École d'été
« Patrimoine et Numérique »

OCR pour la langue arabe

Ilham CHAKER, Arsalane ZARGHILI, Abdelhay ZOIZOU
FST - USMBA Fès

Plan

- Généralités sur les OCRs
- Classification des OCRs
- OCR arabe
- Processus général d'un OCR arabe
- Exemples d'OCRs arabes

Généralités sur les OCRs

Généralités sur les OCRs

Les textes obtenus sous forme d'images ne permettent pas:

- Les modifications des textes
- Les copier-collers
- La recherche automatique
- ….

Généralités sur les OCRs

Reconnaissance optique de caractères

(En anglais: Optical Character Recognition)

Généralités sur les OCRs

OCR: Définition

- OCR: est une technique permettant de convertir une image numérisée d'un texte **imprimé** ou **manuscrit** en une forme qui peut être **reconnue** et **manipulée** par un éditeur de texte.
- L'OCR implique plusieurs disciplines : Traitement d'images, reconnaissance des formes, traitement du langage naturel, intelligence artificielle, etc.

Généralités sur les OCRs

OCR: Historique

- L'OCR a vu le jour dans la première moitié du 20ème siècle
- **Les années 1950:** l'OCR a trouvé son marché et s'est développé non seulement comme technologie mais comme produit commercial,
- **Les années 1960** : développement de la machine de Jacob Rabinow, permettant de lire et trier les adresses postales américaines,
- **En 2005**: la mise à disposition du premier logiciel OCR libre, Tesseract ouvre la voie à une large diffusion de cette technologie
- **En 2008**: océrisation de documents historiques numérisés à travers le projet européen INPACT

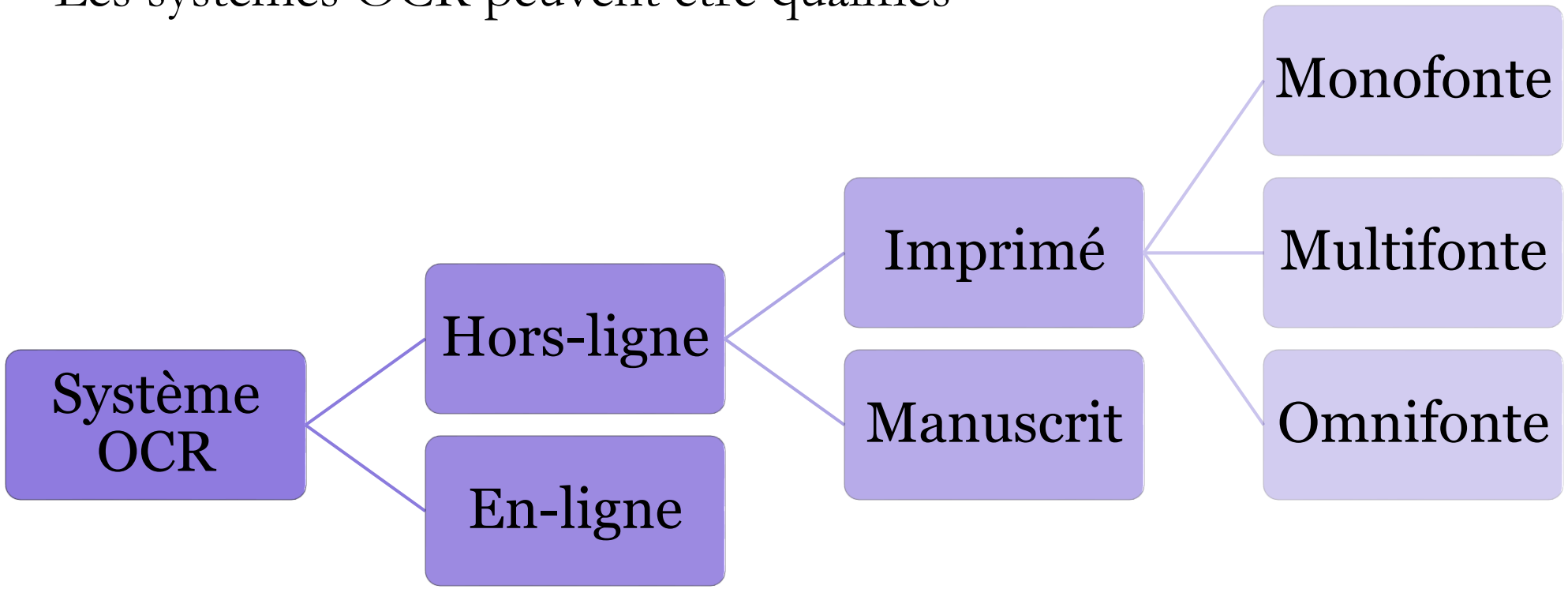
Généralités sur les OCRs

OCR: Domaines d'application

- La lecture de chèques bancaires par la reconnaissance des montants littéraux , des montants numériques manuscrits et le numéro de compte ;
- La lecture automatique de formulaires et de documents administratifs;
- La lecture des adresses postales et le tri automatique des courriers
- L'indexation et l'archivage automatique de documents;
- La recherche d'information dans une base de documents manuscrits historiques
- …

OCR: classification

Les systèmes OCR peuvent être qualifiés



OCR: classification

OCR en-ligne

- Systèmes en ligne = dynamiques
- l'écriture se fait naturellement à l'aide d'un stylet sur une ardoise ou un écran ou par un stylo digital.
- Par conséquent, la reconnaissance de l'écriture est réalisée pendant le processus de l'écriture.

OCR: classification

OCR en-ligne

Ces systèmes sont utilisés dans plusieurs équipements électroniques : Smartphone, iphone, ipad, PDA ou Tablet PC

Écran tactile ou tablette PC, ipad

Smartphone, iphone, Assistant personnel

Stylo doppler, stylo camera

OCR: classification

OCR hors-ligne

- Systèmes hors-lignes = statiques.
- Le texte, qui a été écrit sur un support en papier préalablement, est par la suite numérisé par un scanner ou une caméra, qui permet de convertir l'écriture en image numérique.
- La plupart des recherches se focalisent sur les systèmes OCR hors-lignes.
- La reconnaissance hors-ligne concerne soit:
  - Le texte imprimé
  - Le texte manuscrit

OCR: classification

OCR hors-ligne : texte imprimé

- Dans un texte imprimé, la forme des caractères est définie par un style calligraphique (fonte) qui constitue un modèle pour l'identification.
- La reconnaissance dans ce type d'OCR peut être monofonte, multifonte ou omnifonte
  - Un système est dit monofonte s'il ne peut reconnaître qu'une seule fonte à la fois
  - Un système est dit multifonte s'il est capable de reconnaître divers types de fontes parmi un ensemble de fontes préalablement apprises.
  - Un système omnifonte est capable de reconnaître toute fonte, généralement sans apprentissage préalable. Cependant ceci est quasiment impossible car il existe des milliers de fontes.

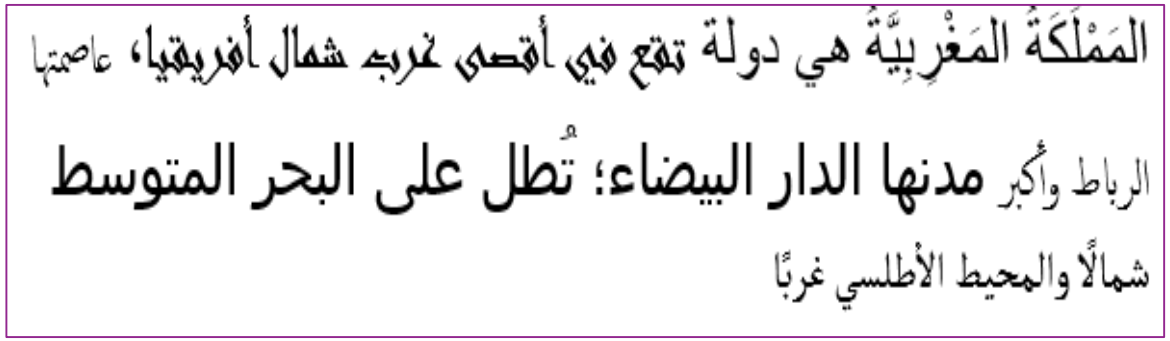
OCR: classification

OCR hors-ligne

Texte Monofonte

المَمْلَكَةُ المَغْرِبِيَّةُ هي دولة تقع في أقصى غرب شمال أفريقيا، عاصمتها الرباط وأكبر مدنها الدار البيضاء؛ تُطل على البحر المتوسط شمالًا والمحيط الأطلسي غربًا

Texte Multifonte



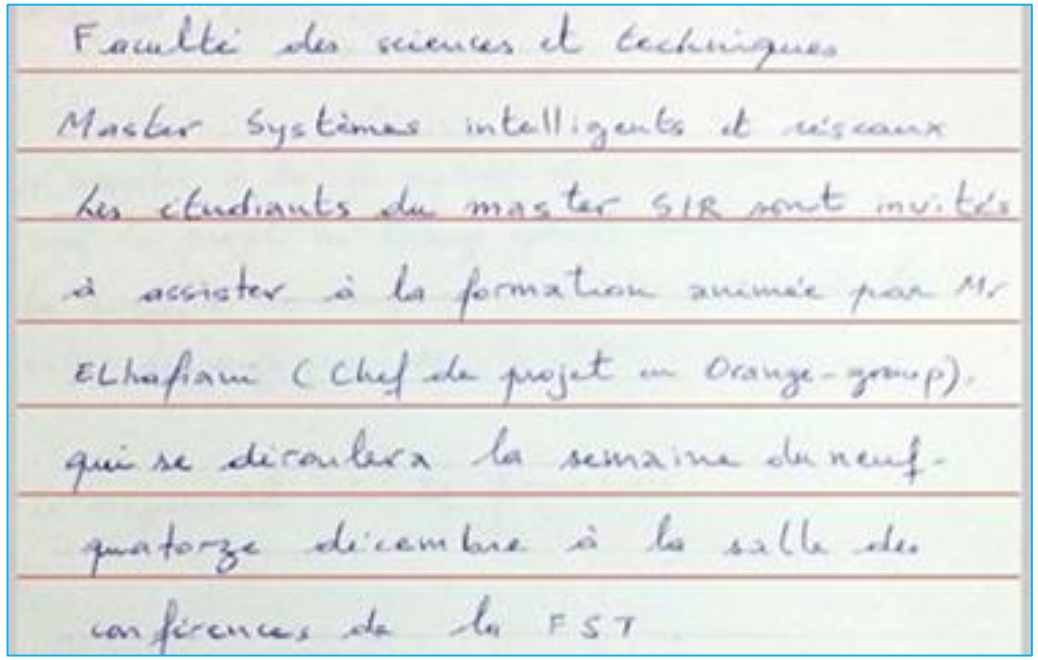
OCR: classification

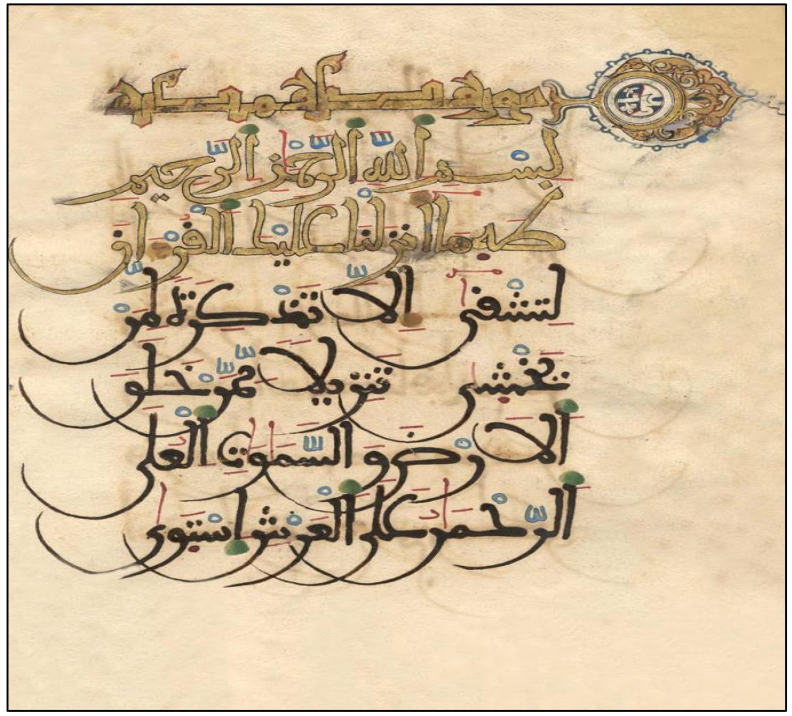
OCR hors-ligne : texte manuscrit

Dans le cas du manuscrit, le graphisme des caractères est inégalement proportionné provenant de la variabilité des scripteurs ou calligraphes. Ce qui complique le processus de traitement de ce type de textes.

OCR: classification

OCR hors-ligne : texte manuscrit


Faculté des sciences et techniques
Master Systèmes intelligents et réseaux
Les étudiants du master SIR sont invités
à assister à la formation animée par Mr
ELhafiani (Chef de projet en Orange-group),
qui se déroulera la semaine du neuf-
quatorze décembre à la salle des
conférences de la FST



OCRs arabes

- Dans les 20 dernières années, des progrès considérables ont été réalisés dans le domaine de la reconnaissance des textes latins
- Pour l'arabe , le problème de la reconnaissance de l'écriture arabe demeure à ce jour non résolu.
- SAKHR est consideré comme le premier OCR arabe apparu en 1997

Les caractéristiques des textes arabes

- La cursivité de l'écriture arabe
- En s'écrivant, les lettres arabes se lient les unes aux autres.
- Cet usage entraîne jusqu'à quatre morphologies différentes d'une même lettre en fonction de son emplacement dans le mot : initiale, médiane, finale et isolée

Name	Isolated	Initial	Medial	Fina
alif	ا	-		ا
baa	ب	بـ	ـبـ	ـب
taa	ت	تـ	ـتـ	ـت
thaa	ث	ثـ	ـثـ	ـث
jiim	ج	جـ	ـجـ	ـج
Haa	ح	حـ	ـحـ	ـح
khaa	خ	خـ	ـخـ	ـخ
daal	د	-		ـد
dhaal	ذ	-		ـذ
raa	ر	-		ـر
zaay	ز	-		ـز
siin	س	سـ	ـسـ	ـس
shiin	ش	شـ	ـشـ	ـش
Saad	ص	صـ	ـصـ	ـص
Daad	ض	ضـ	ـضـ	ـض
Taa	ط	طـ	ـطـ	ـط
Dhaa	ظ	ظـ	ـظـ	ـظ
ayn	ع	عـ	ـعـ	ـع
ghayn	غ	غـ	ـغـ	ـغ
faa	ف	فـ	ـفـ	ـف
qaaf	ق	قـ	ـقـ	ـق
kaaf	ك	كـ	ـكـ	ـك
laam	ل	لـ	ـلـ	ـل
miim	م	مـ	ـمـ	ـم
nuun	ن	نـ	ـنـ	ـن
haa	ه	هـ	ـهـ	ـه
waaw	و	-		ـو
yaa	ي	يـ	ـيـ	ـي

Les caractéristiques des textes arabes

- La présence des points et signes diacritiques

قيل قيل ثمر ثمر شمر شمر

اللُّغَةُ العَرَبِيَّةُ

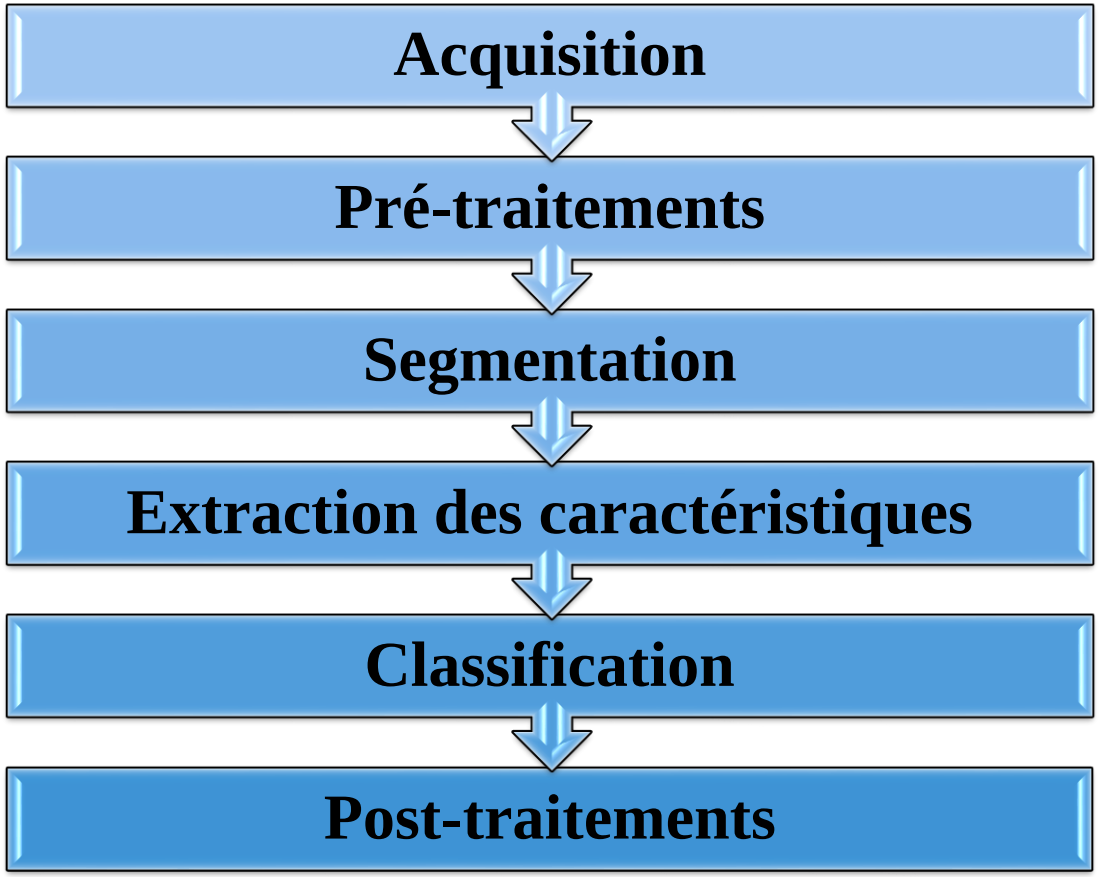
Les caractéristiques des textes arabes

- Les ligatures horizontales et verticales de certains caractères.
- Deux caractères ou plus peuvent être combinés pour construire un nouveau caractère : ligature

لمجة لجنة محمد

Tracé avec ligature	*Tracé normal*	*Séquence*
فـي	فـي	ف + ي
محمد	محمد	م + ح + م + د

Processus général d'un OCR arabe



Processus général d'un OCR arabe

Acquisition

- C'est la phase qui consiste à convertir un document papier à une image numérique en utilisant des capteurs physiques (Scanner, caméra..).
- L'acquisition est caractérisée par la résolution et le niveau d'éclairage.

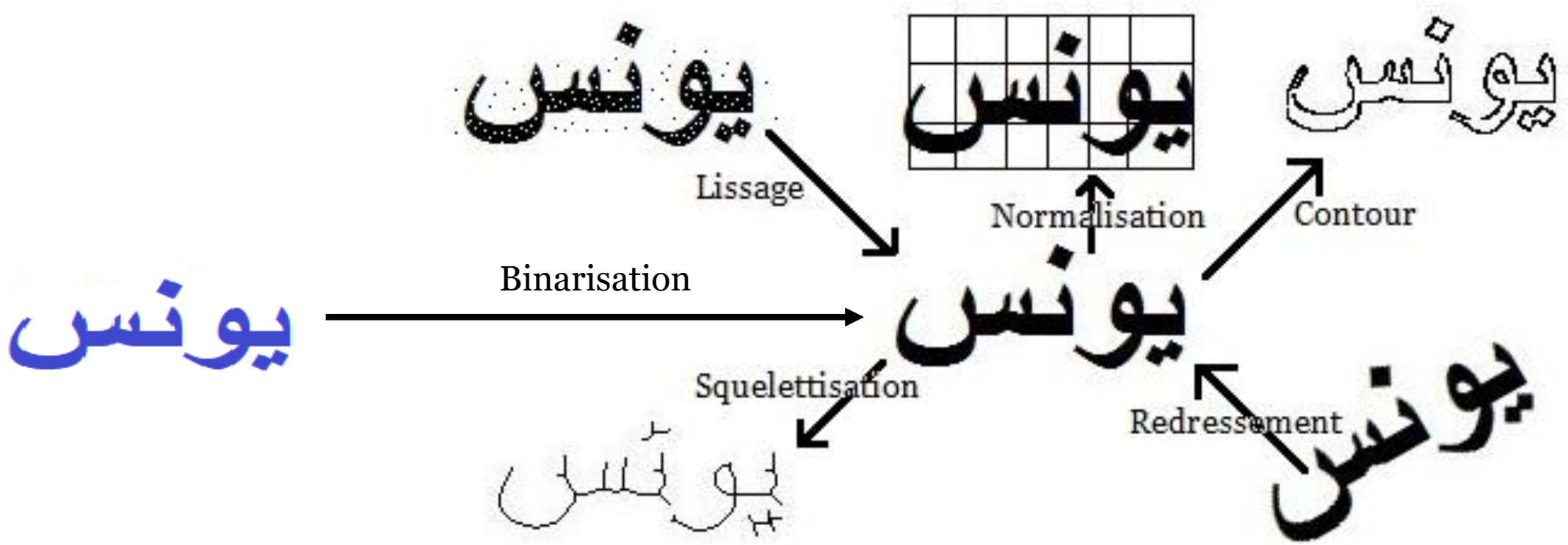
Processus général d'un OCR arabe

Prétraitements

- Cette phase de prétraitements consiste à préparer les données issues du capteur à la phase suivante.
- Il s'agit essentiellement de réduire le bruit (dû aux conditions d'acquisition ou à la qualité du document d'origine)

Processus général d'un OCR arabe

Prétraitements



Processus général d'un OCR arabe

Segmentation

- La segmentation est le processus de division du texte en plusieurs caractères individuels ou en pseudo-mots.
- C'est l'étape la plus cruciale pour un système OCR arabe.
- En général, dans les OCRs on distingue deux types d'approches:
  - L'approche globale
  - L'approche analytique

Processus général d'un OCR arabe

Segmentation: approche analytique

- Le mot est segmenté en caractères.
- La reconnaissance du mot complet sera obtenue par la combinaison des reconnaissances de ses caractères intermédiaires.
- Cette approche peut être utilisée pour la reconnaissance d'un vocabulaire étendu

Processus général d'un OCR arabe

Segmentation: approche globale

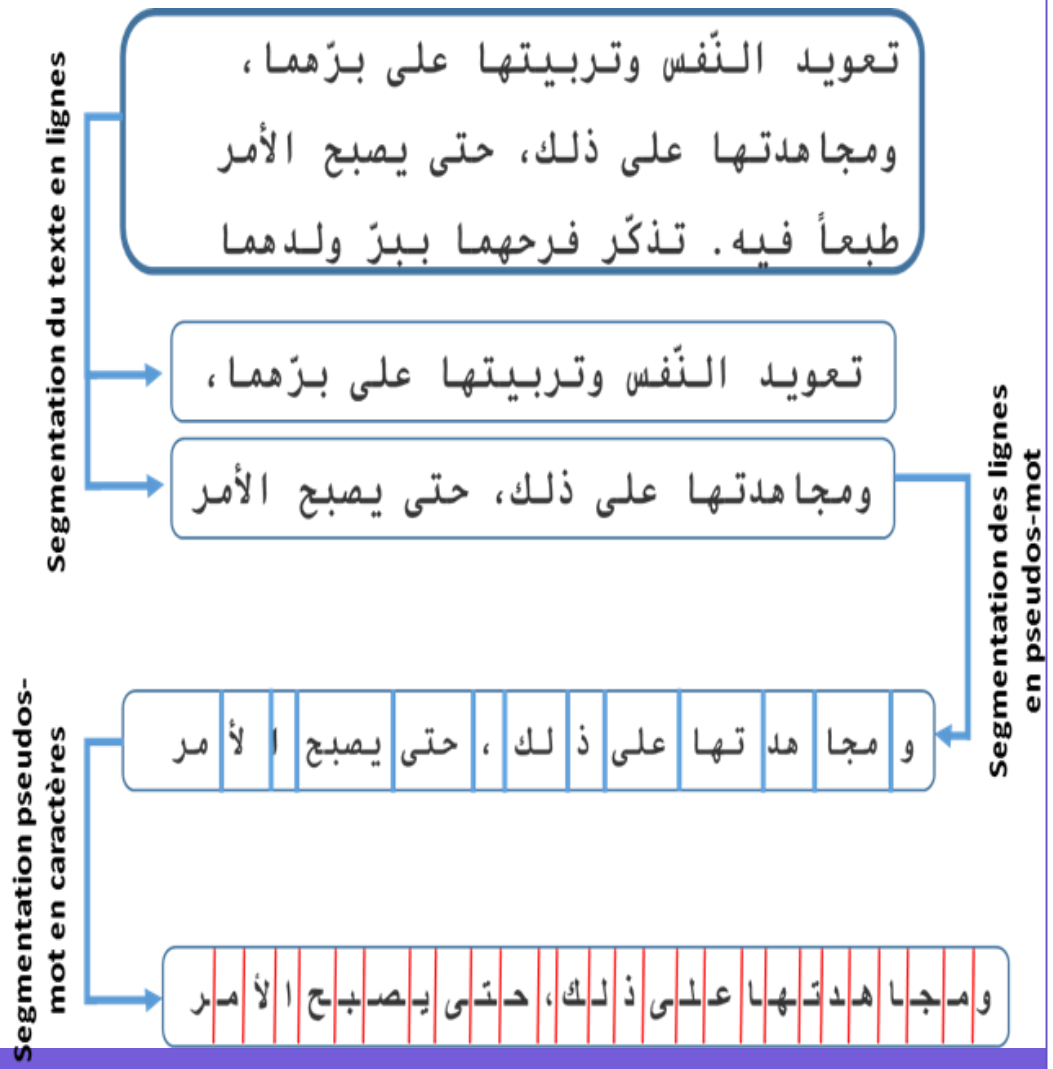
- Dans cette approche, on considère le mot comme une seule entité décrite indépendamment des caractères qui le constituent.
- Cette approche présente l'avantage de garder le caractère dans son contexte avoisinant.
- Cependant cette méthode est pénalisante par la taille mémoire, le temps de calcul et la complexité du traitement qui croient linéairement avec la taille du vocabulaire

Processus général d'un OCR arabe

Segmentation du texte arabe

Le processus de segmentation du texte arabe en caractères peut être divisé en trois niveaux :

1. Le texte est segmenté en plusieurs lignes.
2. Chaque ligne est segmentée en plusieurs pseudo-mots.
3. Chaque pseudo-mot est segmenté en caractères individuels

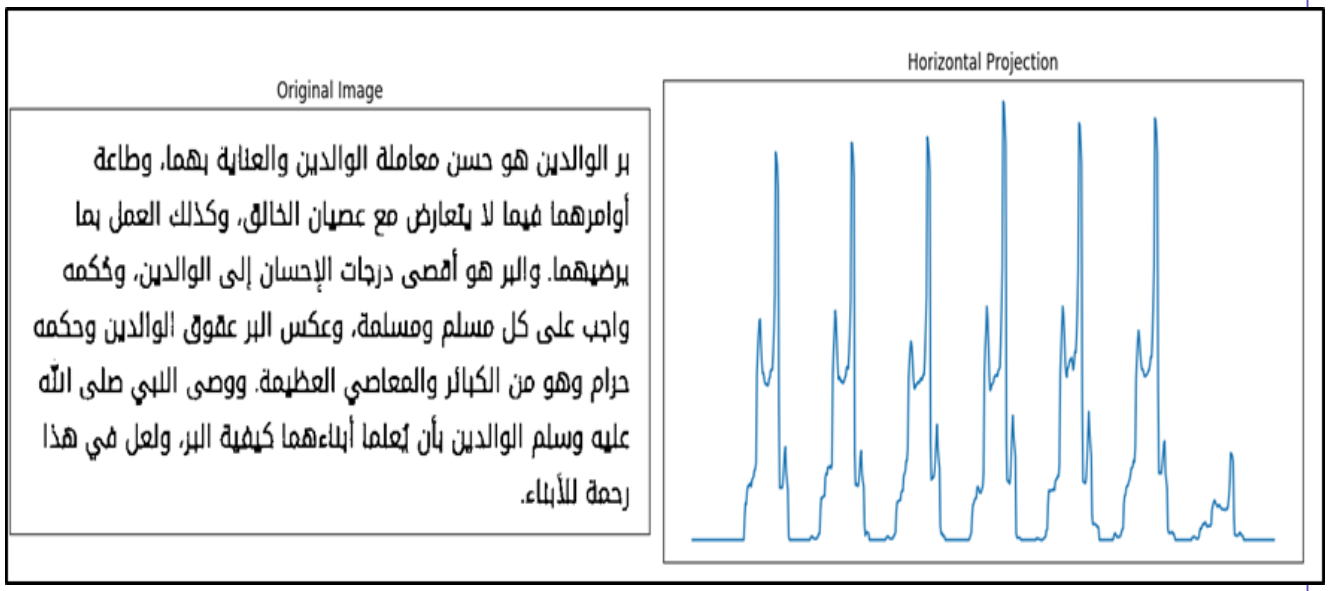


Processus général d'un OCR arabe

Segmentation du texte en lignes

Projection horizontale:

- Calcul du nombre de pixels noirs dans chaque rangée de l'image binaire.
- Chaque annulation de l'histogramme correspond à un vide dans l'image, et représente une zone de segmentation entre les lignes.
- Chaque ligne est séparée de l'image et sauvegardée dans une image binaire.

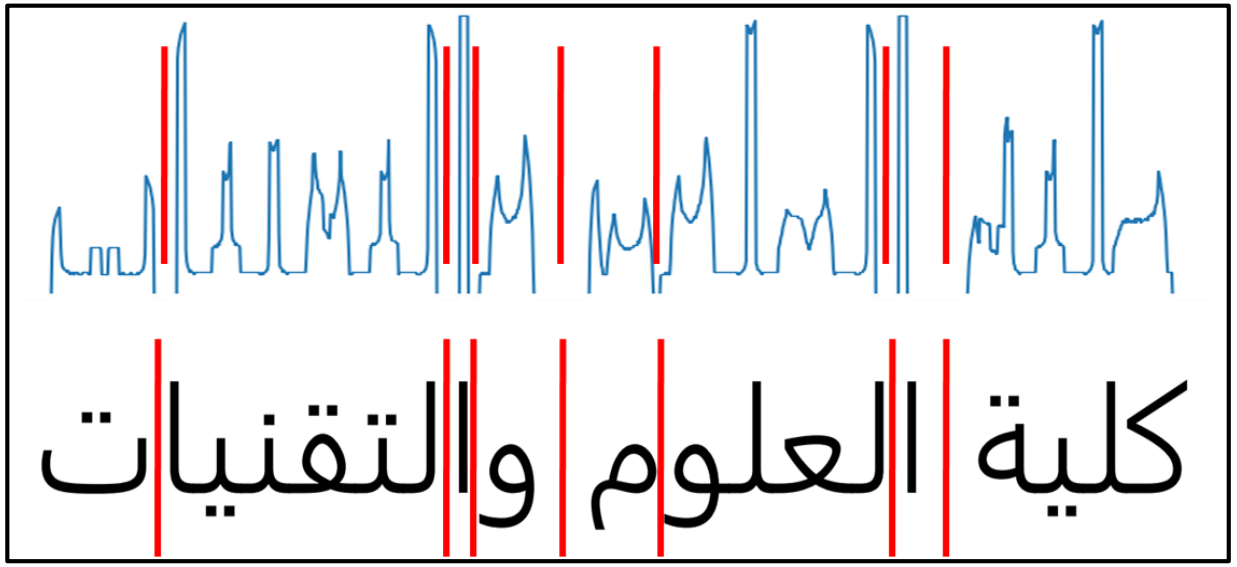


Processus général d'un OCR arabe

Segmentation de la ligne en pseudo-mots

Projection verticale:

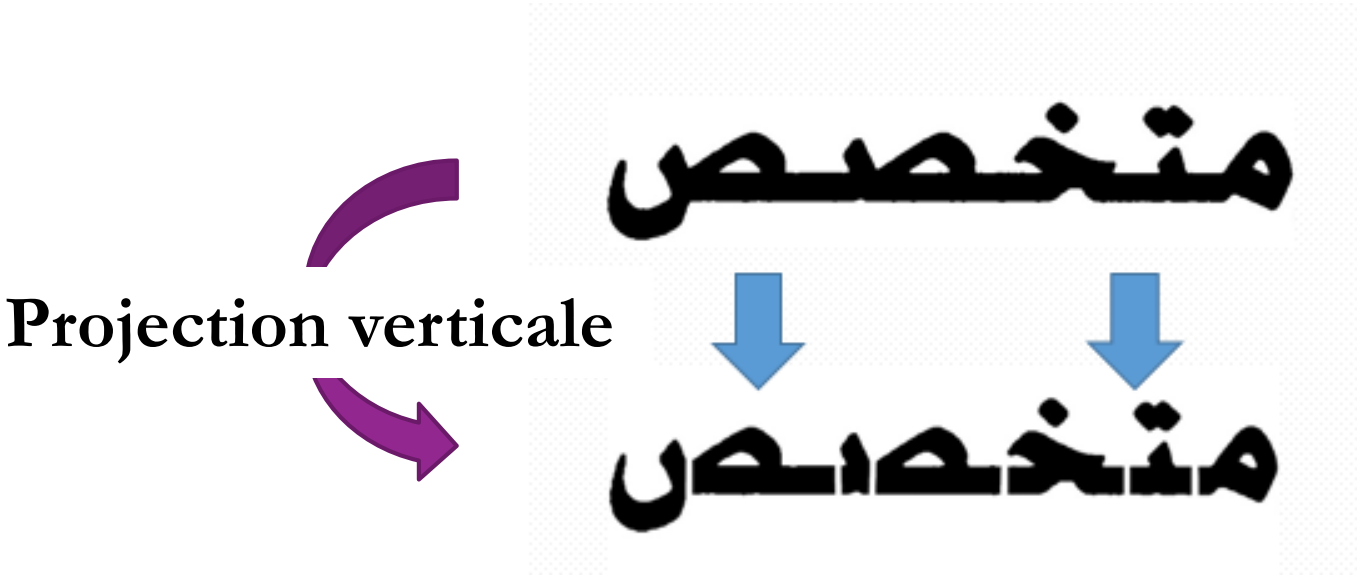
- Calcul du nombre de pixels noirs dans chaque colonne de l'image binaire.
- Chaque annulation de l'histogramme correspond à un espace vide, et représente une zone de segmentation entre les pseudo-mots
- Chaque pseudo-mot est séparé et sauvegardé dans une image binaire.



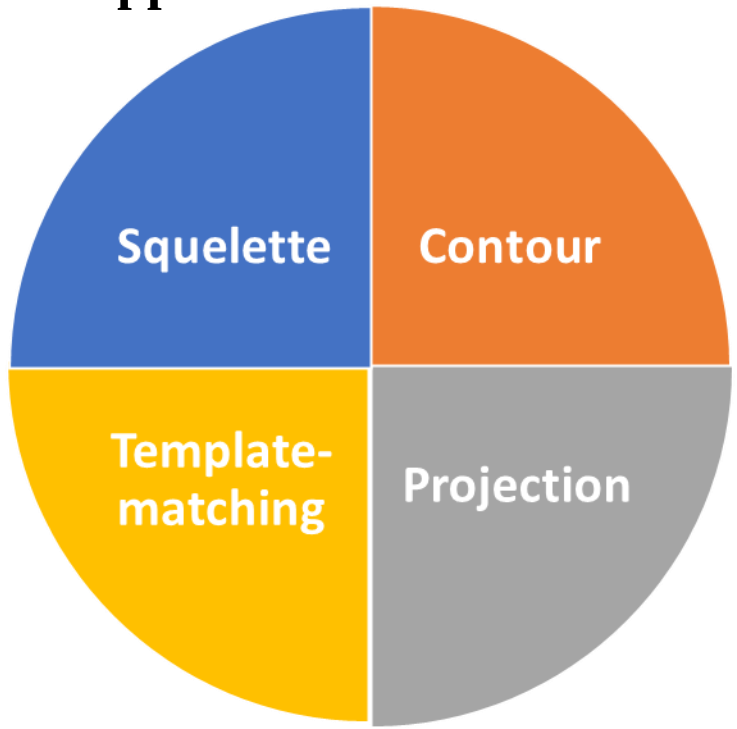
Processus général d'un OCR arabe

Segmentation des pseudo-mots en caractères

.



Plusieurs approches dans la littérature



Processus général d'un OCR arabe

Extraction des caractéristiques

- L'extraction des caractéristiques donne une description synthétique du caractère à reconnaître,
- Le but de cette phase est la sélection de l'information pertinente, discriminante et de dimension limitée pour l'étape de la classification toute en évitant le risque de perte des informations importantes et signifiantes

Processus général d'un OCR arabe

Extraction des caractéristiques

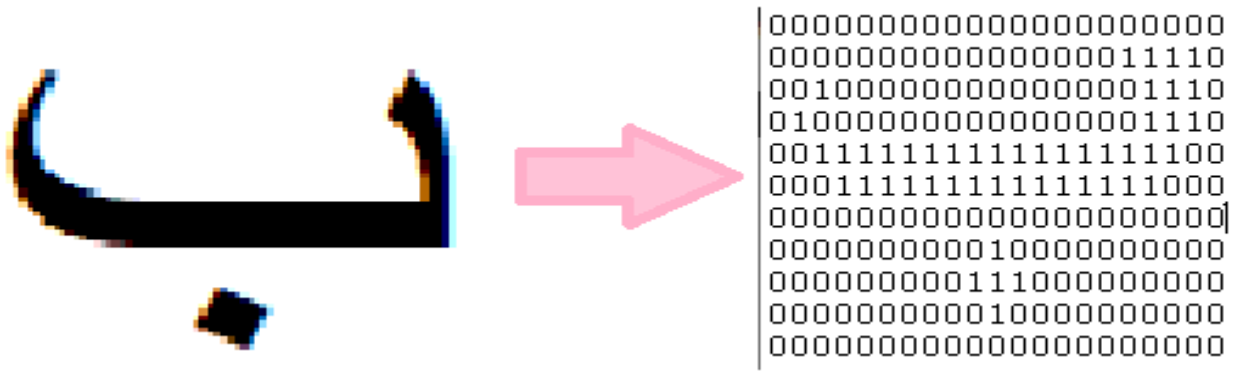
De nombreuses caractéristiques peuvent être extraites à partir des caractères

- Liste de pixels
- Structurelles
- Statistiques

Processus général d'un OCR arabe

Extraction des caractéristiques
Liste des pixels

La méthode la plus élémentaire consiste à construire un vecteur constitué d'autant de composantes qu'il y a de pixels dans l'image.



Processus général d'un OCR arabe

Extraction des caractéristiques
Liste des pixels

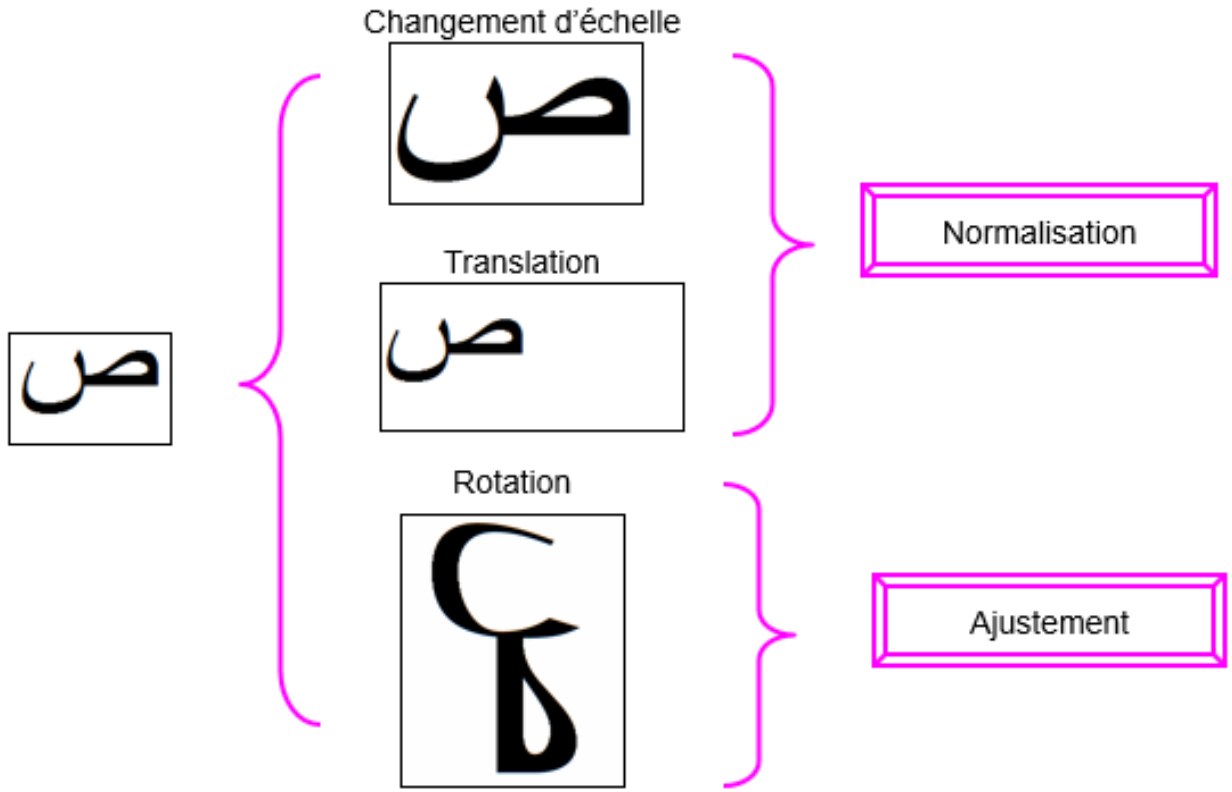
.

Points faibles des caractéristiques pixels:

- Vecteurs de taille trop importante
- Très sensible au bruit
- Variance par rapport à la transition au changement d' échelle et à la rotation

Processus général d'un OCR arabe

Extraction des caractéristiques
Liste des pixels



Processus général d'un OCR arabe

Extraction des caractéristiques

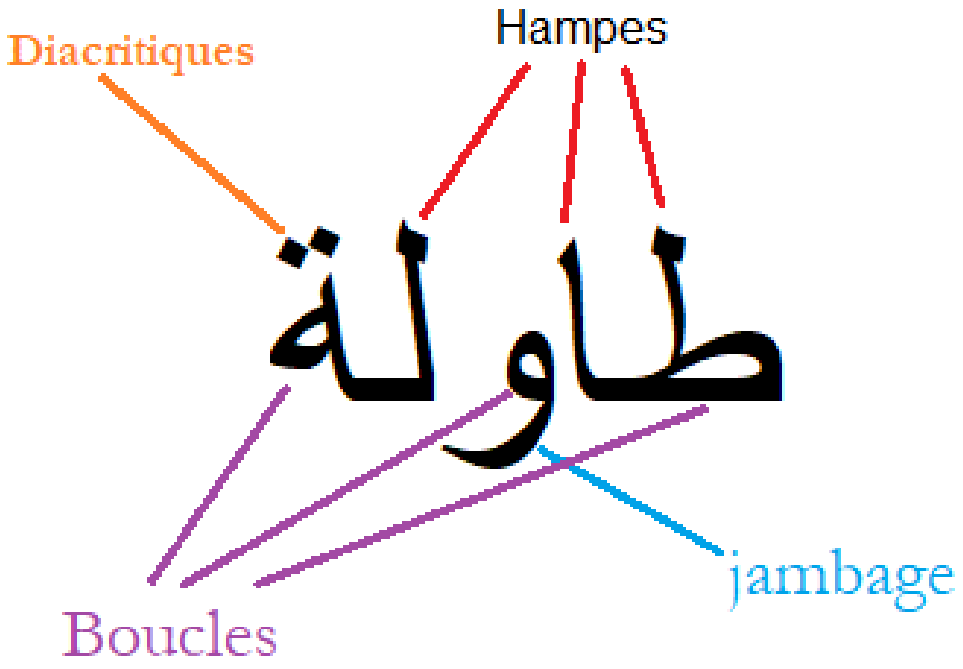
Les caractéristiques structurelles

- Elles décrivent un caractère en terme de sa topologie et sa géométrie
- Parmi ces caractéristiques on peut citer:
  - Le nombre de traits et leurs tailles.
  - Les points d'intersections.
  - Les boucles.
  - Le nombre de points diacritiques et leur position par rapport à la ligne de base.
  - Les jambages (descendants) et les hampes (ascendants)
  - La hauteur et la largeur du caractère.
  - ….

Processus général d'un OCR arabe

Extraction des caractéristiques

Les caractéristiques structurelles



Processus général d'un OCR arabe

Extraction des caractéristiques

Les caractéristiques statistiques ou globales

Elles utilisent des mesures calculées sur le caractère afin de le coder sous forme de vecteur

On cite deux exemples de méthodes définissants ces caractéristiques:

- Une méthode basée sur la transformation de Fourier permettant de décrire le contour externe du caractère,
- Une méthode basée sur les moments de Zernike permettant de décrire la forme du caractère en tant que région,

Ces méthodes sont invariantes à la translation, rotation et changement d'échelle

Processus général d'un OCR arabe

Classification

- L'idée principale de la classification est d'attribuer un caractère non connu

à une classe de caractères prédéfinie à partir des caractéristiques du caractère.

- La classification nécessite généralement deux phases : l'apprentissage des caractères et leur identification

Processus général d'un OCR arabe

Classification

Apprentissage = entraînement (training en anglais)

- L'apprentissage permet au système d'élaborer sa bibliothèque de caractéristiques des différents caractères de la base de données
- Cette bibliothèque de caractéristiques servira dans la phase suivante pour assigner à un caractère inconnu sa classe de caractères

Processus général d'un OCR arabe

Classification

Identification du caractère

- Dans cette étape, on cherche parmi les caractéristiques des caractères de la base de données, celles qui sont les plus proches des caractéristiques du caractère inconnu
- L'identification du caractère se fait à l'aide d'un classifieur qui détermine l'appartenance de la forme à une ou plusieurs classes de caractères

Processus général d'un OCR arabe

Classification

Identification du caractère

- Pas de meilleur classifieur. L'utilisation du classifieur depend de plusieurs facteurs comme: la base de données , le type des caracteristiques etc.
- k-Nearest Neighbour (k-NN) , Bayes Classifier, Neural Networks (NN), Hidden Markov Models (HMM), Support Vector Machines (SVM)…..

Processus général d'un OCR arabe

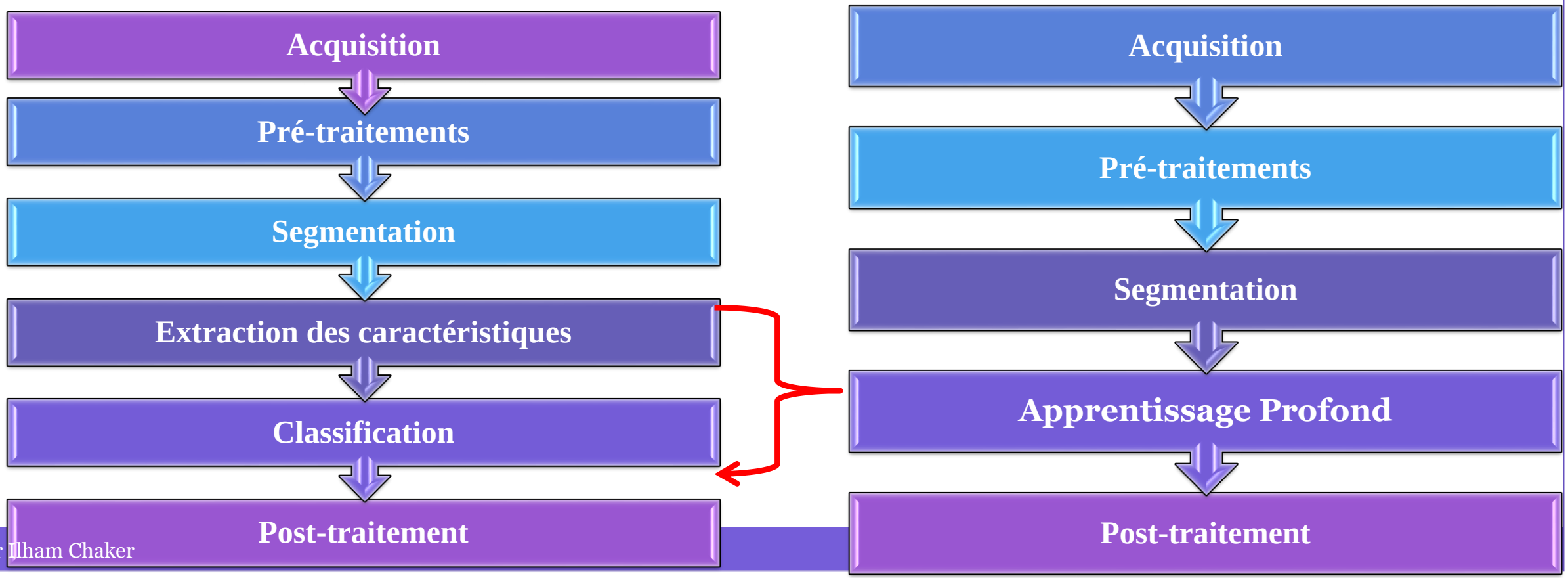
Les OCRs et l'apprentissage profond

- Problèmes du processus classique

  L'étape de l'extraction des caractéristiques est cruciale et critique dans les systèmes de reconnaissance. En effet, un mauvais choix des caractéristiques influence négativement et nettement les résultats même si on utilise un classifeur très performant.

Processus général d'un OCR arabe

Les OCRs et l'apprentissage profond



Processus général d'un OCR arabe

Les OCRs et l'apprentissage profond

- L'apprentissage profond ou deep learning est dérivé du machine learning (apprentissage automatique) où la machine est capable d'apprendre par elle-même.
- L'apprentissage profond est capable de résoudre des problèmes liés à la classification et la reconnaissance des caractères,
- Deep Learning s'appuie sur un réseau de neurones artificiels s'inspirant du cerveau humain. Ce réseau est composé de dizaines voire de centaines de « couches » de neurones, chacune recevant et interprétant les informations de la couche précédente

Processus général d'un OCR arabe

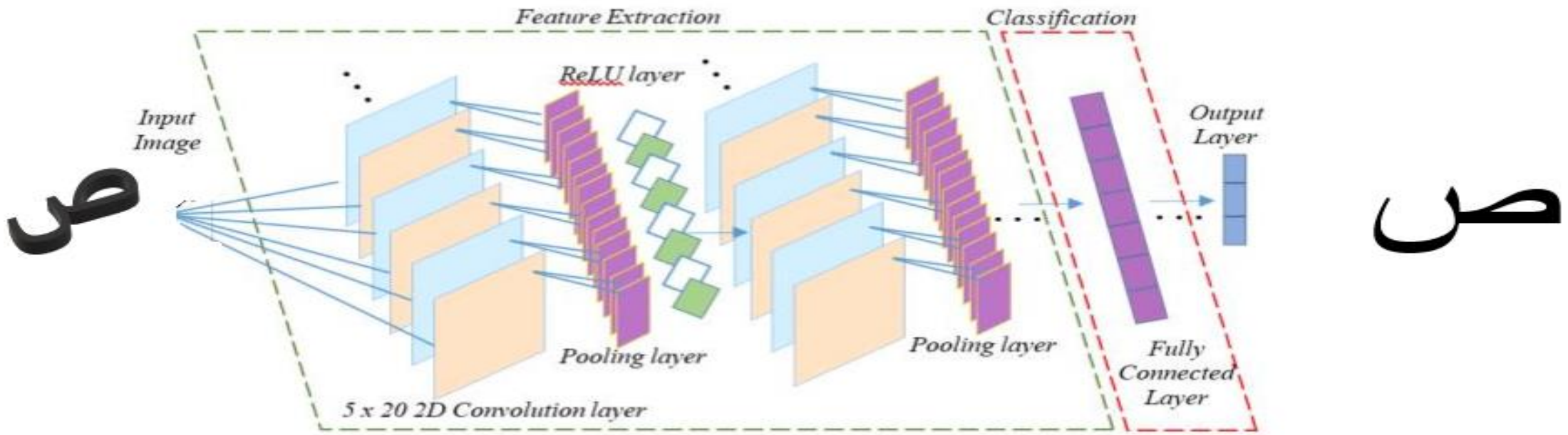
Les OCRs l'apprentissage profond

.

- CNN est l'un des types de réseau de neurones profonds les plus populaires,
- Il peut apprendre et extraire des caractéristiques des images.
- CNN peut reconnaître efficacement les caractères présents dans l'image.

Processus général d’un OCR arabe

Les OCRs l’apprentissage profond



Processus général d'un OCR arabe

Post-traitement

- Le post-traitement est effectué quand le processus de classification aboutit à la génération d'une liste de caractères ou de mots possibles.
- Le but principal est d'améliorer le taux de reconnaissance en faisant des corrections orthographiques ou morphologiques à l'aide de dictionnaires de digrammes, de contraintes de niveaux successifs : lexical, syntaxique ou sémantique.

Exemples d'OCRs arabes

- SAKHR

Sakhr Software
Arabic language technology

- ReadIris Arabic OCR

Readiris™

- NovoVerus

NovoVerus