant de neuf archives européennes différentes sont présentés dans la référence [120].

Les documents historiques issus des recensements, des actes de naissance, des registres familiaux [24] et des registres historiques [121] contiennent des informations structurées dont l'extraction permet aux chercheurs de reconstituer des généalogies et de réaliser des études démographiques [122]. Plusieurs chercheurs se sont intéressés à la reconnaissance de ces types de documents.

2.3. Problèmes traités

Différents problèmes liés à l'analyse des documents historiques sont abordés et résolus dans les recherches résumées dans cet article. Dans cette section, nous mettons en évidence les principales tâches considérées afin de comprendre quelles sont les principales relations d'entrée-sortie qui doivent être résolues par les techniques de Deep Learning proposées [153]. Les principales tâches sont organisées selon une chaîne de traitement traditionnel en (DAR) [106] où les images des documents sont d'abord collectées (Section 2.3.1), puis trois étapes de traitement principales sont effectuées : le prétraitement pour améliorer la qualité des images (Section 2.3.2), l'identification/segmentation des régions avec un contenu homogène grâce à des techniques d'analyse de la mise en page (Section 2.3.3), et enfin la reconnaissance du contenu textuel (Section 2.3.4).

2.3.1. Constitution des collections dans les bibliothèques numériques

Les documents historiques, une fois numérisés, sont généralement stockés dans des bibliothèques numériques et des archives. Afin d'indexer correctement les documents de la collection, les conservateurs ont besoin de connaître certaines informations sur les éléments archivés, telles que l'auteur d'un manuscrit et la date (ou la période) de sa production. Pour estimer le coût de la transcription, il est également utile de connaître le nombre d'enregistrements dans les documents d'archives. Les techniques de Deep Learning traitant de ces sujets sont discutées ci-dessous.

2.3.1.1. Identification de l'auteur

L'identification de l'auteur consiste à associer les lignes et les pages des images de documents historiques manuscrits à leur auteur et écrivain. Ce problème a été abordé par Cilia et al. [124] dans un travail présentant un système de bout en bout pour identifier les écrivains dans les manuscrits médiévaux. Le système proposé se compose d'un modèle en trois étapes pour la détection et la classification des lignes dans le manuscrit, ainsi que l'identification de l'auteur de la page. Les étapes de détection et de classification des lignes reposent sur la combinaison de l'affinage de MobileNetV2 [125] et d'un CNN personnalisé. La troisième étape est constituée d'un combinateur de décision pondéré par vote majoritaire des lignes, qui vise à relier les pages et les auteurs. Les réseaux Siamese peuvent également être

utilisés pour l'identification et la reconnaissance des auteurs, comme discuté dans la référence [126].

2.3.1.2. Datation des manuscrits

Les manuscrits sont répandus dans les bibliothèques et les archives. Contrairement aux œuvres plus modernes, comme les livres imprimés, la date de production des manuscrits est souvent inconnue. Cependant, une datation automatique de ces œuvres est utile pour un archivage précis dans les bibliothèques numériques.

La datation des manuscrits historiques repose sur le problème qui se pose de manière répétée lorsqu'il s'agit de tout manuscrit d'origine et de période de production inconnues. L'idée principale est d'associer les compétences en écriture de l'auteur, son style et sa technique à une certaine période historique. Cela se fait soit par l'analyse de diverses caractéristiques au sein des pages du document, soit par l'écriture des auteurs. Hamid et al. [151] présentent une approche basée sur l'apprentissage profond pour caractériser efficacement l'année de production d'échantillons de documents à partir de l'ensemble de données MPS (Medieval Paleographical Scale) - un ensemble de documents de la fin de la période médiévale (1300-1550). La méthode proposée repose sur un modèle basé sur les convolutions qui extrait des caractéristiques à partir de patchs d'images de documents manuscrits. Selon les auteurs, cette approche surpasse les techniques traditionnelles d'extraction de caractéristiques basées sur le traitement d'images.

Plus précisément, les résultats discutés par les auteurs sont calculés sur l'ensemble de données MPS et les performances du système sont comparées à d'autres méthodes, principalement non basées sur l'apprentissage profond, qui ont été précédemment utilisées pour résoudre la tâche. Le système proposé dans la référence [151] améliore considérablement les résultats par rapport à ces travaux précédents, d'au moins un facteur 2 (jusqu'à un facteur 10), réduisant considérablement les valeurs d'erreur moyenne absolue (MAE) (de valeurs allant de 35,4% à 7,8% jusqu'à 3%), augmentant ainsi la précision pour la tâche proposée lorsqu'elle est évaluée sur l'ensemble de données pris en compte.
Les méthodes d'apprentissage profond pour la datation des manuscrits sont également discutées dans la référence [36], qui met en évidence l'importance et l'efficacité de l'utilisation de réseaux pré-entraînés en tant qu'extracteurs de caractéristiques à partir de documents historiques, puis leur application en affinant pour différentes tâches.

2.3.1.3. Estimation des coûts de transcription

Compte tenu de la tendance croissante à transformer les bibliothèques numériques en lieux où les utilisateurs peuvent trouver rapidement des informations et des livres de manière efficace et efficace, il est essentiel d'estimer le temps nécessaire pour transcrire le contenu textuel. Dans la référence [127], les auteurs proposent un modèle basé sur la segmentation pour estimer le temps nécessaire pour transcrire une grande collection de documents

manuscrits historiques lorsque la transcription est assistée par un système de repérage de mots clés selon l'approche de la requête par chaîne. Le modèle a été validé en comparant ses estimations avec le temps réel nécessaire pour la transcription manuelle de pages à partir de l'ensemble de données Bentham [41].

2.3.2. Prétraitement

Après la numérisation, la première étape dans un pipeline (DAR) traditionnel est le prétraitement des images d'entrée. Cette tâche vise à améliorer la qualité du document, que ce soit pour une meilleure inspection humaine du travail ou pour améliorer la reconnaissance automatique des étapes de traitement ultérieures. En raison de la nature des documents historiques, cette étape est particulièrement pertinente dans ce domaine, comme le démontrent les différents articles relatifs au prétraitement des documents historiques. Deux opérations principales sont réalisées pour le prétraitement : l'amélioration de l'image du document (y compris la réduction du bruit et la restauration de l'image) et la binarisation du document.

2.3.2.1. Amélioration de l'image

Le débruitage des documents historiques est l'une des étapes les plus difficiles dans le domaine du traitement d'images et de la vision par ordinateur. Dans le cas de la restauration de documents historiques, l'objectif principal est d'améliorer la qualité de l'image pour faciliter les étapes ultérieures d'analyse du document. Malgré les avancées récentes dans l'exactitude de la reconnaissance des caractères isolés à l'aide de réseaux neuronaux profonds, les systèmes OCR échouent presque à reconnaître les motifs de caractères lorsqu'ils sont gravement dégradés, en particulier ceux des documents historiques.

2.3.2.2. Binarisation de l'image

Une autre tâche importante dans le prétraitement des images de document est la binarisation. Cela consiste à transformer une image couleur ou en niveaux de gris en une image en noir et blanc. Cette tâche est souvent réalisée pour minimiser l'impact des dégradations physiques du document telles que le fond non uniforme, les tâches, l'encre décolorée, l'encre qui traverse la page et l'éclairage inégal sur les images de document. Le processus de binarisation sépare le contenu du document de ces facteurs de bruit en classifiant chaque pixel comme avant-plan ou arrière-plan.

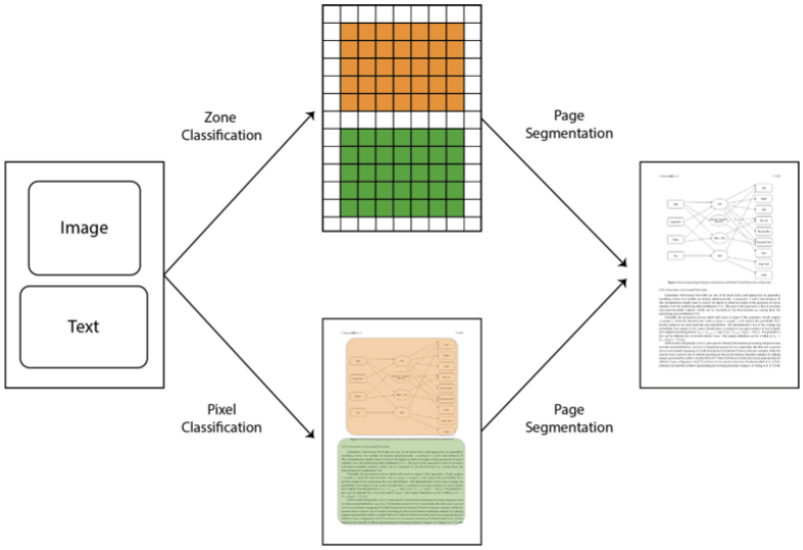


Figure 8. Classification de zone (ou patch) par rapport à la classification de pixel.

2.3.3. Analyse de la mise en page

L'analyse de la mise en page est le processus permettant d'identifier et de reconnaître l'organisation physique et logique des documents numérisés. La chaîne de traitement typique dans la plupart des applications (DAR) commence par des opérations de prétraitement visant à nettoyer les images bruitées, puis se poursuit par l'analyse de la mise en page et la reconnaissance des caractères/symboles. Les techniques d'apprentissage automatique ont été largement utilisées pour l'analyse de la mise en page [130] et ceci reste un banc d'essai important pour les applications des techniques d'apprentissage profond en (DAR). Les sous-tâches décrites dans cette section sont organisées de manière descendante. Nous décrivons d'abord les techniques permettant d'identifier les tableaux dans les pages entières, puis les approches visant à localiser les lignes de texte entières ou leurs lignes de base.

2.3.3.1. Détection/Reconnaissance des tableaux

Ces dernières années, l'accès en ligne aux documents d'archives a considérablement augmenté. La disponibilité de documents tels que des registres ou des livres de recensement contenant à la fois du texte et des tableaux, pouvant être entièrement dessinés et écrits à la main ou imprimés, a stimulé la recherche sur l'extraction d'informations à partir de documents tabulaires.

2.3.3.2. Segmentation des lignes de texte

La segmentation des lignes de texte est une tâche centrale pour l'analyse des manuscrits et autres documents historiques. Alors que la segmentation des lignes de texte dans les documents imprimés est relativement facile, la segmentation du texte manuscrit historique est

particulièrement difficile en raison de l'orientation irrégulière des lignes de texte et de la présence de divers artefacts (comme des lettres capitales enluminées et des notes marginales, c'est-à-dire des notes ajoutées à la page). Un autre aspect des manuscrits est qu'ils ont été rédigés en essayant d'utiliser le moins de pages possible ; par conséquent, les lignes de texte sont plus rapprochées les unes des autres par rapport aux documents modernes. La segmentation des lignes de texte vise à identifier la zone correspondant à chaque ligne de texte (Figure 2) et a été abordée avec plusieurs approches basées sur l'apprentissage profond. (a) Échantillon (b) lignes de texte (c) lignes de base (Figure 2). Différentes approches pour l'identification des lignes de texte dans le manuscrit (a) : segmentation des lignes de texte (b) et détection des lignes de base (c) [131]. Chen et al. [132] proposent un simple CNN n'ayant qu'une seule couche de convolution.

(a) Échantillon (b) lignes de texte (c) lignes de base

Figure 9 : différentes approches pour l'identification des lignes de texte dans le manuscrit

Alaasam et al. [149] effectuent une classification des patchs en utilisant un réseau Siamese qui compare une paire de patchs et produit en sortie leur similarité. De cette manière, il est possible de regrouper les patchs de chaque page en trois classes : texte principal, texte d'accompagnement (annotations) et arrière-plan. L'approche est testée sur 8 pages (contenant environ 4000 patchs) issues de différents manuscrits historiques arabes. Les résultats ont été évalués par rapport à la référence [133] et à la référence [134]. En particulier, Bukhari et al. [133] utilisent une approche basée sur un MLP, tandis que la référence [134] présente un FCN pour l'analyse de la mise en page des lignes courbes. En travaillant au niveau des patchs plutôt qu'au niveau des pages, la méthode proposée surpasse ces travaux à la fois dans la segmentation du texte principal et du texte d'accompagnement, améliorant également la précision pour l'analyse de mises en page particulières, ce qui porte le score F1 pour l'analyse du texte principal et du texte d'accompagnement respectivement de 95% à 98,5% et de 94% à 96,8% [149].

Les différentes lignes de texte sont localisées dans les documents historiques indiens dans la référence [150] en utilisant un modèle profond basé sur un Mask R-CNN avec une architecture

ResNet-50. La différence entre la segmentation d'instances (qui sépare les objets individuels dans la page, par exemple chaque ligne de texte) et la segmentation sémantique (qui vise à identifier les pixels appartenant à un type d'objet donné, par exemple les lignes de texte) est prise en compte et discutée dans l'article. Par rapport à d'autres approches, plusieurs classes sont considérées parmi les types d'objets : segment de ligne de caractères, limite de page, trou, ligne de limite, composant de caractère et dégradation physique.

2.3.3.3. Détection de la ligne de base

La détection de la ligne de base est une approche légèrement différente de la segmentation des lignes de texte précédemment discutée. Dans le cas de cette dernière, la tâche consiste à identifier les lignes de texte individuelles (Figure 2b), tandis que dans le cas de la détection de la ligne de base, l'objectif est d'identifier la ligne située en dessous de chaque ligne de texte (Figure 2c). La principale raison de considérer ces deux tâches différentes réside dans le coût d'annotation pour l'étiquetage des données utilisées pour l'entraînement. Dessiner une ligne sous une ligne de texte est clairement plus facile (et plus rapide) que d'encadrer tous les pixels appartenant à la ligne.

Un réseau neuronal convolutif basé sur U-Net est utilisé dans la référence [135] pour la détection de la ligne de base et la reconnaissance des lignes de texte dans les images de documents historiques. Plus précisément, le modèle BL-net proposé est un U-net résiduel, testé sur différentes configurations et nombres de couches cachées. Il a été évalué en prédisant les lignes de base sur des images à différentes échelles et avec différents ensembles de données, et donne des performances égales ou supérieures aux travaux comparés, atteignant des résultats d'une précision proche de 99% (une amélioration de 1% par rapport au meilleur résultat obtenu dans HistDoc [135]).

2.3.4. Reconnaissance des caractères et des symboles

La reconnaissance des caractères consiste à comprendre et à reconnaître les caractères linguistiques de différents idiomes, qu'ils soient manuscrits ou imprimés. Dans les premières années d'application des réseaux de neurones artificiels pour effectuer la reconnaissance des caractères, la plupart des méthodes portaient sur les langues occidentales (souvent l'anglais) [38]. Cela était clairement dû aux grands ensembles de données disponibles pour la recherche. Au cours des dernières années, plusieurs programmes de numérisation dans de nombreux pays du monde ont rendu possible l'étude et l'application de différentes techniques de reconnaissance de caractères sur différentes langues et écritures.

2.3.4.1. Repérage de mots-clés

Une tâche étroitement liée à la reconnaissance de texte est le repérage de mots-clés, où les occurrences d'un mot de requête sont localisées dans des documents numériques. L'utilisation du repérage de mots-clés est appropriée dans le domaine des documents

historiques, car les systèmes de reconnaissance peuvent échouer en raison de la mauvaise qualité de l'écriture manuscrite et du manque de disponibilité de dictionnaires adaptés pour vérifier les résultats de la reconnaissance de l'écriture manuscrite. Une application des architectures CNN pour le repérage de mots-clés dans des documents manuscrits est proposée dans la référence [136] et est testée sur différents ensembles de données historiques.

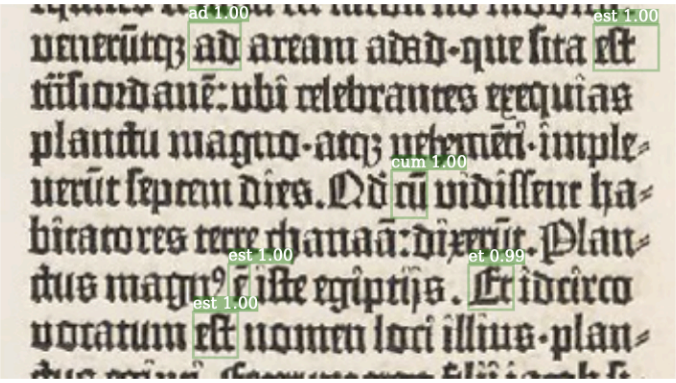


Figure 10. Repérage de mots-clés dans un fragment de document historique.

3.3.4.2. Détection de texte

Les cartes historiques contiennent souvent à la fois des informations textuelles et graphiques qui représentent différentes caractéristiques géographiques ou politiques et des échelles spatiales. De tels documents historiques complexes présentent plusieurs défis uniques : le texte peut apparaître dans presque n'importe quelle orientation, de nombreuses tailles et avec des espacements importants entre les éléments graphiques ou même d'autres textes à proximité. Dans ce contexte, il est important d'effectuer la détection et la reconnaissance de texte sans supposer que le texte est organisé de manière régulière en lignes de texte. De nombreux travaux antérieurs présentent ou utilisent des algorithmes artisanaux pour faire face à de telles complexités, les systèmes d'apprentissage profond peuvent être utilisés efficacement pour extraire du texte à partir d'images complexes de cartes historiques[150].

Localiser et reconnaître les informations textuelles dans les images naturelles est l'objectif principal de la détection de texte de scène, un sujet de recherche qui a suscité une attention importante ces dernières années dans le domaine de l'IA pour la reconnaissance automatique de documents (par exemple, références [65, 66]). Utiliser ces techniques pour reconnaître le texte dans les documents historiques graphiques est une direction de recherche pertinente à prendre en compte.

2.3.4.3. Reconnaissance de caractères

La recherche sur la reconnaissance de texte manuscrit dans les langues occidentales n'est pas encore complètement résolue. L'un des sujets de recherche les plus importants dans ce domaine est le développement de techniques pour former un système de reconnaissance de

texte manuscrit (HTR) avec peu de données étiquetées. Chammas et al. [138] démontrent comment former un système HTR avec peu de données étiquetées. Plus précisément, ils entraînent un réseau de neurones convolutifs récurrents profonds sur seulement 10 % de données de lignes de texte étiquetées manuellement à partir d'un ensemble de données, puis ils effectuent une procédure d'entraînement incrémental qui couvre les 90 % restants. Les performances sont améliorées par un processus d'augmentation de données à plusieurs échelles.

2.4. Architectures neuronales et leurs applications

Depuis la proposition des premières architectures d'apprentissage profond basées sur des réseaux neuronaux convolutifs [143], plusieurs techniques ont été proposées au fil des ans. Dans cette section, nous souhaitons concentrer sur les approches qui se sont avérées adaptées à la reconnaissance de documents historiques. À cette fin, nous discutons d'abord dans la section 4.1 des différentes relations entrée-sortie utiles dans la reconnaissance de documents historiques, selon l'analyse effectuée dans la section 3. Ensuite, nous décrivons les différentes architectures qui peuvent être utilisées pour former les relations entrée-sortie susmentionnées dans la section 4.2. Dans la section 4.3, nous résumons les types d'architectures qui ont été utilisés pour obtenir les relations entrée-sortie souhaitées.

2.4.1. Relations Entrée-Sortie

Divers types d'informations peuvent être calculés à partir de pages complètes. Le prétraitement (binarisation, restauration d'image) est une tâche typique où l'entrée et la sortie ont la même taille et correspondent à la page entière. Dans ce cas, l'opération souhaitée correspond à un filtrage d'image et peut être utilisée pour la binarisation des images de documents, la génération d'images et l'étiquetage des pixels, entre autres tâches. La plupart des techniques d'analyse de mise en page identifient des régions avec un contenu uniforme, comme des paragraphes. Dans certains contextes, le contenu textuel dans le document est plus clairsemé, comme dans les cartes historiques ou la recherche de mots clés. Dans ces cas, les informations de sortie sont basées sur des boîtes englobantes appropriées autour des éléments d'intérêt. Lorsque l'objectif global est d'extraire les informations textuelles des documents historiques, nous avons alors besoin de la transcription du texte, qui est abordée dans plusieurs travaux. Des informations globales (valeurs numériques) correspondant à chaque page peuvent également être calculées, par exemple pour dater le document [124] ou pour identifier le nombre d'enregistrements dans les registres.

Certaines méthodes prennent en entrée un patch d'image qui est déplacé sur l'image d'entrée et qui peut produire en sortie un patch modifié pour la réduction du bruit de l'image [119], la restauration de documents ou la restauration de lettres. Les informations de sortie calculées à partir de chaque patch peuvent également être utilisées pour la datation des documents [151] ou l'identification de l'écrivain [124]. Il est important de remarquer ici que plusieurs approches basées sur les réseaux de neurones convolutionnels suivent implicitement un

paradigme similaire, car dans les premières couches, les filtres de convolution sont déplacés sur l'image pour extraire les mêmes caractéristiques dans différentes positions (Section 4.2.1). La reconnaissance de texte peut également être abordée en recherchant le contenu dans une région avec différentes approches qui sont souvent similaires à celles proposées pour la reconnaissance de texte dans des pages entières. La dernière combinaison de relations entrée-sortie que nous considérons est l'information textuelle qui est transformée en un document historique simulé pour effectuer une augmentation de données pour l'entraînement des algorithmes d'apprentissage profond.

2.4.2. Architectures d'apprentissage profond

Au cours des dernières années, plusieurs architectures d'apprentissage profond ont été proposées pour répondre à différents domaines d'application. Étant donné que la plupart des problèmes liés à l'analyse de documents historiques concernent la reconnaissance de documents numérisés, il est raisonnable qu'une grande partie des travaux reposent sur différentes variantes de réseaux neuronaux convolutionnels (Sections 4.2.1.4.2.6). Pour aborder la correspondance image-image (ou plus généralement le filtrage), des approches basées sur les auto-encodeurs ont également été utilisées (Section 4.2.5). Lorsqu'il s'agit du contenu textuel des documents, qui peut être considéré comme un flux séquentiel d'informations, les réseaux neuronaux récurrents sont des architectures naturelles à considérer, en particulier les approches récentes basées sur les LSTM (Section 4.2.7). La dernière famille d'architectures utilisées pour les documents historiques est celle des réseaux génératifs antagonistes, qui ont été utilisés pour la génération de jeux de données d'entraînement (Section 4.2.8).

2.4.2.1. Réseaux Neuronaux Convolutionnels

Les CNN ont été largement utilisés pour le traitement et la reconnaissance d'images de documents historiques, et les couches de convolution de base constituent l'épine dorsale de plusieurs architectures plus complexes qui sont résumées ci-dessous. Les CNN sont très efficaces pour réduire le nombre de paramètres sans perdre en puissance de représentation des modèles appris, c'est pourquoi la complexité des images peut être réduite dans des cartes de caractéristiques informatives, mais de taille plus petite, extraites par les couches de convolution.

L'apprentissage d'une extraction de caractéristiques utiles dans les premières couches des CNN nécessite un ensemble d'entraînement considérablement large. Étant donné que les ensembles de données disponibles dans le domaine des documents historiques (Section 5.1) ne sont souvent pas assez grands, des stratégies de transfert d'apprentissage sont nécessaires pour résoudre avec succès les différents problèmes décrits dans ce chapitre. L'idée clé est de transférer et de généraliser l'apprentissage d'une tâche à une autre similaire. Le concept de transfert d'apprentissage est particulièrement utile lorsque la quantité de données disponibles concernant un problème spécifique n'est pas suffisante pour entraîner

un CNN profond à partir de zéro. Dans de tels scénarios, un réseau pré-entraîné comme ImageNet peut être utilisé soit comme extracteur de caractéristiques à partir des images étudiées, soit en adaptant ses dernières couches pour affiner le réseau lui-même. Hamid et al. [151] comparent différentes pré-entraînements de CNN pour la datation des manuscrits, Weinman et al. [123] utilisent ResNet50 comme colonne vertébrale convolutive pour la détection et la reconnaissance de texte dans les cartes historiques. L'ensemble des tâches abordées comprend la reconnaissance de caractères, la classification de style, la datation des manuscrits, la segmentation sémantique et la recherche basée sur le contenu.

2.4.2.2. Réseaux neuronaux Siamese

Les réseaux neuronaux Siamese sont des réseaux neuronaux contenant deux composants sous-réseau ou plus, identiques [140]. Ces composants partagent les mêmes poids tout en travaillant de concert sur deux entrées différentes pour calculer des sorties comparables, et ils sont généralement utilisés pour comparer des instances similaires dans différents ensembles de types. L'idée principale derrière les réseaux Siamese est qu'ils peuvent apprendre des descripteurs de données utiles qui peuvent ensuite être utilisés pour comparer les entrées des sous-réseaux respectifs. Les réseaux Siamese sont utilisés dans la référence [149] pour extraire des lignes de texte en comparant des fragments d'images de documents et en estimant leur similarité.

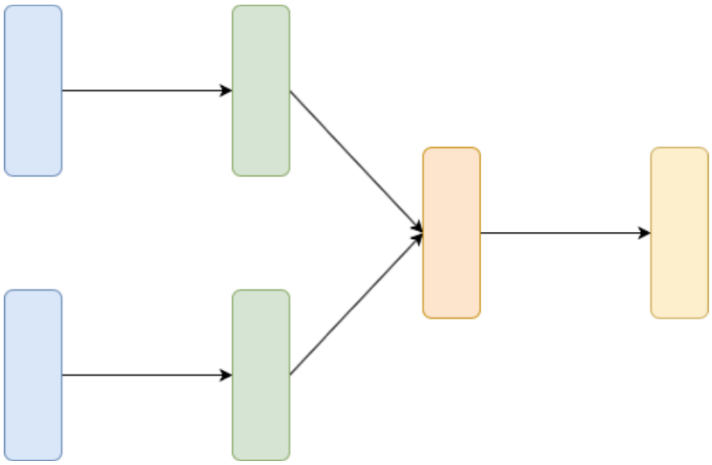


Figure 11 : Architecture de réseau de neurones Siamese simple.

2.4.2.3. Réseaux entièrement convolutionnels

Le concept de base derrière les réseaux entièrement convolutionnels (FCN) est de ne contenir que des couches convolutionnelles [141]. Les FCN ne possèdent pas de couches entièrement connectées à la fin de l'architecture, qui sont généralement utilisées pour la classification. Au lieu de cela, ces réseaux utilisent des couches convolutionnelles pour effectuer une classification pixel par pixel, calculant une sortie ayant la même largeur et la même hauteur que l'image d'entrée, mais avec un nombre de canaux égal au nombre de classes. Des modèles respectant les caractéristiques mentionnées ci-dessus ont également été utilisés

dans le domaine du traitement et de la compréhension des documents historiques, comme résumé ci-dessous. Prusty et al. [150] utilisent par exemple des FCN pour la segmentation d'instances de lignes de texte et d'autres zones dans les documents historiques.

2.4.2.4. U-Nets

U-Net est un type de FCN très adapté à la segmentation d'images, initialement proposé pour la segmentation d'images biomédicales. Tout comme les autres FCN, son objectif est de prédire la classe de chaque pixel de l'image. L'architecture du réseau est schématisée à la Figure 5. Il se compose d'un chemin contractant (côté gauche) et d'un chemin expansif (côté droit). Le chemin contractant est constitué d'un réseau convolutionnel standard. Il comprend l'application répétée de deux convolutions non rembourrées de 3 × 3, chacune activée par ReLU, et d'une opération de réduction d'échantillonnage de 2 × 2 par max pooling - chaque étape de réduction d'échantillonnage double le nombre de canaux de caractéristiques. Le réseau compte 23 couches de convolution [129].

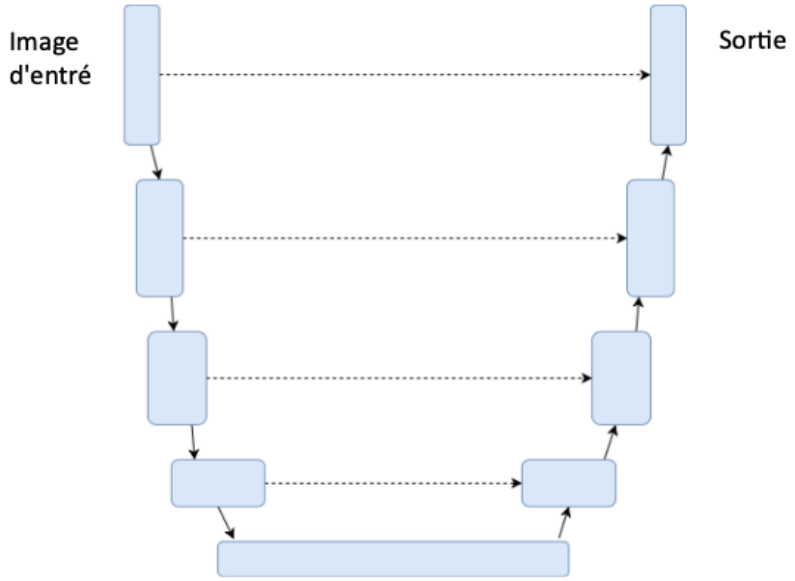


Figure 12 : Schéma général de l'architecture Unet

Cette architecture a été conçue et créée pour résoudre la tâche de segmentation d'images dans le domaine médical. Par la suite, étant donné l'efficacité de cette méthode pour la tâche de segmentation, l'architecture a été utilisée dans de nombreux autres domaines. L'un de ces domaines est sans aucun doute l'analyse des images de documents historiques et en particulier la segmentation des documents.

2.4.2.5. Réseaux Encodeur-Décodeur

La détection d'objets dans les images a suscité un intérêt croissant ces dernières années grâce à l'introduction de puissantes techniques basées sur l'apprentissage profond. Les méthodes traditionnelles reposaient sur des algorithmes spécialement conçus pour la

proposition de régions et l'extraction de caractéristiques à partir de l'image, suivis de classifieurs entraînables. En revanche, les architectures modernes abordent la sélection de régions, l'extraction de caractéristiques et la classification au sein d'une seule architecture entraînable.

Les applications des modèles de détection d'objets pour les documents historiques vont de la recherche de mots-clés dans les premières œuvres imprimées [21] en utilisant Faster R-CNN, à la localisation de différentes lignes de texte dans les documents historiques indiens en considérant une architecture Mask R-CNN [150].

2.4.2.6. Modèles profonds pour la détection d'objets

Les auto-encodeurs et les réseaux encodeur-décodeur sont des réseaux neuronaux artificiels qui apprennent à compresser et à encoder efficacement des données, ainsi qu'à les reconstruire à partir de la représentation encodée réduite pour obtenir une représentation aussi proche que possible de l'entrée d'origine. Ces réseaux neuronaux réduisent les dimensions des données en apprenant à ignorer le bruit dans les données. La forme la plus simple d'un auto-encodeur est un réseau neuronal feed-forward non récurrent similaire aux couches individuelles dans les perceptrons multicouches, avec une couche d'entrée, une couche de sortie et une ou plusieurs couches cachées les reliant, où la couche de sortie a le même nombre de neurones que la couche d'entrée, dans le but de reconstruire ses entrées en minimisant la différence entre l'entrée et la sortie, comme illustré dans la Figure 6.

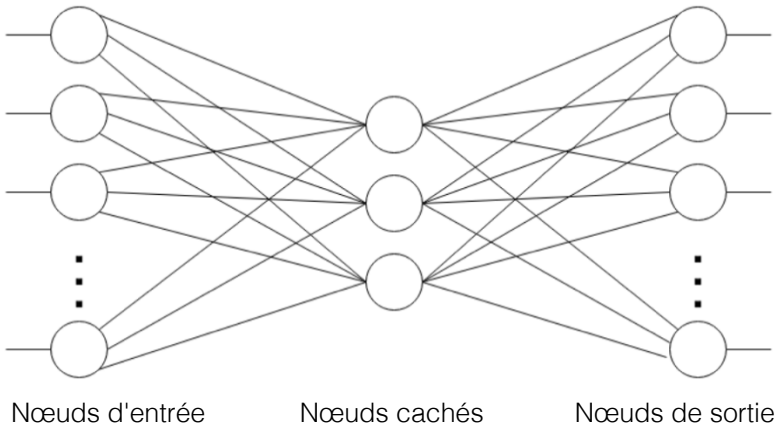


Figure 13 : Architecture d'auto-encodeur.

Comme mentionné précédemment, les auto-encodeurs sont des modèles d'apprentissage non supervisé qui exploitent les réseaux neuronaux pour la tâche de l'apprentissage de représentations. Plus précisément, ils sont conçus comme une architecture de réseau neuronal imposant un rétrécissement dans celui-ci, ce qui force une représentation compressée des connaissances de l'entrée d'origine. Les auto-encodeurs convolutionnels, tout comme FCN et U-Net, ont les mêmes dimensions en entrée et en sortie. Les auto-

encodeurs convolutionnels sont utilisés pour apprendre des caractéristiques à partir d'images de documents historiques et les classifier, comme décrit dans la référence [142].

2.4.2.7. Réseaux neuronaux récurrents

Les réseaux neuronaux récurrents (RNN) sont un type de modèle de réseau neuronal contenant une couche cachée auto-connectée. Ce type de modèle est capable de traiter une séquence de longueur arbitraire en appliquant de manière récursive une fonction de transition à son vecteur d'état caché interne de la séquence d'entrée[150]. L'activation de l'état caché pour chaque pas de temps est calculée en fonction de l'entrée actuelle et de l'état caché précédent, l'un des avantages de cette connexion récurrente est qu'une "mémoire" des entrées précédentes reste dans l'état interne du réseau, lui permettant de tirer parti du contexte passé.

2.4.2.8. Réseaux antagonistes génératifs

Les réseaux antagonistes génératifs (GAN) sont l'une des approches les plus récemment utilisées en modélisation générative, où deux modèles sont entraînés simultanément : un générateur G et un discriminateur D. Le discriminateur vise généralement à classer ses entrées comme étant soit une sortie du générateur, soit des échantillons réels provenant de la distribution de données sous-jacente p(x). L'objectif du générateur est donc de produire des sorties trompeuses, qui peuvent être classées par le discriminateur comme provenant de la distribution de données sous-jacente.

Un tel modèle est fréquemment utilisé dans le traitement et la génération de certains documents historiques spécifiques, comme décrit plus en détail ci-dessous. Cai et al. [137] présentent un modèle pour deux tâches principales : la première est réalisée sur le transfert de style pour des échantillons de caractères chinois traditionnels à polices multiples, tandis que la deuxième tâche est réalisée sur l'apprentissage par transfert pour des échantillons de caractères chinois historiques en ajoutant des échantillons générés par le modèle TH-GAN des auteurs.

2.4.3. Combinaisons Entrée-Sortie et Architectures de Réseaux Neuronaux Associées

Nous avons analysé dans la section 4.1 les combinaisons les plus courantes d'informations d'entrée et de sortie, et dans la section 4.2 les architectures profondes les plus utilisées lors du traitement de documents historiques. La Figure 7 récapitule les types d'architectures qui ont été utilisées pour mettre en œuvre la relation d'entrée-sortie attendue dans la littérature analysée dans cet article. Ce schéma synthétique peut être utilisé comme référence pour indiquer quels types d'architectures peuvent être considérés pour résoudre des tâches spécifiques.

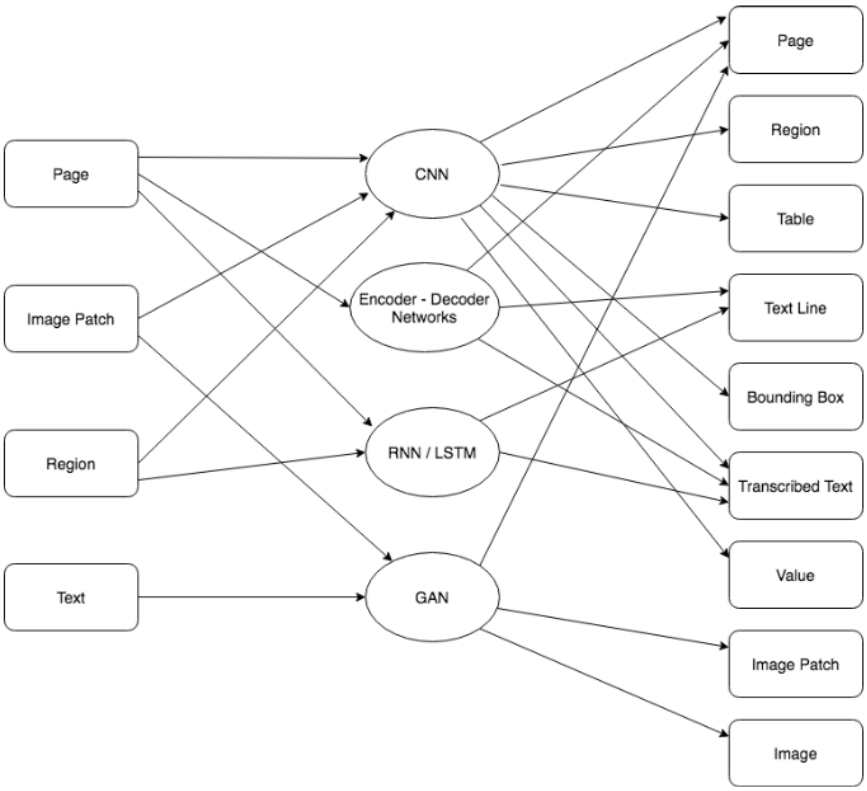


Figure 14 : Combinaisons d'entrée-sortie les plus fréquentes et architectures de réseaux de neurones associées.

2.5. Environnement expérimental

La conception efficace et efficiente de solutions basées sur les techniques d'apprentissage profond nécessite un solide environnement expérimental qui comprend deux composantes principales : des ensembles de données étiquetées et des plateformes d'entraînement. Un autre élément qui alimente cette recherche est l'organisation de compétitions spécifiques, qui sont également essentielles pour constituer des ensembles de données de valeur. Lorsque la collecte de grands ensembles de données n'est pas possible, des stratégies d'apprentissage par transfert et la génération de données d'entraînement synthétiques peuvent également être envisagées. La recherche sur l'analyse et la reconnaissance des documents historiques ne fait pas exception, comme résumé dans cette section.

2.5.1. Ensembles de données

Au cours des dernières années, plusieurs ensembles de données ont été proposés pour soutenir la recherche sur la reconnaissance des documents historiques, comme discuté ci-dessous et résumé dans le Tableau 1. Il est important de souligner que tous ces ensembles de données ne sont pas qu'une simple collection d'images et de transcriptions, mais sont étudiés et annotés par des chercheurs en sciences humaines, experts dans leur domaine respectif.

Dataset	Task	Type of Doc.	Type of GT	Number of Items
George Washington	Keyword Spotting	Handwritings	Word-level annotations	20 pages 656 text lines
Esposalles	Handwriting Recognition	Archival Documents	Word-level annotations	173 pages, 1747 registers, and 5447 lines
READ-BAD	Baseline Detection	Archival Documents	Page-level annotations	2036 pages
DIVA-HisDB	Layout Analysis	Medieval Manuscripts	Page Layout	150 pages, 3 manuscripts
St.Gall	Handwriting Rec./Layout Analysis	Handwritten Latin Manuscripts	Word and line level annotations	60 pages, 1410 text-lines, 11,597 words
Parzival	Handwriting Rec./Layout Analysis	Handwritten German Manuscripts	Word and line level annotations	47 pages, 4477 text-lines, 23,478 words
VML-HD	Character Recognition	Arabic Scripts	Character and sub-words level bounding boxes	680 pages, 1731 sub-words, 1731 sub-words
Pinkas Dataset	Page Segmentation	Hebrew Manuscripts	Word, line and page-level annotations	30 pages, 13,744 words
Kuzushiji - MNIST	Character	Kuzushiji Characters	Character-level annotations	49 character classes
GRK-Papyri	Writer Identification	Handwritten Papyri	Writer id	50 pages, 10 writers
Lontar Sunda	Binarization, Recognition	Sudanese palm leaf Manuscripts	Binarized images, word and character level annotations	66 pages, 27 collections
SleukRith	Binarization,	Khmer palm	Character, word and	657 pages
Sunda AMADI _LontarSet	Recognition Binarization	leaf Manuscripts Balinese palm leaf manuscripts	line-level annotations Binarized images, word and character level annotations	100 pages
Muscima++	Music Recognition	Handwritten music Pages	Notation graph	140 pages, 91,254 symbols

Tableau 7 : Caractéristiques des principaux ensembles de données utilisés dans l'analyse d'images de documents historiques.

Les documents très anciens rédigés sur différents supports attirent également l'attention de plusieurs chercheurs ces dernières années. GRK-Papyri [148] est un ensemble de données d'écriture grecque sur papyrus utilisé pour tester les algorithmes d'identification des auteurs. Les éléments de l'ensemble de données sont sélectionnés par des experts du domaine de la papyrologie et se composent de 50 échantillons d'écriture en grec sur des papyrus datant approximativement du 6e siècle après J.-C., produits par 10 scribes différents. Les feuilles de palmier sont un autre support pertinent pour l'écriture dans les cultures asiatiques. Lontar Sunda [145] est un ensemble de données contenant des manuscrits sundanais écrits à la main sur des feuilles de palmier datant du XVe siècle. Il comprend 66 pages provenant de 27 collections de Garut, Java occidental, Indonésie. Les informations de verification incluent des images binarisées, des annotations au niveau des mots et des annotations au niveau des

caractères. L'ensemble de données SleukRith [146] contient des manuscrits khmers sur feuilles de palmier, comprenant 657 pages de manuscrits khmers sur feuilles de palmier sélectionnées de manière aléatoire parmi différentes collections de qualité variable.

2.5.2. Plateformes expérimentales

En plus de la disponibilité de grands ensembles d'entraînement dans de nombreux domaines et de l'accès à une puissance de calcul relativement bon marché fournie par les GPU, le troisième composant qui a alimenté l'utilisation accrue des architectures d'apprentissage profond dans plusieurs domaines est le développement de puissantes bibliothèques open-source pour l'entraînement des modèles.
Dans le domaine de la reconnaissance des documents historiques, plusieurs approches ont utilisé des bibliothèques standard telles que PyTorch [143] (par exemple, dans la référence [137]), TensorFlow [144] et Keras [145] (par exemple, les références [22,77]). Dans d'autres cas, des outils spécifiques à la reconnaissance des documents historiques ont récemment été proposés afin de fournir aux chercheurs des outils facilement accessibles et efficaces pour gérer les différentes étapes nécessaires au développement d'applications efficaces.

eScriptorium [146] est une plateforme open source pour l'analyse et l'annotation de documents historiques. Elle permet aux utilisateurs intéressés de télécharger des collections de documents, de les transcrire et de les segmenter manuellement ou automatiquement à l'aide d'un moteur OCR. HistCorp [147] est une plateforme de distribution de corpus (collectés à partir de corpus historiques) et d'autres ressources et outils utiles dans un format uniforme et normalisé. Elle est conçue pour la distribution de textes historiques de différentes périodes et genres pour plusieurs langues européennes. HInDoLA [152] est une plateforme unifiée basée sur le cloud pour l'annotation, la visualisation et l'analyse de mise en page de manuscrits historiques basée sur l'apprentissage automatique [151]. Elle propose une interface utilisateur graphique d'annotation adaptée, un tableau de bord d'analyse graphique et des interfaces avec certains modules basés sur l'apprentissage automatique.

Un élément important dans le développement d'applications efficaces d'apprentissage automatique est la possibilité de mesurer de manière appropriée les performances de la solution proposée. Un outil d'évaluation ouvert pour l'analyse de la mise en page est présenté dans la référence [148]. Cet outil normalise l'évaluation des tâches d'analyse de mise en page au niveau des pixels. Il est disponible à la fois sous la forme d'une application Java autonome et d'un service web RESTful.

2.5.3. Compétitions

Comme mentionné à la fin de la section précédente, l'évaluation de manière standardisée des techniques de reconnaissance de documents est essentielle pour mesurer les progrès de la recherche en reconnaissance de documents historiques. Les compétitions organisées lors

des principales conférences dans le domaine ont été essentielles à la fois pour la collecte de jeux de données de référence et pour la conception de mesures de performance.

Une des premières compétitions sur la reconnaissance de documents historiques a été organisée par Antonacopoulos et al. lors du 2e atelier international sur l'imagerie et le traitement des documents historiques (HIP2013) qui s'est tenu en conjonction avec l'ICDAR 2013, avec pour tâche l'analyse de mise en page des journaux (HNLA 2013) [149]. Avec la multiplication des applications de l'apprentissage approfondi pour l'analyse d'images de documents, comme discuté dans la Section 1, il n'est pas surprenant que plusieurs compétitions aient été organisées lors de l'ICDAR 2019. La compétition ICDAR 2019 sur la recherche d'images pour les documents manuscrits historiques [153] étudie les performances de la recherche à grande échelle d'images de documents historiques basée sur le style d'écriture [152]. Les ensembles de données fournis par des institutions du patrimoine culturel et des bibliothèques numériques contiennent au total 20 000 images de documents réparties en trois classes principales (manuscrits, lettres, chartes et documents juridiques) réalisées par environ 10 000 rédacteurs.

De ce bref résumé des compétitions récentes sur l'analyse de documents historiques, il est une fois de plus évident que le nombre de documents disponibles pour l'entraînement et le test des méthodes basées sur l'apprentissage approfondi a considérablement augmenté ces dernières années.

2.5.4. Génération de données synthétiques

L'augmentation de données est aujourd'hui considérée comme l'une des techniques les plus efficaces lorsqu'il s'agit de petits ensembles de données ou de données déséquilibrées. Cette technique consiste à augmenter de manière synthétique la quantité de données disponibles, de sorte qu'elles puissent être utilisées pour entraîner des modèles d’apprentissage profond qui bénéficieront probablement de ce processus.

Comme mentionné précédemment, l'augmentation de données implique l'existence et l'utilisation de données synthétiquement générées. Comme son nom l'indique, un ensemble de données synthétiques contient des données générées à l'aide d'outils logiciels spécifiques. Ces données ne sont donc pas collectées lors d'une enquête ou d'une expérience réelle : leur principal objectif est d'être suffisamment flexibles et riches pour améliorer les performances des modèles d’apprentissage profond.

2.6. Discussion

La révolution de l'apprentissage en profondeur a touché plusieurs domaines de recherche et l'analyse et la reconnaissance de documents historiques ne font pas exception. Un grand nombre d'articles ont été publiés sur ce sujet au cours des dernières années, avec une augmentation significative du nombre de tâches abordées et des techniques considérées.

Alors que les recherches précédentes sur la reconnaissance automatique d'images de documents historiques se concentraient principalement sur la reconnaissance de texte et la détection de mots-clés, les applications récentes couvrent des sujets nouveaux allant des manuscrits datant de la localisation de texte sur des cartes. Plusieurs travaux se concentrent désormais sur l'identification de lignes de texte individuelles, tandis que précédemment, l'analyse de la mise en page était davantage axée sur la segmentation de régions dans des documents principalement imprimés.

Les tâches abordées par les méthodes analysées dans ce chapitre sont si diverses qu'il est difficile d'établir une comparaison significative des résultats obtenus, car différentes données et différentes mesures de performance sont proposées lors de la description des techniques destinées à résoudre des tâches similaires. En général, les méthodes basées sur l'apprentissage en profondeur parviennent, dans presque tous les cas analysés, à approcher voire à améliorer les performances et la faisabilité de l'état de l'art. La plus grande difficulté, toujours rencontrée, réside certainement dans la recherche d'une quantité de données variées, cohérentes et suffisantes pour entraîner des modèles basés sur l’apprentissage profond.

Chapitre 03 : La contribution

Introduction

Ce chapitre présente notre contribution qui se concentre sur l'idée de utiliser les caractéristiques locales extraites à l'aide des méthodes SIFT (Scale-Invariant Feature Transform) et ORB (Oriented FAST and Rotated BRIEF) [70], ainsi que les caractéristiques d'un modèle pré-entraîné de réseau de neurones convolutifs (CNN) [72] tel que VGG (Visual Geometry Group). L'objectif [152] est de construire un vecteur représentatif de l'image, qui sera ensuite utilisé par un modèle implémenté avec PyTorch pour évaluer les performances de la datation des documents historiques.

3.1. Introduction

La datation des documents historiques [82] est une tâche complexe qui nécessite une analyse minutieuse des caractéristiques présentes dans ces documents, tels que le style d'écriture, le type de papier utilisé et d'autres éléments contextuels. L'utilisation de techniques de vision par ordinateur et de l'apprentissage automatique peut apporter une contribution précieuse à cette tâche, en permettant une évaluation objective et précise de l'âge des documents. Les méthodes SIFT et ORB sont couramment utilisées pour extraire les caractéristiques locales d'une image, notamment en ce qui concerne la détection de points d'intérêt et la description des descripteurs. Ces méthodes offrent une robustesse aux transformations géométriques et permettent d'extraire des informations discriminantes à partir des images. Cependant, elles peuvent ne pas être suffisamment sensibles aux caractéristiques spécifiques des documents historiques, telles que les styles d'écriture anciens ou les variations de l'encre.

D'autre part, les modèles CNN pré-entraînés, tels que VGG, sont capables d'apprendre des représentations abstraites et discriminantes à partir des images grâce à leur capacité à extraire des caractéristiques de haut niveau. Ces modèles ont été entraînés sur de grandes bases de données et peuvent fournir des informations précieuses sur le contenu des images. Cependant, ils peuvent ne pas être spécifiquement adaptés à la datation des documents historiques et peuvent nécessiter une adaptation pour cette tâche spécifique. Ainsi, notre approche consiste à combiner ces deux types de caractéristiques, les caractéristiques locales et les caractéristiques extraites par un modèle CNN pré-entraîné, afin de construire un vecteur représentatif de document. Ce vecteur sera utilisé par un modèle implémenté avec PyTorch pour évaluer les performances de la datation des documents historiques en utilisant l'ensemble de données KERTAS [154].

L'ensemble de données KERTAS est une ressource précieuse qui regroupe des manuscrits historiques arabes datant de différentes périodes de l'histoire islamique. Ces documents présentent des défis uniques en termes de datation en raison de l'ambiguïté des dates d'écriture et de la variation des styles d'écriture au fil du temps. Notre approche vise à exploiter ces caractéristiques spécifiques des documents historiques pour améliorer la précision de la datation.

Dans la suite de ce chapitre, nous détaillerons notre méthode de combinaison des caractéristiques, en expliquant les étapes de prétraitement et d'extraction des caractéristiques locales, ainsi que la manière dont nous intégrons ces caractéristiques avec le modèle CNN pré-entraîné. Nous présenterons également les résultats expérimentaux obtenus sur l'ensemble de données KERTAS, en mettant en évidence les performances de notre approche pour la datation des documents historiques.

3.2. Description de l'ensemble de données

3.2.1. Definition

KERTAS est un ensemble de données pour les manuscrits historiques arabes, se compose de plus de 2000 images numériques de haute qualité et haute résolution acquises du 1er au 14e siècle AH. Chaque classe contient des manuscrits du même siècle ; par conséquent, il y a 14 classes dans l'ensemble de données. Un résumé de la répartition numérique des documents dans KERTAS et le nombre d'images que nous avons utilisées pour l'entraînement et les tests sont présentés dans le Tableau 15. nous avons utilisé 80% de la base de données pour l'entraînement et 20% pour les tests.

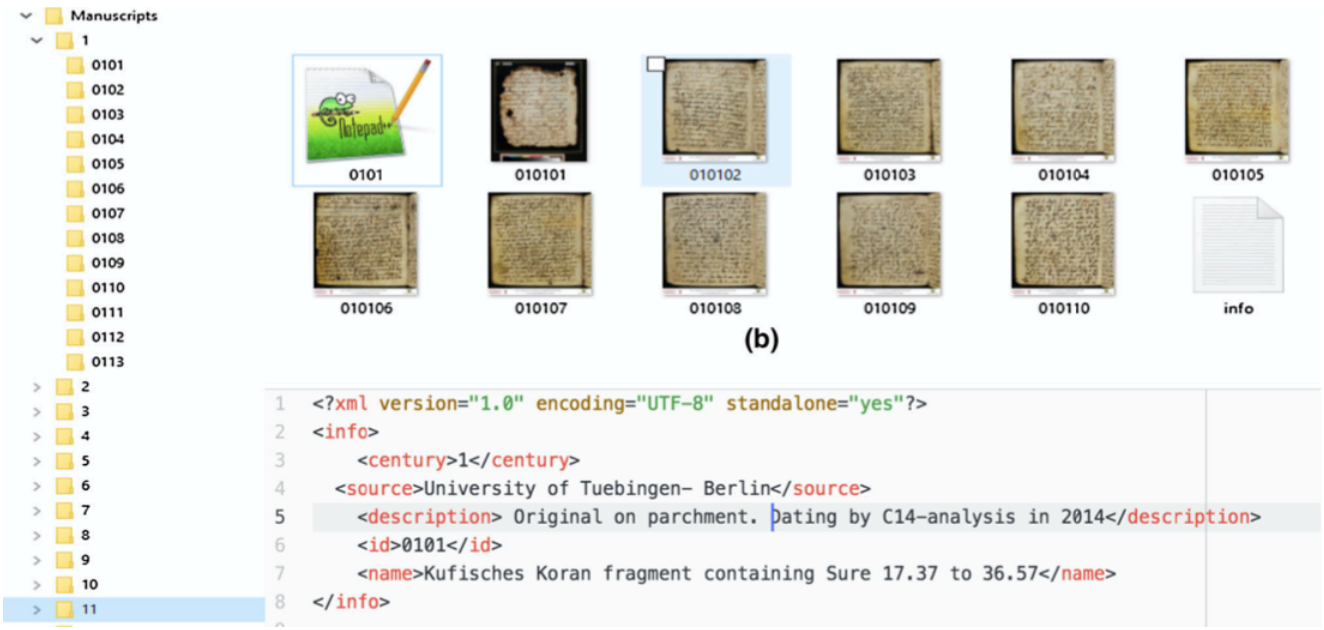


Figure 15 : Structure du répertoire pour la base de données KERTAS

Siècle	Nombre de documents	Entraînement	Test
1	60	48	12
2	47	37	10
3	144	116	28
4	592	474	118
5	164	132	32
6	119	95	24
7	184	147	37
8	110	88	22
9	153	123	30
10	73	59	14
11	169	135	34
12	147	118	29
13	119	95	24
14	17	14	3

Tableau 8. Résumé de la distribution numérique des documents dans l'enseble de données KERTAS.

3.2.2. Ensemble de données KERTAS

L'ensemble de données KERTAS [154] a été spécialement conçu pour faciliter l'entraînement et le test efficaces d'algorithmes de détection de l'auteur et de l'âge des documents. Cela étant dit, l'ensemble de données est suffisamment diversifié et volumineux pour être également utile dans le test d'autres algorithmes tels que les algorithmes de segmentation de documents, d'extraction de lignes et de mots. Les images sélectionnées pour faire partie de l'ensemble de données sont également choisies parmi un ensemble diversifié de documents couvrant des manuscrits sur les mathématiques, la physique, l'histoire islamique, la métaphysique, etc. Cette diversité des manuscrits offre des images uniques et stimulantes qui peuvent être utilisées pour tester les limites des algorithmes en cours de test. Certains images de manuscrits sur les mathématiques contiennent des dessins de figures et de formes, tandis que d'autres contiennent des tableaux et des listes intégrés dans le texte.

Figure 16 : Manuscrit coranique ancien détenu par la bibliothèque de l'Université de Birmingham

Ces figures et tableaux sont généralement tracés avec une encre de couleur différente, plus claire que le texte. De plus, l'ensemble de données contient également des images de pages

avec des commentaires dans les marges écrits par différents auteurs et dans différents styles d'écriture. La Figure 6a, b montre des exemples des deux cas. Ces images peuvent constituer un défi particulier pour les algorithmes d'identification de l'auteur et peuvent aider à la conception d'algorithmes robustes d'identification de l'auteur.

Nom des sources	Nombre de manuscrits
BRILL via QNL	57
Université de Tübingen, Berlin	1
Dar al Makhtotat Sanaa Yémen	1
Institut de culture orientale, Université de Tokyo	8
Bibliothèque de l'Université de Princeton	4
Bienvenue Bibliothèque	2
Université de Yale Cambridge	4
Bibliothèque de l'Université	3
Site Web de sensibilisation islamique	13
La Bibliothèque royale, nationale	12

Tableau 9 : ensembles de données KERTAS

Nom	Langue	Taille
Caractère syriaque	syriaque	60 000 caractères
IBN SINA	arabe	51 folios, 20722 CC
Historique de Barcelone	Espagnol	244 livres, 174 images
Ensemble de données sur les mariages	Néerlandais médiéval	2858 chartes
(BH2 M)	arabe	2505 images, 135 livre

Tableau 10 : Ensembles de données comparables à KERTAS

3.3. Extraction des caractéristiques

3.3.1. Extraction des caractéristiques locales

L'approche consiste à extraire les descripteurs SIFT et ORB de chaque image, puis à les quantifier en utilisant un processus de codification. La quantification des descripteurs permet de réduire leur nombre et de les représenter sous forme de vecteurs de codes, qui servent à former un codebook pour chaque type de caractéristique. Le codebook est construit en regroupant les vecteurs de codes à l'aide d'un algorithme de clustering, tel que le K-means.

Pour créer un vecteur de représentation d'image à l'aide d'un codebook de descripteurs SIFT, vous pouvez suivre les étapes suivantes :

1. Extraction des descripteurs SIFT/ORB : L'utilisation de l'algorithme SIFT/ORB pour détecter et extraire les descripteurs locaux de chaque image de l' ensemble de données. Chaque descripteur SIFT/ORB représente une région d'intérêt de l'image.

2. Quantification des descripteurs : L'application d'une technique de quantification pour réduire la dimensionnalité des descripteurs SIFT/ORB. La méthode K-means utilsée pour former un codebook en regroupant les descripteurs similaires.

3. Codification des descripteurs : Pour chaque image, on attribue à chaque descripteur SIFT/ORB extrait le code (ou l'index) correspondant du codebook. Ainsi, chaque descripteur sera représenté par un code qui indique son appartenance à un certain groupe de descripteurs similaires.

4. Construction du vecteur de représentation : Comptage de nombre d'occurrences de chaque code dans l'image et la création d'un vecteur de représentation en utilisant ces comptages. Chaque dimension du vecteur correspondra à un code du codebook.

5. Normalisation du vecteur : Une normalisation sur le vecteur de représentation pour réduire les variations dues aux différences de taille ou d'intensité des images. Avec l'utilisation de technique de normalisation par amplitude.

En utilisant ce processus, nous pouvons obtenir deux vecteurs de représentation compact et informatif pour chaque image, en se basant sur les descripteurs SIFT et ORB et leur association avec le codebook. Ces deux vecteurs de représentation avec le vecteur de caractéristiques globale issu de CNN pré-entrainé, seront utilisés par un modèle implémenté avec PyTorch.

3.3.2. Extraction des caractéristiques globales

Pour extraire les caractéristiques globales de l'image en utilisant un modèle CNN pré-entraîné comme ResNet / VGG, on va suivre les étapes suivantes :

1. Prétraitement de l'image : Avant de passer l'image au modèle CNN, nous devons effectuer un prétraitement pour mettre l'image à l'échelle et normaliser les valeurs des pixels conformément aux exigences du modèle.

2. Passage de l'image à travers le modèle : Les couches convolutives du modèle extrairont automatiquement les caractéristiques globales de l'image à différents niveaux d'abstraction.

3. Extraction des caractéristiques : Récupération des sorties des couches convolutives du modèle. Ces sorties représentent les caractéristiques extraites de l'image à différents niveaux de profondeur du modèle.

4. Réduction de dimension : réduction de la dimension des caractéristiques extraites, par l'application de technique de l'analyse en composantes principales (PCA).

Le vecteur issu de l'extraction des caractéristiques globales par un modèle CNN pré-entraîné VGG sera utilisé avec les resultats de l'extraction des caractéristiques locales SIFT et ORB.

3.3.3. Combinaison des vecteurs de caractéristiques locales et globales

L'utilisation de techniques d'extraction de caractéristiques locales et globales est essentielle pour obtenir des représentations complètes et riches des images. Les méthodes traditionnelles, telles que SIFT (Scale-Invariant Feature Transform) et ORB (Oriented FAST and Rotated BRIEF), permettent d'extraire des informations locales robustes à partir d'images. D'autre part, les réseaux de neurones convolutionnels (CNN) pré-entraînés, tels que VGG (Visual Geometry Group), sont capables de capturer des caractéristiques globales de haut niveau dans les images

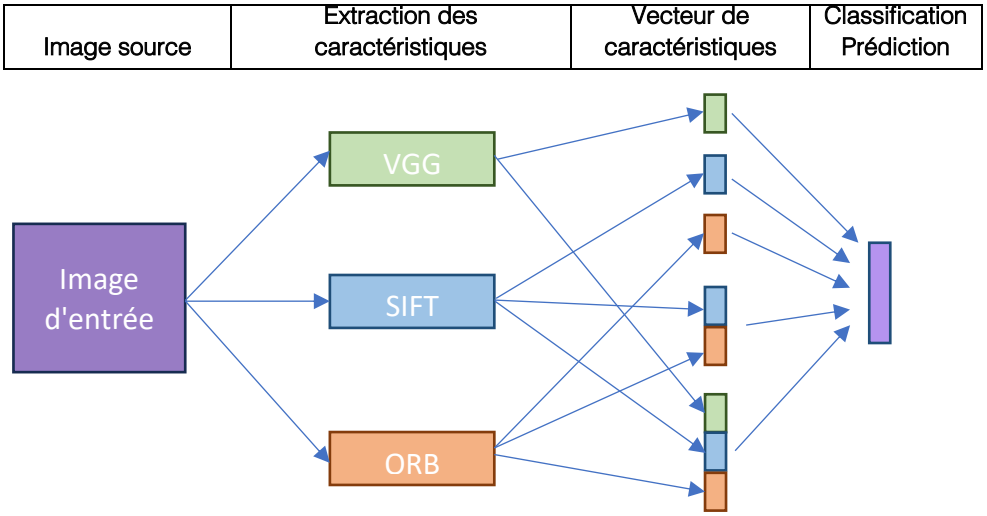


Figure 17 : Combinaison des vecteurs de caractéristiques locales et globales

Dans ce contexte, l'utilisation des vecteurs descripteurs issus de l'extraction des caractéristiques locales et globales devient une approche prometteuse pour améliorer les performances des modèles d'apprentissage automatique en exploitant les avantages complémentaires de ces deux types de caractéristiques. En combinant les informations locales précises avec les caractéristiques globales de haut niveau, pour obtenir une représentation plus riche et discriminante des images.

3.4. Proposition de model d'entrainement

Dans le domaine de l'analyse d'images et de la datation des documents historiques, l'utilisation de modèles basés sur des réseaux de neurones convolutif est devenue une approche courante. PyTorch, une bibliothèque populaire d'apprentissage automatique, offre des outils puissants pour la création des modèles d'entraînement.

Un modèle PyTorch est une architecture de réseau de neurones qui prend des caractéristiques en entrée et effectue des opérations de classification et/ou de régression pour prédire des labels ou des valeurs numériques. Dans le contexte de la classification et de

la datation des documents historiques, le modèle tabulaire peut être utilisé pour prédire l'âge d'un document à partir de ses caractéristiques visuelles extraites précédemment.

Le modèle est composé de plusieurs couches, notamment des couches linéaires, des couches d'activation et des couches de sortie. Les couches linéaires effectuent des opérations de multiplication matricielle pour combiner les caractéristiques d'entrée, tandis que les couches d'activation introduisent des non-linéarités dans le modèle, permettant ainsi de capturer des relations complexes entre les caractéristiques et la cible de prédiction.

Pour évaluer les performances du modèle, on utilise des mesures d'évaluation telles que l'exactitude (accuracy), la précision (precision), le rappel (recall) et la mesure F1. Ces mesures permettent de quantifier la capacité du modèle à classer correctement les documents historiques et à estimer précisément leur âge.

Dans ce chapitre de contribution, nous proposons d'utiliser un modèle PyTorch pour la classification et la détermination de l'âge des documents historiques. Nous combinons les vecteurs descripteurs extraits des caractéristiques locales (SIFT, ORB) et globales (CNN pré-entraîné VGG) pour former une représentation combinée de chaque document (Figure 17). Cette représentation sera ensuite utilisée comme entrée pour le modèle tabulaire afin de prédire la classe et l'âge du document.

L'objectif est d'évaluer l'efficacité de cette approche de modèle pour la classification et la datation des documents historiques, en utilisant l'ensemble de données KERTAS. Les résultats obtenus nous permettront de mieux comprendre la performance de ce modèle dans le contexte spécifique de l'analyse des documents historiques.

3.4.1. Configuration du modele Pytorch

Les différentes parties de modele :

1. La classe `MulticlassClassification` est définie comme une sous-classe de `nn.Module`, qui est une classe de base pour tous les modules de réseaux de neurones dans PyTorch.

2. Dans la méthode `__init__`, la classe définit les couches linéaires et les autres modules nécessaires pour le réseau. Les paramètres `num_feature` et `num_class` sont respectivement le nombre d'entrées de caractéristiques (ou de dimensions) et le nombre de classes de sortie pour la classification.

3. Le réseau est composé de trois couches linéaires (`nn.Linear`) avec des tailles de sortie différentes. La première couche a `num_feature` entrées et 512 sorties, la deuxième couche a 512 entrées et 128 sorties, et la troisième couche a 128 entrées et 64 sorties. La dernière

couche (`layer_out`) a 64 entrées et `num_class` sorties, correspondant aux probabilités de chaque classe de sortie.

4. Les activations ReLU (`nn.ReLU()`) sont utilisées après chaque couche linéaire pour introduire de la non-linéarité dans le réseau.

5. Des couches de dropout (`nn.Dropout(p=0.2)`) sont utilisées après la deuxième et la troisième couche linéaire pour régulariser le réseau en désactivant aléatoirement certains neurones lors de l'entraînement, ce qui peut aider à prévenir le surapprentissage.

6. Des couches de normalisation par lot (`nn.BatchNorm1d`) sont utilisées après chaque couche linéaire pour normaliser les activations des couches précédentes, ce qui peut accélérer et stabiliser l'apprentissage.

7. Dans la méthode `forward`, les opérations de propagation avant (forward propagation) du réseau sont définies. Les entrées `x` sont passées à travers chaque couche linéaire, suivies de la normalisation par lot, de l'activation ReLU et de la couche de dropout correspondantes.

8. Finalement, les sorties du réseau sont calculées en passant les activations de la dernière couche linéaire (`layer_out`) sans utiliser d'activation supplémentaire.

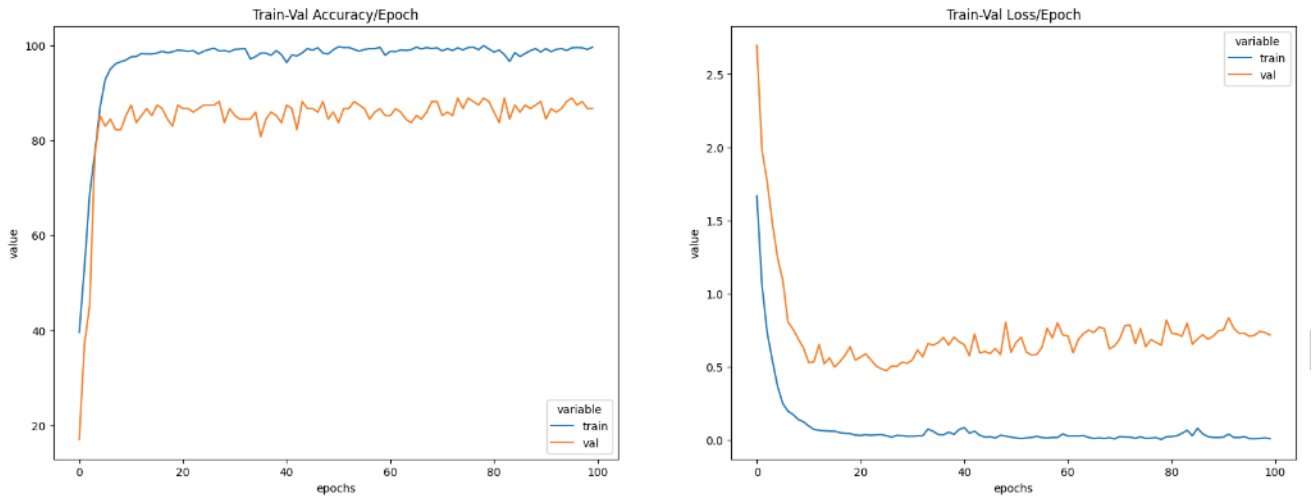
9. Les sorties du réseau sont renvoyées en tant que résultat de la méthode `forward`.

Ce modèle peut être utilisé pour entraîner un réseau de neurones pour la classification multi-classes en utilisant des données d'entrée avec `num_feature` caractéristiques et en prédisant les probabilités des `num_class` classes de sortie.

5. Résultats

Descripteurs des caractéristiques	Classifieur	Dimenssion de vecteur	Nombre de vecteurs	Accuracy Test
SIFT (Clustred)	PyTorch	512	1686	90.23%
ORB (Clustred)		512	1686	60.05%
VGG19		256	1686	53.46%
SIFT+ORB		1024	1686	85.20%
SIFT+ORB+VGG		1280	1686	73.26%

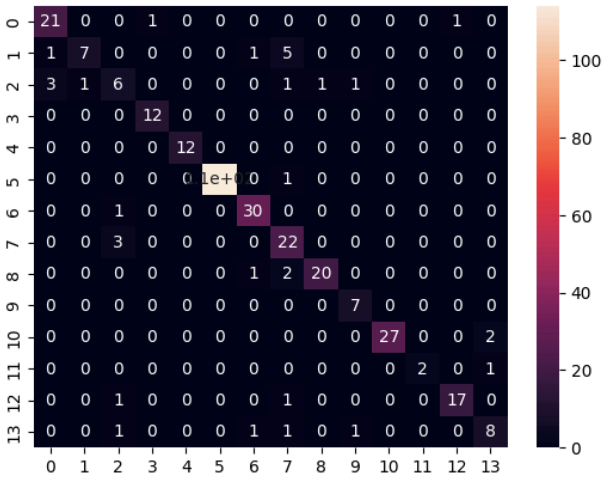
Tableau 11 : Comparaison des performances



Training-Validation Perte/Epoque

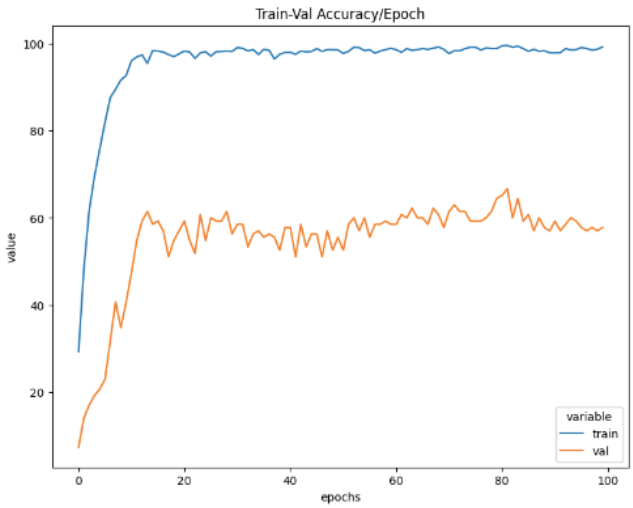
	precision	recall	f1-score	support
0.0	0.84	0.91	0.87	23
1.0	0.88	0.50	0.64	14
2.0	0.50	0.46	0.48	13
3.0	0.92	1.00	0.96	12
4.0	1.00	1.00	1.00	12
5.0	1.00	0.99	1.00	115
6.0	0.91	0.97	0.94	31
7.0	0.67	0.88	0.76	25
8.0	0.95	0.87	0.91	23
9.0	0.78	1.00	0.88	7
10.0	1.00	0.93	0.96	29
11.0	1.00	0.67	0.80	3
12.0	0.94	0.89	0.92	19
13.0	0.73	0.67	0.70	12
accuracy			0.90	338
macro avg	0.87	0.84	0.84	338
weighted avg	0.91	0.90	0.90	338

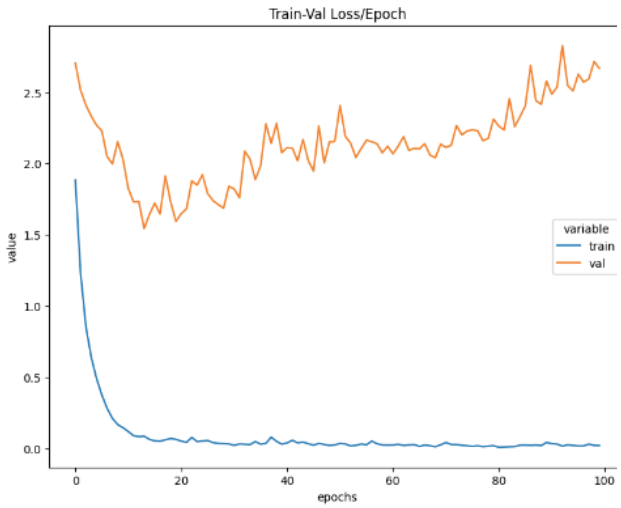
Rapport de classification



Matrice de confusion du réseau proposé

Clustred SIFT / PyTorch

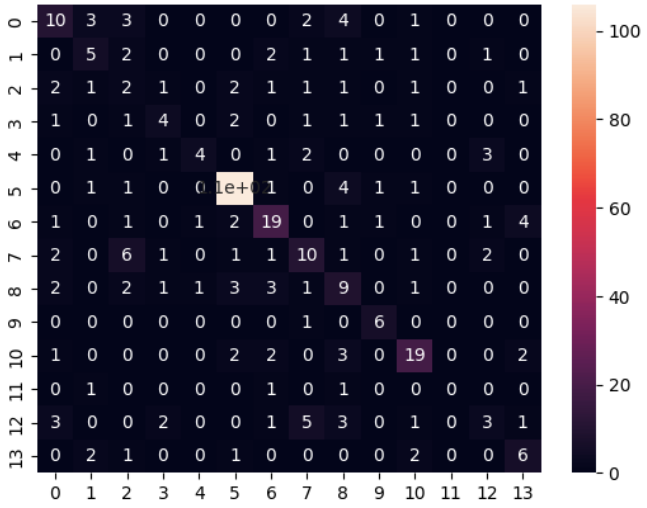




Training-Validation Perte/Epoque

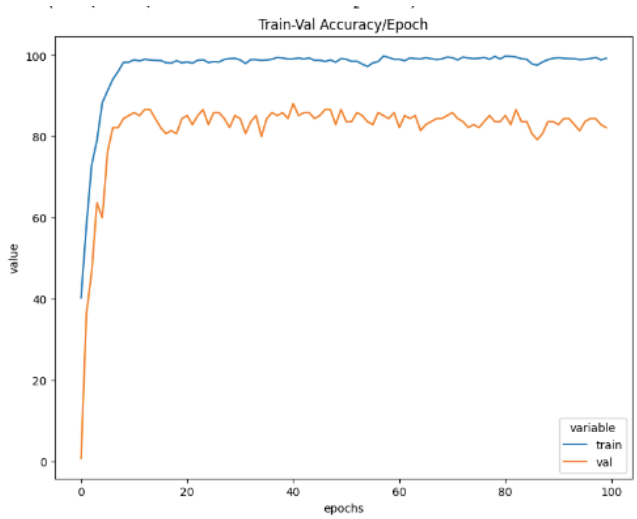
	precision	recall	f1-score	support
0.0	0.45	0.43	0.44	23
1.0	0.36	0.36	0.36	14
2.0	0.11	0.15	0.12	13
3.0	0.40	0.33	0.36	12
4.0	0.67	0.33	0.44	12
5.0	0.89	0.92	0.91	115
6.0	0.59	0.61	0.60	31
7.0	0.42	0.40	0.41	25
8.0	0.31	0.39	0.35	23
9.0	0.60	0.86	0.71	7
10.0	0.66	0.66	0.66	29
11.0	0.00	0.00	0.00	3
12.0	0.30	0.16	0.21	19
13.0	0.43	0.50	0.46	12
accuracy			0.60	338
macro avg	0.44	0.44	0.43	338
weighted avg	0.60	0.60	0.59	338

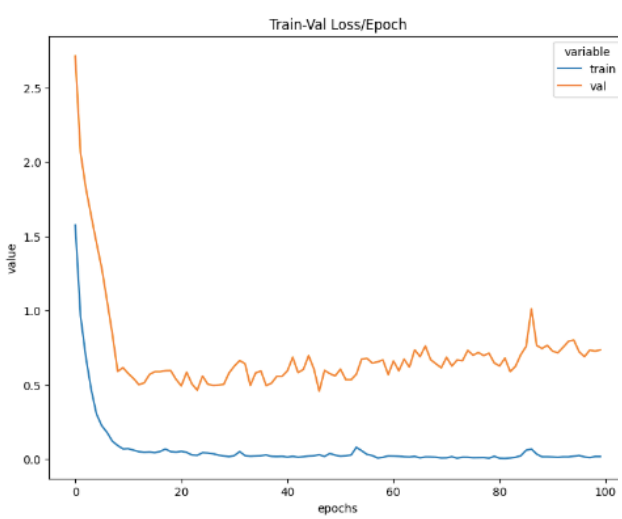
Rapport de classification



Matrice de confusion du réseau proposé

Clustred ORB / PyTorch

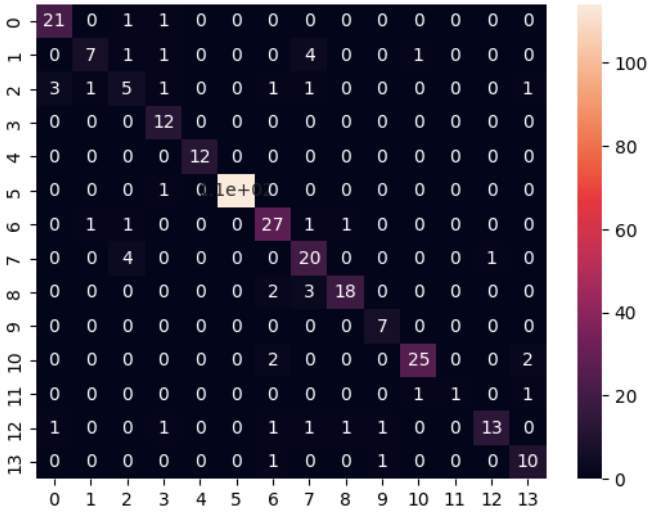




Training-Validation Perte/Epoque

	precision	recall	f1-score	support
0.0	0.84	0.91	0.87	23
1.0	0.78	0.50	0.61	14
2.0	0.42	0.38	0.40	13
3.0	0.71	1.00	0.83	12
4.0	1.00	1.00	1.00	12
5.0	1.00	0.99	1.00	115
6.0	0.79	0.87	0.83	31
7.0	0.67	0.80	0.73	25
8.0	0.90	0.78	0.84	23
9.0	0.78	1.00	0.88	7
10.0	0.93	0.86	0.89	29
11.0	1.00	0.33	0.50	3
12.0	0.93	0.68	0.79	19
13.0	0.71	0.83	0.77	12
accuracy			0.86	338
macro avg	0.82	0.78	0.78	338
weighted avg	0.87	0.86	0.86	338

Rapport de classification



Matrice de confusion du réseau proposé

Clustred SIFT+ORB / PyTorch

6. Conclusions

La comparaison des méthodes d'extraction des caractéristiques locales et globales, y compris SIFT, ORB, VGG, ainsi que les combinaisons SIFT+ORB et SIFT+ORB+VGG, a permis de déterminer leurs performances respectives dans l'estimation de la date d'un document historique à partir de la dataset KERTAS, en utilisant un modèle CNN implémenté avec PyTorch. Les résultats ont révélé que la méthode SIFT a obtenu la meilleure précision, avec un score de 90,23%. Cette méthode a réussi à extraire efficacement les caractéristiques distinctives des images de documents historiques, ce qui lui a permis d'estimer avec précision la date associée à chaque document.

La méthode ORB a affiché une précision inférieure, avec un score de 60,05%. Bien que cette méthode ait montré une vitesse d'exécution plus rapide, elle a eu du mal à capturer les détails complexes des images de documents historiques, ce qui a conduit à une précision plus faible.

La méthode VGG, basée sur un réseau de neurones profonds, a obtenu une précision de 53,46%. Bien que les modèles d'apprentissage profond aient démontré leur capacité à extraire des caractéristiques riches et complexes, ils ont rencontré des difficultés dans l'estimation.

En combinant les méthodes SIFT et ORB, nous avons observé une amélioration de la précision, atteignant un score de 85,20%. Cette combinaison a exploité les forces complémentaires des deux méthodes, permettant ainsi d'obtenir de meilleurs résultats dans l'estimation de la date des documents historiques. Cependant, lorsque la méthode VGG a été ajoutée à la combinaison SIFT+ORB, la précision a diminué pour atteindre 73,26%. Cela suggère que l'incorporation de la méthode VGG n'a pas apporté d'amélioration significative à la précision de l'estimation de la date des documents historiques dans ce contexte spécifique.

En résumé, notre travail met en évidence l'efficacité de la méthode SIFT dans l'estimation précise de la date des documents historiques de l'ensemble de données KERTAS. La combinaison SIFT+ORB a également montré de bons résultats, tandis que l'ajout de la méthode VGG n'a pas contribué de manière significative à l'amélioration des performances.

Ces résultats fournissent des informations précieuses pour guider les travaux futurs sur l'estimation de la date des documents historiques. Ils soulignent également l'importance de choisir la méthode appropriée en fonction des caractéristiques spécifiques des données et des objectifs de la tâche.

Des recherches supplémentaires sont nécessaires pour explorer de nouvelles combinaisons de méthodes et affiner les approches existantes afin de développer des modèles encore plus performants pour l'estimation de la date des documents historiques. L'utilisation de techniques avancées d'apprentissage automatique et l'exploration d'autres caractéristiques spécifiques aux documents historiques pourraient être des directions prometteuses à suivre.

Bibliographie

Bibliographie

[1] Y. Y. Tang, S.-W. Lee, and C. Y. Suen, 'Automatic document process¬ing: a survey,' Pattern recognition, vol. 29, no. 12, pp. 1931.1952, 1996.

[2] D. Ghosh, T. Dube, and A. Shivaprasad, 'Script recognition—a review,' IEEE Transactions on pattern analysis and machine intelligence, vol. 32, no. 12, pp. 2142.2161, 2010.

[3] D. Sinwar, V. S. Dhaka, N. Pradhan, and S. Pandey, 'Offline script recognition from handwritten and printed multilingual documents: a survey,' International Journal on Document Analysis and Recognition (IJDAR), vol. 24, no. 1, pp. 97-121, 2021.

[4] A. Chaudhuri, K. Mandaviya, P. Badelia, and S. K. Ghosh, 'Optical character recognition systems,' in Optical Character Recognition Sys-tems for Different Languages with Soft Computing. Springer, 2017, pp. 9-41.

[5] 'Optical character recognition systems,' in Optical Character Recognition Systems for Different Languages with Soft Computing. Springer, 2017, pp. 9-41.

[6] N. Islam, Z. Islam, and N. Noor, 'A survey on optical character recognition system,' arXiv preprint arXiv:1710.05703, 2017.

[7] R. Smith et al., 'Tesseract ocr engine,' Lecture. Google Code. Google Inc, 2007. [Online]. Available: https://github.com/tesseract-ocr/ tesseract#tesseract-ocr

[8] D. Doermann, K. Tombre et al., 'Handbook of document image pro-cessing and recognition'. Springer, 2014, vol. 1.

[9] C. C. Tappert, C. Y. Suen, and T. Wakahara, 'The state of the art in online handwriting recognition,' IEEE Transactions on pattern analysis and machine intelligence, vol. 12, no. 8, pp. 787-808, 1990.

[10] R. Smith, D. Antonova, and D.-S. Lee, 'Adapting the tesseract open source ocr engine for multilingual ocr,' in Proceedings of the Interna-tional Workshop on Multilingual OCR, 2009, pp. 1.8.

[11] N. Samadiani and H. Hassanpour, 'A neural network-based approach for recognizing multi-font printed english characters,' Journal of Elec¬trical Systems and Information Technology, vol. 2, no. 2, pp. 207-218, 2015.

[12] F. Slimane, R. Ingold, and J. Hennebert, 'Icdar2017 competition on multi-font and multi-size digitally represented arabic text,' in 201714th IAPR International Conference on Document Analysis and Recognition (ICDAR), vol. 1. IEEE, 2017, pp. 1466-1472.

[13] A. A. Aburas and M. E. Gumah, 'Arabic handwriting recognition: Challenges and solutions,' in 2008 International Symposium on In-formation Technology, vol. 2. IEEE, 2008, pp. 1.6.

[14] H. Singh, R. K. Sharma, and V. Singh, 'Online handwriting recognition systems for indic and non-indic scripts: a review,' Artificial Intelligence Review, vol. 54, no. 2, pp. 1525-1579, 2021.

[15] Y. Zheng, H. Li, and D. Doermann, 'Text identification in noisy doc-ument images using markov random model,' in Seventh International Conference on Document Analysis and Recognition,. Proceedings. IEEE, 2003, pp. 599-603.

[16] A. Farahmand, H. Sarrafzadeh, and J. Shanbehzadeh, 'Document image noises and removal methods,' International MultiConference of Engineers and Computer Scientists 2013. Proceedings., vol. 1, 2013.

[17] M. Agrawal and D. Doermann, 'Stroke-like pattern noise removal in binary document images,' in 2011 International Conference on Document Analysis and Recognition. IEEE, 2011, pp. 17-21.

[18] J. Liang, D. Doermann, and H. Li, 'Camera-based analysis of text and documents: a survey,' International Journal of Document Analysis and Recognition (IJDAR), vol. 7, no. 2, pp. 84-104, 2005.

[19] Z. Zhu, L. Gao, Y. Li, Y. Huang, L. Du, N. Lu, and X. Wang, 'Ntable: A dataset for camera-based table detection,' in International Conference on Document Analysis and Recognition. Springer, 2021, pp. 117-129.

[20] M. Li, L. Cui, S. Huang, F. Wei, M. Zhou, and Z. Li, 'Tablebank: Table benchmark for image-based table detection and recognition,' in Proceedings ofThe 12th language resources and evaluation conference,2020,pp. 1918-1925.

[21] H. Singh, R. K. Sharma, and V. Singh, 'Online handwriting recognition systems for indic and non-indic scripts: a review,' Artificial Intelligence Review, vol. 54, no. 2, pp. 1525-1579, 2021.

[22] W. Bieniecki, S. Grabowski, and W. Rozenberg, 'Image preprocessing for improving ocr accuracy,' in 2007 international conference on perspective technologies and methods in MEMS design. IEEE, 2007, pp. 75-80.

[23] H. P. Le and G. Lee, 'Noise removal from binarized text images,' in 2010 The 2nd International Conference on Computer and Automation Engineering (ICCAE), vol. 3. IEEE, 2010, pp. 586-589.

[24] L. Xu, E. Oja, and P. Kultanen, 'A new curve detection method: randomized hough transform (rht),' Pattern recognition letters, vol. 11, no. 5, pp. 331.338, 1990.

[25] A. Chakraborty and M. Blumenstein, 'Marginal noise reduction in his-torical handwritten documents-a survey,' in 2016 12th IAPR Workshop on Document Analysis Systems (DAS). IEEE, 2016, pp. 323.328.

[26] S.-C. Pei, M. Tzeng, and Y.-Z. Hsiao, 'Enhancement of uneven lighting text image using line-based empirical mode decomposition,' in 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2011, pp. 1249-1252.

[27] F. Z. A. Bella, M. El Rhabi, A. Hakim, and A. Laghrib, 'Reduction of the non-uniform illumination using nonlocal variational models for document image analysis,' Journal of the Franklin Institute, vol. 355, no. 16, pp. 8225-8244, 2018.

[28] C. Simon, I. K. Park et al., 'Correcting geometric and photometric distortion of document images on a smartphone,' Journal ofElectronic Imaging, vol. 24, no. 1, p. 013038, 2015.

[29] W. Niblack,' An introduction to digital image processing'. Birkerod: Strandberg Publishing Company, 1985.

[30] J. Sauvola and M. Pietikainen, 'Adaptive document image binarization,' Pattern Recognit, vol. 33, 2000. [Online]. Available: https://doi.org/10.1016/S0031. 3203(99)00055- 2

[31] N. Otsu, 'A threshold selection method from gray-level histograms,'Trans Syst Man Cybern, vol. 9,1979. [Online]. Available: https://doi.org/10.1109/TSMC.1979.4310076

[32] F. Kasmin, A. Abdullah, and A. S. Prabuwono, 'Ensemble of steerable local neighbourhood grey-level information for binarization,' Pattern Recognit Lett., vol. 98, 2017. [Online]. Available: https: //doi.org/10.1016/j.patrec.2017.07.014

[33] W. Xiong, J. Xu, Z. Xiong, J. Wang, and M. Liu, 'Degraded historical document image binarization using local features and support vector machine (svm),' Optik, vol. 164, 2018. [Online]. Available: https://doi.org/10.1016/jijleo.2018.02.072

[34] I. B. Messaoud, H. El Abed, H. Amiri, and V. Margner, 'New method for the selection of binarization parameters based on noise features of historical documents,' in Proceedings of the 2011 Joint Workshop on Multilingual OCR and Analytics for Noisy Unstructured Text Data, 2011, pp. 1.8.

[35] K. Kise, A. Sato, and M. Iwata, 'Segmentation of page images using the area voronoi diagram,' Computer Vision and Image Understanding, vol. 70, no. 3, pp. 370-382, 1998.

[36] F. Y. Shih and S.-S. Chen, 'Adaptive document block segmentation and classification,' IEEE transactions on systems, man, and cybernetics, part B (cybernetics), vol. 26, no. 5, pp. 797-802, 1996.

[37] H. Wei, M. Baechler, F. Slimane, and R. Ingold, 'Evaluation of svm, mlp and gmm classifiers for layout analysis of historical documents,' in 2013 12th International Conference on Document Analysis and Recognition. IEEE, 2013, pp. 1220-1224.

[38] S. Marinai, M. Gori, and G. Soda, 'Artificial neural networks for document analysis and recognition,' IEEE Transactions on pattern analysis and machine intelligence, vol. 27, no. 1, pp. 23.35, 2005.

[39] A. Antonacopoulos, 'Page segmentation using the description of the background,' Computer Vision and Image Understanding, vol. 70, no. 3, pp. 350-369, 1998.

[40] K. Y. Wong, R. G. Casey, and F. M. Wahl, 'Document analysis system,' IBM journal of research and development, vol. 26, no. 6, pp. 647-656, 1982.

[41] L. O'Gorman, 'The document spectrum for page layout analysis,' IEEE Transactions on pattern analysis and machine intelligence, vol. 15, no. 11, pp. 1162.1173, 1993.

[42] D. J. Ittner and H. S. Baird, 'Language-free layout analysis,' in Proceedings of 2nd International Conference on Document Analysis and Recognition (ICDAR'93). IEEE, 1993, pp. 336-340.

[43] A. Dias, 'Minimum spanning trees for text segmentation,' in Proc. of Fifth Annual Symposium on Document Analysis and Information Retrieval, Las Vegas, Nevada, 1996.

[44] M. Aiello, C. Monz, L. Todoran, M. Worring et al., 'Document understanding for a broad class of documents,' International Journal on Document Analysis and Recognition, vol. 5, no. 1, pp. 1.16, 2002.

[45] Y. Lu, Z. Wang, and C. L. Tan, 'Word grouping in document images based on voronoi tessellation,' in Document Analysis Systems VI: 6th International Workshop, DAS 2004, Florence, Italy, September 8-10, 2004. Proceedings 6. Springer, 2004, pp. 147-157.

[46] M. Arivazhagan, H. Srinivasan, and S. Srihari, 'A statistical approach to line segmentation in handwritten documents,' in Document recognition and retrieval XIV, vol. 6500. SPIE, 2007, pp. 245-255.

[47] S. Jindal and G. S. Lehal, 'Line segmentation of handwritten gurmukhi manuscripts,' in Proceeding of the workshop on document analysis and recognition, 2012, pp. 74-78.

[48] R. Saabni, A. Asi, and J. El-Sana, 'Text line extraction for historical document images,' Pattern Recognit Lett, vol. 35, 2014.

[49] C. A. Boiangiu, M. C. Tanase, and R. Ioanitescu, 'Handwritten documents text line segmentation based on information energy,' Int J Comput Commun Control, vol. 9, 2014.

[50] N. Arvanitopoulos and S. Siisstrunk, 'Seam carving for text line extraction on color and grayscale historical manuscripts,' in 2014 14th International Conference on Frontiers in Handwriting Recognition. IEEE, 2014, pp. 726-731.

[51] U. Garain, T. Paquet, and L. Heutte, 'On foreground-background sep-aration in low quality color document images,' in Eighth International Conference on Document Analysis and Recognition (ICDAR'05), 2005, pp. 585-589 Vol. 2.
[52] S. S. Bukhari, F. Shafait, and T. M. Breuel, 'Text-line extraction using a convolution of isotropic gaussian filter with a set of line filters,' in 2011 International Conference on Document Analysis and Recognition. IEEE, 2011, pp. 579-583.
[53] K. Wong, R. Casey, and F. Wahl, 'Document analysis systems,' IBM J Res Dev, vol. 26, 1982.
[54] D. J. Kennard and W. A. Barrett, 'Separating lines of text in free-form handwritten historical documents,' in Second International Conference on Document Image Analysis for Libraries (DIAL'06). IEEE, 2006, pp. 12.pp.
[55] R. Gomathi, R. S. Uma, and S. Mohanval, 'Segmentation of touching, overlapping, skewed and short handwritten text lines,' Int J Comput Appl, vol. 49, 2012.
[56] E. J. Almazan, R. Tal, Y. Qian, and J. H. Elder, 'Mcmlsd: A dynamic programming approach to line segment detection,' in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 2031.2039.
[57] E. Hussain, A. Hannan, and K. Kashyap, 'A zoning based feature extraction method for recognition of handwritten assamese characters,' Int J Comput Sci Technol, vol. 6, 2015.
[58] L. Likforman-Sulem, A. Hanimyan, and C. Faure, 'A hough based algorithm for extracting text lines in handwritten documents,' in Proceedings of 3rd international conference on document analysis and recognition, vol. 2. IEEE, 1995, pp. 774-777.
[59] G. Louloudis, B. Gatos, I. Pratikakis, and C. Halatsis, 'Text line and word segmentation of handwritten documents,' Pattern Recognit, vol. 42, 2009.
[60] S. Jetley, S. Belhe, V. K. Koppula, and A. Negi, 'Two-stage hybrid binarization around fringe map based text line segmentation for docu-ment images,' in Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012). IEEE, 2012, pp. 343.346.
[61] D. Brodic, 'Text line segmentation with water flow algorithm based on power function,' J Electr Eng, vol. 66, 2015.
[62] Y.-H. Tseng and H.-J. Lee, 'Recognition-based handwritten chinese character segmentation using a probabilistic viterbi algorithm,' Pattern Recognition Letters, vol. 20, no. 8, pp. 791.806, 1999.
[63] D. C. Ciresan, U. Meier, L. M. Gambardella, and J. Schmidhuber, 'Convolutional neural network committees for handwritten character classification,' in 2011 International conference on document analysis and recognition. IEEE, 2011, pp. 1135-1139.
[64] D. Chen, J.-M. Odobez, and H. Bourlard, 'Text detection and recogni-tion in images and video frames,' Pattern recognition, vol. 37, no. 3, pp. 595-608, 2004.
[65] J. Park, G. Lee, E. Kim, J. Lim, S. Kim, H. Yang, M. Lee, and S. Hwang, 'Automatic detection and recognition of korean text in outdoor signboard images,' Pattern Recognition Letters, vol. 31, no. 12, pp. 1728-1739, 2010.
[66] P. Shivakumara, W. Huang, and C. L. Tan, 'Efficient video text detection using edge features,' in 2008 19th International Conference on Pattern Recognition. IEEE, 2008, pp. 1.4.
[67] M. Li and C. Wang, 'An adaptive text detection approach in images and video frames,' in 2008 IEEE International Joint Conference on Neu¬ral Networks (IEEE World Congress on Computational Intelligence). IEEE, 2008, pp. 72.77.
[68] S. P. Deore and A. Pravin, 'Histogram of oriented gradients based off-line handwritten devanagari characters recognition using svm, k-nn and nn classifiers.' Rev. d'Intelligence Artif., vol. 33, no. 6, pp. 441.446, 2019.
[69] A. L. Spitz, 'Determination of the script and language content of doc-ument images,' IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 19, no. 3, pp. 235-245, 1997.
[70] N. Kato, M. Suzuki, S. Omachi, H. Aso, and Y. Nemoto, 'A hand-written character recognition system using directional element feature and asymmetric mahalanobis distance,' IEEE transactions on pattern analysis and machine intelligence, vol. 21, no. 3, pp. 258-262, 1999.
[71] P Natarajan, Z. Lu, R. Schwartz, I. Bazzi, and J. Makhoul, 'Multilin-gual machine printed ocr,' International Journal of Pattern Recognition and Artificial Intelligence, vol. 15, no. 01, pp. 43.63, 2001.
[72] M. Amrouch, M. Rabi, and Y. Es-Saady, 'Convolutional feature learning and cnn based hmm for arabic handwriting recognition,' in International conference on image and signal processing. Springer, 2018, pp. 265-274.
[73] H. Choudhury, S. Mandal, S. Devnath, S. M. Prasanna, and S. Sun- daram, 'Combining hmm and svm based stroke classifiers for online assamese handwritten character recognition,' in 2015 Annual IEEE India Conference (INDICON). IEEE, 2015, pp. 1-6.
[74] L. Guichard, A. H. Toselli, and B. Coiiasnon, 'Handwritten word verification by svm-based hypotheses re-scoring and multiple thresh-olds rejection,' in 2010 12th International Conference on Frontiers in Handwriting Recognition. IEEE, 2010, pp. 57-62.

[75] Z. Chi and K. Wong, 'A two-stage binarization approach for document images,' in Proceedings of 2001 International Symposium on Intelli¬gent Multimedia, Video and Speech Processing. ISIMP 2001 (IEEE Cat. No. 01EX489). IEEE, 2001, pp. 275-278.
[76] S. Gunter and H. Bunke, 'Ensembles of classifiers for handwritten word recognition,' Int. J. Doc. Anal. Recognit., vol. 5, 2003. [Online]. Available: https://doi.org/10.1007/s10032. 002. 0088-2
[77] O. Alsharif and J. Pineau, 'End-to-end text recognition with hybrid hmm maxout models,' arXiv preprint arXiv:1310.1811, 2013.
[78] M. Carbonell, J. Mas, M. Villegas, A. Fornes, and J. Llados, 'End- to-end handwritten text detection and transcription in full pages,' in 2019 International conference on document analysis and recognition workshops (ICDARW), vol. 5. IEEE, 2019, pp. 29-34.
[79] J. Chung and T. Delteil, 'A computationally efficient pipeline approach to full page offline handwritten text recognition,' in 2019 Interna¬tional conference on document analysis and recognition workshops (ICDARW), vol. 5. IEEE, 2019, pp. 35-40.
[80] T. Mondal, U. Bhattacharya, S. K. Parui, K. Das, and D. Man- dalapu, 'On-line handwriting recognition of indian scripts - the first benchmark,' in 2010 12th International Conference on Frontiers in Handwriting Recognition, 2010, pp. 200-205.
[81] R. Hussain, A. Raza, I. Siddiqi, K. Khurshid, and C. Djeddi, 'A comprehensive survey of handwritten document benchmarks: structure, usage and evaluation,' EURASIP Journal on Image and Video Process¬ing, vol. 2015, no. 1, pp. 1-24, 2015.
[82] K. Nikolaidou, M. Seuret, H. Mokayed, and M. Liwicki, 'A survey of historical document image datasets,' arXiv preprint arXiv:2203.08504, 2022
[83] X. Zhong, E. ShafieiBavani, and A. Jimeno Yepes, 'Image-based table recognition: data, model, and evaluation,' in European Conference on Computer Vision. Springer, 2020, pp. 564-580.
[84] N. Siegel, N. Lourie, R. Power, and W. Ammar, 'Extracting scientific figures with distantly supervised neural networks,' in Proceedings of the 18th ACM/IEEE on joint conference on digital libraries, 2018, pp. 223.232.
[85] B. Pfitzmann, C. Auer, M. Dolfi, A. S. Nassar, and P. W. Staar, 'Doclaynet: A large human-annotated dataset for document-layout analysis,' arXiv preprint arXiv:2206.01062, 2022.
[86] L. Gao, Y. Huang, H. Dejean, J.-L. Meunier, Q. Yan, Y. Fang, F. Kleber, and E. Lang, 'Icdar 2019 competition on table detection and recognition (ctdar),' in 2019 International Conference on Document Analysis and Recognition (ICDAR). IEEE, 2019, pp. 1510-1515.
[87] J. Fang, X. Tao, Z. Tang, R. Qiu, and Y. Liu, 'Dataset, ground-truth and performance metrics for table detection evaluation,' in 2012 10th IAPR International Workshop on Document Analysis Systems. IEEE, 2012, pp. 445-449.
[88] M. Li, Y. Xu, L. Cui, S. Huang, F. Wei, Z. Li, and M. Zhou, 'Docbank: A benchmark dataset for document layout analysis,' arXiv preprint arXiv:2006.01038, 2020.
[89] M. Gobel, T. Hassan, E. Oro, and G. Orsi, 'Icdar 2013 table compe-tition,' in 2013 12th International Conference on Document Analysis and Recognition. IEEE, 2013, pp. 1449-1453.
[90] X. Zhong, J. Tang, and A. J. Yepes, 'Publaynet: largest dataset ever for document layout analysis,' in 2019 International Conference on Document Analysis and Recognition (ICDAR). IEEE, 2019, pp. 1015-1022.
[91] T. Furukawa, 'Recognition of laser-printed characters based on creation of new laser-printed characters datasets,' in International Conference on Document Analysis and Recognition. Springer, 2021, pp. 407-421.
[92] H. S. Kawoosa, M. Singh, M. M. Joshi, and P. Goyal, 'Ncert5k-iitrpr: A benchmark dataset for non-textual component detection in school books,' in International Workshop on Document Analysis Systems. Springer, 2022, pp. 461-475.
[93] J. J. Hull, 'A database for handwritten text recognition research,' IEEE Trans. Pattern Anal. Mach. Intell., vol. 16, 1994. [Online]. Available: https://doi.org/10.1109/34.291440
[94] R. A. Wilkinson, J. Geist, S. Janet, P. J. Grother, C. J. Burges, R. Creecy, B. Hammond, J. J. Hull, N. Larsen, T. P. Vogl et al., 'The first census optical character recognition system conference'. US Depart-ment of Commerce, National Institute of Standards and Technology, 1992, vol. 184.
[95] Y. LeCun, 'The mnist database of handwritten digits,' http://yann. lecun. com/exdb/mnist/, 1998.
[96] U. V. Marti and H. Bunke, 'The iam-database: An english sentence database for offline handwriting recognition,' Int. J. Doc. Anal. Recognit, vol. 5, 2002. [Online]. Available: https: //doi.org/10.1007/s100320200071
[97] E. Augustin, M. Carre, E. Grosicki, J.-M. Brodin, E. Geoffrois, and F. Preteux, 'Rimes evaluation campaign for handwritten mail processing,' in International Workshop on Frontiers in Handwriting Recognition (IWFHR'06),, 2006, pp. 231-235.
[98] in International Workshop on Frontiers in Handwriting Recognition ('IWFHR'06'),, 2006, pp. 231-235.

[99] H. Alamri, J. Sadri, C. Y. Suen, and N. Nobile, 'A novel comprehensive database for arabic off-line handwriting recognition,' in Proceedings of 11th international conference on frontiers in handwriting recognition, ICFHR, vol. 8, 2008, pp. 664-669.
[100] R. Pardeshi, B. Chaudhuri, M. Hangarge, and K. Santosh, 'Automatic handwritten indian scripts identification,' in 2014 fourteenth interna-tional conference on frontiers in handwriting recognition. IEEE, 2014, pp. 375-380.
[101] G. A. Daniyar Nurseitov, Kairat Bostanbekov, Maksat Kanatov, Anel Alimova, Abdelrahman Abdallah, 'Classification of Handwritten Names of Cities and Handwritten Text Recognition using Various Deep Learning Models,' Advances in Science, Technology and Engineering Systems Journal, vol. 5, no. 5, pp. 934-943, 2020.
[102] D.-H. KIM, Y.-S. Hwang, S.-T. Park, E.-J. Kim, S.-H. Paek, and S.-Y. BANG, 'Handwritten korean character image database pe92,' IEICE transactions on information and systems, vol. 79, no. 7, pp. 943.950, 1996.
[103] I. Pratikakis, B. Gatos, and K. Ntirogiannis, 'H-dibco 2010-handwritten document image binarization competition,' in 2010 12th International Conference on Frontiers in Handwriting Recognition. IEEE, 2010, pp. 727-732.
[104] K. Ntirogiannis, B. Gatos, and I. Pratikakis, 'Icfhr2014 competition on handwritten document image binarization (h-dibco 2014),' in 2014 14th International conference on frontiers in handwriting recognition. IEEE, 2014, pp. 809-813.
[105] A. Antonacopoulos, B. Gatos, and D. Bridson, 'Icdar2005 page segmentation competition,' in Eighth International Conference on Document Analysis and Recognition (ICDAR'05). IEEE, 2005, pp. 75-79.
[106] Marinai, S. 'Introduction to Document Analysis and Recognition. In Machine Learning in Document Analysis and Recognition'; Marinai, S., Fujisawa, H., Eds.; Springer: Berlin/Heidelberg, Germany, 2008; Volume 90, pp. 1–20, doi:10.1007/978-3.540-76280-5_1.
[107] Lecun, Y.; Bottou, L.; Bengio, Y.; Haffner, P. 'Gradient-based learning applied to document recognition'. Proc. IEEE 1998, 86, 2278–2324.
[108] Suryani, M.; Paulus, E.; Hadi, S.; Darsa, U.A.; Burie, J.C. 'The handwritten sundanese palm leaf manuscript dataset from 15th century'. In Proceedings of the 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), Kyoto, Japan, 9–15 November 2017; Volume 1, pp. 796–800.
[109] Valy, D.; Verleysen, M.; Chhun, S.; Burie, J.C. 'A new Khmer Palm leaf manuscript dataset for document analysis and recognition': SleukRith set. In Proceedings of the 4th International Workshop on Historical Document Imaging and Processing, Kyoto, Japan, 10–11 November 2017; pp. 1–6.
[110] Stromer, D.; Christlein, V.; Maier, A.; Zippert, P.; Helmecke, E.; Hausotte, T.; Huang, X. 'Non-destructive Digitization of Soiled Historical Chinese Bamboo Scrolls'. In Proceedings of the 13th IAPR International Workshop on Document Analysis Systems, DAS, Vienna, Austria, 24–27 April 2018; pp. 55–60, doi:10.1109/DAS.2018.37.
[111] Mohammed, H.; Marthot-Santaniello, I.; Margner, V. GRK-Papyri: 'A Dataset of Greek Handwriting on Papyri for the Task of Writer Identification'. In Proceedings of the 2019 International Conference on Document Analysis and Recognition, ICDAR 2019, Sydney, Australia, 20–25 September 2019; pp. 726–731, doi:10.1109/ICDAR.2019.00121.
[112] Alaasam, R.; Kurar, B.; El-Sana, J. 'Layout Analysis on Challenging Historical Arabic Manuscripts using Siamese Network'. In Proceedings of the 2019 International Conference on Document Analysis and Recognition (ICDAR), Sydney, Australia, 20–25 September 2019, pp. 738–742.
[113] Clanuwat, T.; Lamb, A.; Kitamoto, A. KuroNet: 'Pre-Modern Japanese Kuzushiji Character Recognition with Deep Learning'. In Proceedings of the 2019 International Conference on Document Analysis and Recognition, ICDAR 2019, Sydney, Australia, 20–25 September 2019; pp. 607–614, doi:10.1109/ICDAR.2019.00103.
[114] Hamid, A.; Bibi, M.; Moetesum, M.; Siddiqi, I.' Deep Learning Based Approach for Historical Manuscript Dating'. In Proceedings of the 2019 International Conference on Document Analysis and Recognition, ICDAR 2019, Sydney, Australia, 20–25 September 2019; pp. 967–972, doi:10.1109/ICDAR.2019.00159.
[115] Toselli, A.; Romero, V.; Sanchez, J.A.; Vidal, E. 'Making Two Vast Historical Manuscript Collections Searchable and Extracting Meaningful Textual Features Through Large-Scale Probabilistic Indexing'. In Proceedings of the 2019 International Conference on Document Analysis and Recognition, ICDAR 2019, Sydney, Australia, 20–25 September 2019; pp. 108–113, doi:10.1109/ICDAR.2019.00026.
[116] Yin, X.; Aldarrab, N.; Megyesi, B.; Knight, K. 'Decipherment of Historical Manuscript Images'. In Proceedings of the 2019 International Conference on Document Analysis and Recognition, ICDAR 2019, Sydney, Australia, 20–25 September 2019; pp. 78–85, doi:10.1109/ICDAR.2019.00022.
[117] Watanabe, K.; Takahashi, S.; Kamaya, Y.; Yamada, M.; Mekada, Y.; Hasegawa, J.; Miyazaki, S.' Japanese Character Segmentation for Historical Handwritten Official Documents Using Fully Convolutional Networks'. In Proceedings of the 2019 International Conference on Document Analysis and Recognition, ICDAR 2019, Sydney, Australia, 20–25 September 2019; pp. 934–940, doi:10.1109/ICDAR.2019.00154.
[118] Ziran, Z.; Pic, X.; Innocenti, S.U.; Mugnai, D.; Marinai, S. 'Text alignment in early printed books combining deep learning and dynamic programming'. Pattern Recognit. Lett. 2020, 133, 109–115, doi:10.1016/j.patrec.2020.02.016.

[119] Grüning, T.; Labahn, R.; Diem, M.; Kleber, F.; Fiel, S. Read-bad: 'A new dataset and evaluation scheme for baseline detection in archival documents'. In Proceedings of the 2018 13th IAPR International Workshop on Document Analysis Systems (DAS), Vienna, Austria, 24–27 April 2018; pp. 351–356.
[120] Saini, R.; Dobson, D.; Morrey, J.; Liwicki, M.; Liwicki, F. ICDAR 2019 'Historical Document Reading Challenge on Large Structured Chinese Family Records'. In Proceedings of the 2019 International Conference on Document Analysis and Recognition, ICDAR 2019, Sydney, Australia, 20–25 September 2019; pp. 1499–1504, doi:10.1109/ICDAR.2019.00241.
[121] Capobianco, S.; Marinai, S. 'Deep neural networks for record counting in historical handwritten documents'. Pattern Recognit. Lett. 2019, 119, 103–111, doi:10.1016/j.patrec.2017.10.023.
[122] Weinman, J.; Chen, Z.; Gafford, B.; Gifford, N.; Lamsal, A.; Niehus-Staab, L. 'Deep Neural Networks for Text Detection and Recognition in Historical Maps'. In Proceedings of the 2019 International Conference on Document Analysis and Recognition, ICDAR 2019, Sydney, Australia, 20–25 September 2019; pp. 902–909, doi:10.1109/ICDAR.2019.00149.
[123] Cilia, N.D.; Stefano, C.D.; Fontanella, F.; Marrocco, C.; Molinara, M.; di Freca, A.S.' An end-to-end deep learning system for medieval writer identification'. Pattern Recognit. Lett. 2020, 129, 137–143, doi:10.1016/j.patrec.2019.11.025.
[124] Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; Chen, L.C. Mobilenetv2: 'Inverted residuals and linear bottlenecks'. In Proceedings of the IEEE conference on computer vision and pattern recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 4510–4520.
[125] Adak, C.; Marinai, S.; Chaudhuri, B.B.; Blumenstein, M. 'Offline Bengali Writer Verification by PDF-CNN and Siamese Net'. In Proceedings of the 13th IAPR International Workshop on Document Analysis Systems, DAS 2018, Vienna, Austria, 24–27 April 2018; pp. 381–386.
[126] He, S.; Sammara, P.; Burgers, J.; Schomaker, L. 'Towards Style-Based Dating of Historical Documents'. In Proceedings of the 14th International Conference on Frontiers in Handwriting Recognition, ICFHR 2014, Crete, Greece, 1–4 September 2014; pp. 265–270, doi:10.1109/ICFHR.2014.52.
[127] Sánchez, J.A. Bentham Dataset R0. 'Available online': https://zenodo.org/record/44519 (15 October 2020).
[128] Nguyen, K.; Nguyen, C.; Hotta, S.; Nakagawa, M. A 'Character Attention Generative Adversarial Network for Degraded Historical Document Restoration'. In Proceedings of the 2019 International Conference on Document Analysis and Recognition, ICDAR 2019, Sydney, Australia, 20–25 September 2019; pp. 420–425, doi:10.1109/ICDAR.2019.00074.
[129] Uzan, L.; Dershowitz, N.; Wolf, L. 'Qumran Letter Restoration by Rotation and Reflection Modified PixelCNN'. Iapr Int. Conf. Doc. Anal. Recognit. 2017, 1, 23–29.
[130] Lafferty, J.; McCallum, A.; Pereira, F.C. Conditional random fields: 'Probabilistic models for segmenting and labeling sequence data. Proceedings of the Eighteenth International Conference on Machine Learning (ICML)', Williams College, Williamstown, MA, USA, 28 June –1 July 2001.
[131] Chen, K.; Seuret, M.; Hennebert, J.; Ingold, R. 'Convolutional neural networks for page segmentation of historical document images'. In Proceedings of the 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), Kyoto, Japan, 9–15 November 2017; Volume 1, pp. 965–970.
[132] Pastor-Pellicer, J.; Afzal, M.Z.; Liwicki, M.; Castro-Bleda, M.J. 'Complete system for text line extraction using convolutional neural networks and watershed transform'. In Proceedings of the 2016 12th IAPR Workshop on Document Analysis Systems (DAS), Santorini, Greece, 11–14 April 2016; pp. 30–35.
[133] Barakat, B.K.; El-Sana, J. 'Binarization Free Layout Analysis for Arabic Historical Documents Using Fully Convolutional Networks'. In Proceedings of the 2018 IEEE 2nd International Workshop on Arabic and Derived Script Analysis and Recognition (ASAR), London, UK, 12–14 March 2018; pp. 151–155.
[134] Fink, M.; Layer, T.; Mackenbrock, G.; Sprinzl, M. 'Baseline Detection in Historical Documents Using Convolutional U-Nets'. In Proceedings of the 13th IAPR International Workshop on Document Analysis Systems, DAS, Vienna, Austria, 24–27 April 2018; pp. 37–42, doi:10.1109/DAS.2018.34.
[135] Grüning, T.; Leifert, G.; Strauß, T.; Michael, J.; Labahn, R. 'A two-stage method for text line detection in historical documents'. IJDAR 2019, 22, 285–302, doi:10.1007/s10032.019-00332.1.
[136] Frinken, V.; Fischer, A.; Manmatha, R.; Bunke, H. 'A Novel Word Spotting Method Based on Recurrent Neural Networks'. IEEE Trans. Pattern Anal. Mach. Intell. 2012, 34, 211–224, doi:10.1109/TPAMI.2011.113. 62. Romero, V.; Fornés, A.; Serrano, N.; Sánchez, J.; Toselli, A.H.; Frinken, V.; Vidal, E.; Lladós, J. The ESPOSALLES database: An ancient marriage license corpus for off-line handwriting recognition. Pattern Recognit. 2013, 46, 1658–1669, doi:10.1016/j.patcog.2012.11.024.
[137] Nagai, A. 'On the Improvement of Recognizing Single-Line Strings of Japanese Historical Cursive'. In Proceedings of the 2019 International Conference on Document Analysis and Recognition, ICDAR 2019, Sydney, Australia, 20–25 September 2019; pp. 621–628, doi:10.1109/ICDAR.2019.00105.

[138] Xu, Y.; He, W.; Yin, F.; Liu, C. 'Page Segmentation for Historical Handwritten Documents Using Fully Convolutional Networks'. In Proceedings of the 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), Kyoto, Japan, 9–15 November 2017; Volume 1, pp. 541–546.
[139] Ly, N.T.; Nguyen, C.T.; Nakagawa, M. 'Training an End-to-End Model for Offline Handwritten Japanese Text Recognition by Generated Synthetic Patterns'. In Proceedings of the 2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR), Niagara Falls, NY, USA, 5–8 August 2018; pp. 74–79.
[140] Wei, H.; Seuret, M.; Chen, K.; Fischer, A.; Liwicki, M.; Ingold, R. 'Selecting autoencoder features for layout analysis of historical documents'. In Proceedings of the 3rd International Workshop on Historical Document Imaging and Processing, Gammarth, Tunisia, 22 August 2015; pp. 55–62.
[141] Chen, K.; Seuret, M.; Liwicki, M.; Hennebert, J.; Ingold, R. 'Page segmentation of historical document images with convolutional autoencoders'. In Proceedings of the 13th International Conference on Document Analysis and Recognition (ICDAR), Tunis, Tunisia, 23–26 August 2015; pp. 1011–1015.
[142] Martínek, J.; Lenc, L.; Král, P.; Nicolaou, A.; Christlein, V. 'Hybrid Training Data for Historical Text OCR'. In Proceedings of the 15th International Conference on Document Analysis and Recognition (ICDAR 2019), Sydney, Australia, 20–25 September 2019; pp. 565–570, doi:10.1109/ICDAR.2019.00096.
[143] Simistira, F.; Bouillon, M.; Seuret, M.; Würsch, M.; Alberti, M.; Ingold, R.; Liwicki, M. Icdar2017 'competition on layout analysis for challenging medieval manuscripts'. In Proceedings of the 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), Kyoto, Japan, 9–15 November 2017; Volume 1, pp. 1361–1370.
[144] Mehri, M.; Heroux, P.; Mullot, R.; Moreux, J.P.; Coüasnon, B.; Barrett, B.' ICDAR2019 Competition on Historical Book Analysis'—HBA2019. In Proceedings of the 2019 International Conference on Document Analysis and Recognition, ICDAR 2019, Sydney, Australia, 20–25 September 2019; pp. 1488–1493, doi:10.1109/ICDAR.2019.00239.
[145] Gaur, S.; Sonkar, S.; Roy, P.P. 'Generation of synthetic training data for handwritten Indic script recognition'. In Proceedings of the 2015 13th International Conference on Document Analysis and Recognition (ICDAR), Tunis, Tunisia, 23–26 August 2015; pp. 491–495.
[146] Fischer, A.; Visani, M.; Kieu, V.C.; Suen, C.Y. 'Generation of learning samples for historical handwriting recognition using image degradation'. In Proceedings of the 2nd International Workshop on Historical Document Imaging and Processing, HIP@ICDAR 2013, Washington, DC, USA, 24 August 2013; Frinken, V.; Barrett, B.; Manmatha, R.; Märgner, V., Eds.; ACM: New York, NY, USA, 2013; pp. 73–79, doi:10.1145/2501115.2501123.
[147] Journet, N.; Visani, M.; Mansencal, B.; Van-Cuong, K.; Billy, A. DocCreator: 'A New Software for Creating Synthetic Ground-Truthed Document Images'. J. Imag. 2017, 3, 62, doi:10.3390/jimaging3040062. 119. Liwicki, F.S.; Liwicki, M. Deep learning for historical document analysis. In Handbook of Pattern Recognition and Computer Vision; Chen, C.H., Ed.; World Scientific: Singapore City, Singapore, 2020; pp. 287–303.
[148] Alberti,M.;Bouillon,M.;Ingold,R.;Liwicki,M.'Open Evaluation Tool for Layout Analysis of Document', Images. In Proceedings of the 1st International Workshop on Open Services and Tools for Document Analysis, 14th IAPR International Conference on Document Analysis and Recognition, OST@ICDAR 2017, Kyoto, Japan, 9–15 November 2017; pp. 43–47, doi:10.1109/ICDAR.2017.311.
[149] Antonacopoulos,A.;Clausner,C.;Papadopoulos,C.;Pletschacher,S.'ICDAR 2013 Competition on Historical Newspaper Layout Analysis (HNLA 2013)'. In Proceedings of the 2013 12th International Conference on Document Analysis and Recognition, Washington, DC, USA, 35–28 August 2013; pp. 1454–1458.
[150] Omayio, E. O., Indu, S., & Panda, J. (2022). 'Historical manuscript dating: traditional and current trends'. Multimedia Tools and Applications, 81(22), 31573.31602.
[151] Chammas, M., Makhoul, A., Demerjian, J., & Dannaoui, E. (2022). 'A deep learning based system for writer identification in handwritten Arabic historical manuscripts'. Multimedia Tools and Applications, 81(21), 30769-30784.
[152] Adam, K., Al-Maadeed, S., & Akbari, Y. (2022). 'Hierarchical fusion using subsets of multi-features for historical arabic manuscript dating'. Journal of Imaging, 8(3), 60.
[153] Lombardi, F., & Marinai, S. (2020). 'Deep learning for historical document analysis and recognition'—A survey. Journal of Imaging, 6(10), 110.
[154] Adam, K., Baig, A., Al-Maadeed, S., Bouridane, A., & El-Menshawy, S. (2018). 'KERTAS: dataset for automatic dating of ancient Arabic manuscripts'. International Journal on Document Analysis and Recognition (IJDAR), 21, 283.290.
[155] NIGAM, SHIVANGI; Verma, Shekhar; Nagabhushan, P (2023): Document Analysis and Recognition: A survey. TechRxiv. Preprint. https://doi.org/10.36227/techrxiv.22336435.v1