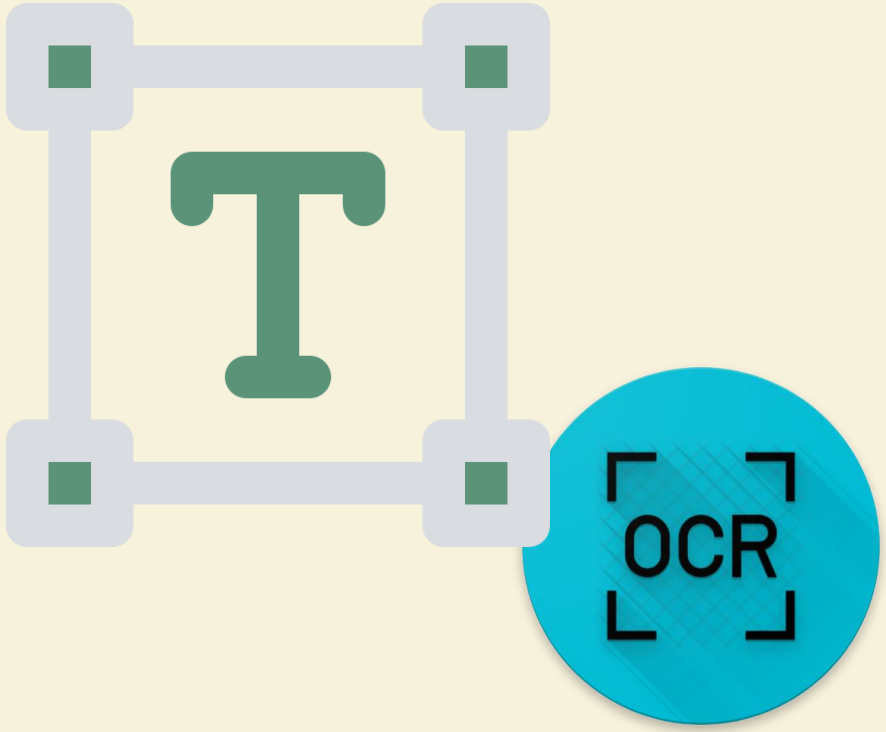
Détection de texte à partir d'images

OCR



Réalisé par : Radi Oumaima

PLAN

1. Introduction
2. C’est quoi OCR ?
3. L’objectif d’OCR
4. EasyOCR
5. L’approche proposée
6. Démarche suivie
7. Conclusion

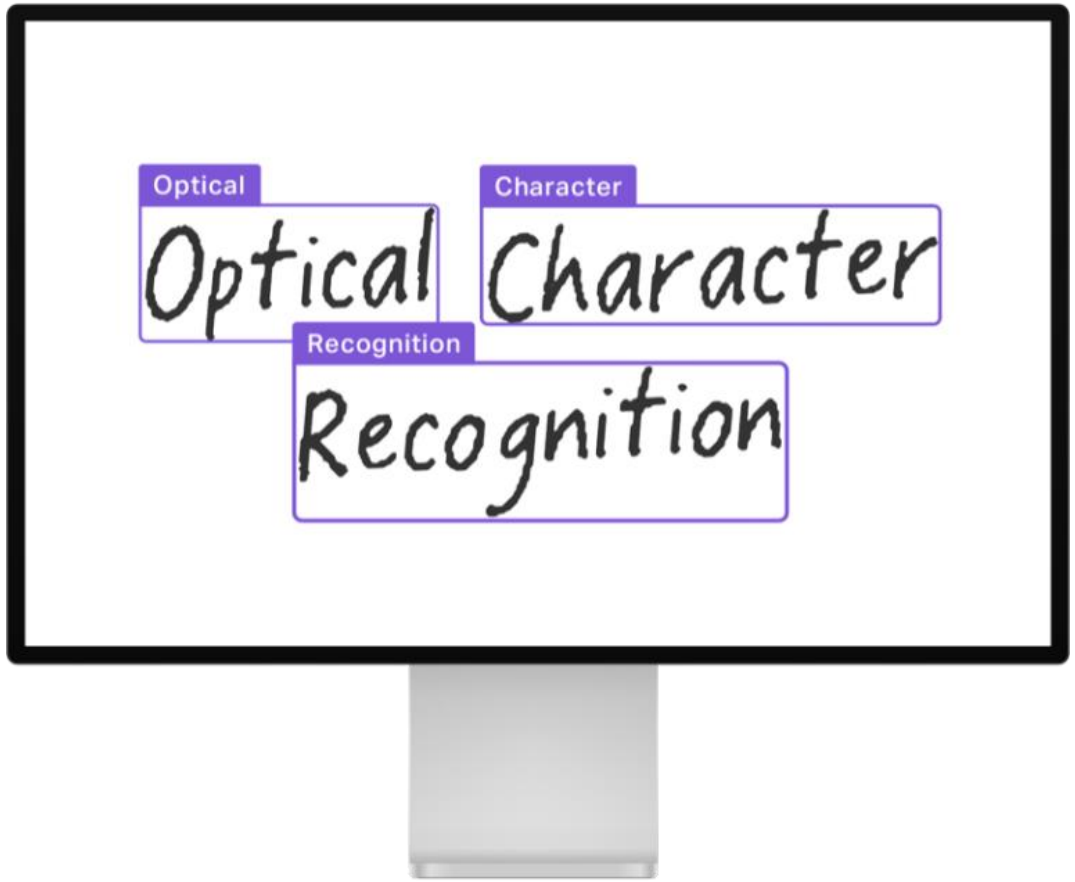
INTRODUCTION

La reconnaissance de caractère est un domaine actif de recherche pour la science informatique depuis la fin des années.
Au début, on pensait qu'il s'agissait d'un problème facile, mais il apparut qu'il s'agissait d'un sujet beaucoup plus intéressant. Il faudra encore de nombreuses décennies aux ordinateurs, s'ils y parviennent un jour, pour lire tous les documents avec la même précision que les êtres humains.

C’est quoi OCR ?

OCR: Optical Character Recognition

L'OCR est anciennement connue sous le nom de reconnaissance optique de caractères, qui est aujourd'hui révolutionnaire pour le monde numérique. L'OCR est en fait un processus complet dans lequel les images/documents qui sont présents dans un monde numérique sont traités et à partir du texte sont traités comme du texte modifiable normal.



L’objectif de l’OCR

L'OCR est une technologie qui vous permet de convertir différents types de documents, tels que des documents papier numérisés, des fichiers PDF ou des images capturées par un appareil photo numérique en données modifiables et consultables.

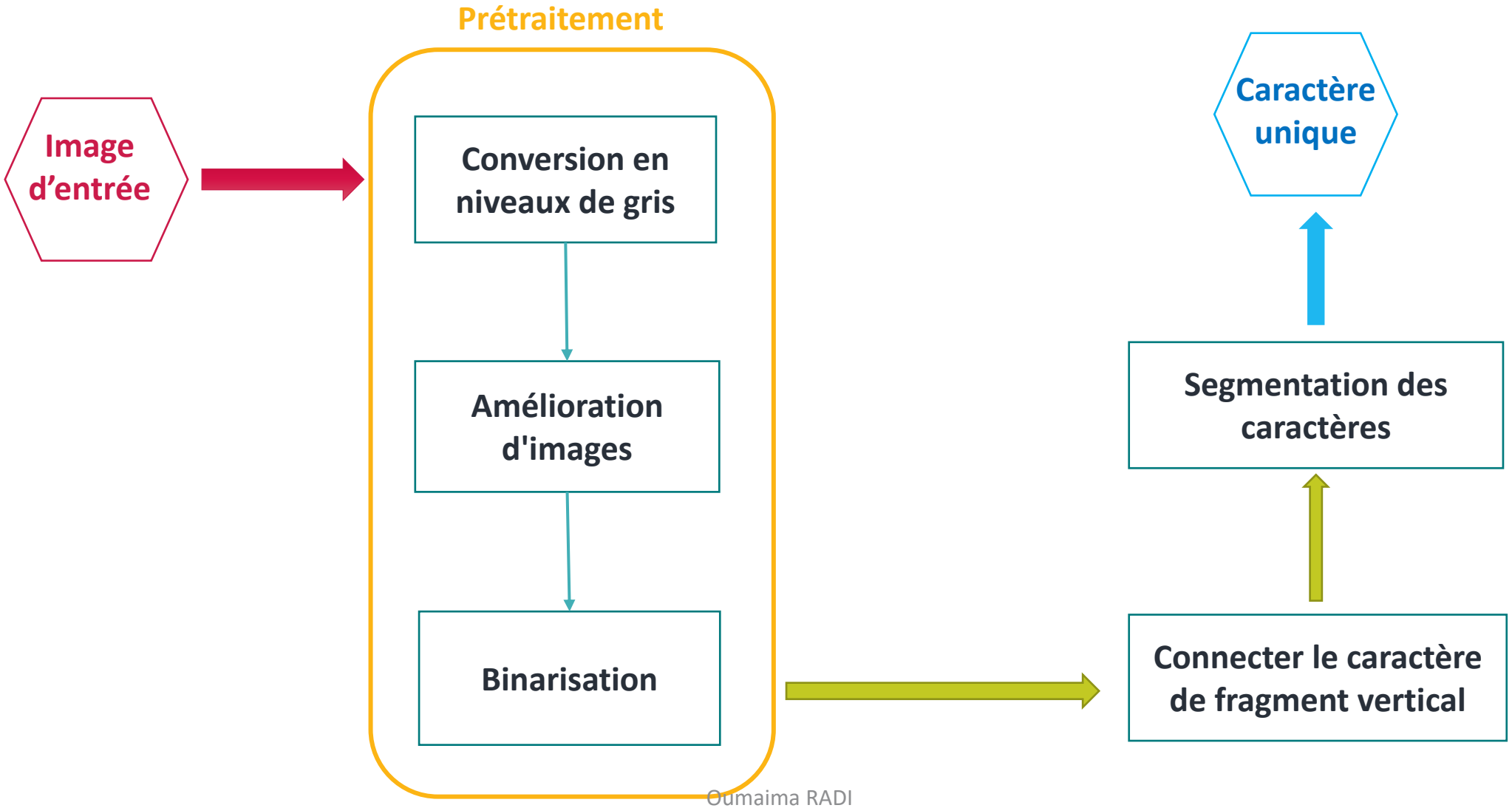
C’est quoi easyOCR ?

easyOCR : Qu'est-ce qu'EasyOCR ?

EasyOCR est en fait un package python qui contient PyTorch en tant que gestionnaire principal.

EasyOCR, comme tout autre OCR (tesseract de Google ou autre) détecte le texte à partir d'images, mais dans ma référence, en l'utilisant, j'ai trouvé que c'était le moyen le plus simple de détecter du texte à partir d'images également lorsque la bibliothèque d'apprentissage en profondeur haut de gamme (PyTorch) le prend en charge dans le backend, ce qui rend sa précision plus crédible.
EasyOCR prend en charge plus de 42 langues à des fins de détection.
EasyOCR est créé par la société nommée Jaided AI company.

L’approche proposée



L’approche proposée

1- Prétraitement ou Pré-analyse

- Le but est d'améliorer éventuellement la qualité de l'image.

  Ceci peut inclure le redressement d'images inclinées ou déformées, des corrections de contraste, binarisation de l’image ,le passage en mode bicolore (noir et blanc, ou plutôt papier et encre), la détection de contours.

2- Segmentation

- Segmentation en lignes et en caractères (ou Analyse de page) : vise à isoler dans l'image les lignes de texte et les caractères à l'intérieur des lignes.

  Cette phase peut aussi détecter le texte souligné, les cadres, les images.

  Donc la segmentation permet d’isoler dans l’image les différentes composantes (illustrations, blocs de texte, marges, etc.).

L’approche proposée

3- Reconnaissance

Reconnaissance proprement dite des caractères : après normalisation (échelle, inclinaison), une instance à reconnaitre est comparée à une bibliothèque de formes connues, et on retient pour l'étape suivante la forme la plus « proche » (ou les N formes les plus proches), selon une distance ou une vraisemblance (likelihood).

Les techniques de reconnaissance se classent en quelques grands types:

-Classification par Caractéristiques
-Méthodes métriques
-Méthodes statistiques

4- Post-traitement

En utilisant des méthodes linguistiques et contextuelles pour réduire le nombre d'erreurs de reconnaissance : systèmes à base de règles, ou méthodes statistiques basées sur des dictionnaires de mots, de syllabes, de N-grammes (séquences de caractères ou de mots). Dans les systèmes industriels, des techniques spécialisées pour certaines zones de texte (noms, adresses postales) peuvent utiliser des bases de données pour éliminer les solutions incorrectes.
Génération du format de sortie, avec la mise en page pour les meilleurs systèmes.

Démarche Suivie

- ✓ Installer les dépendances principales
- ✓ Importation de bibliothèques
- ✓ Lecture d'images
  - Par URL
  - Localement
- ✓ Extraire le texte de l'image
  - Texte en anglais
  - Texte arabe
- ✓ Résultats de dessin sur des images
  - Exemple 1
  - Exemple 2
  - Gérer plusieurs lignes de texte

CONCLUSION

Aujourd'hui l'OCR démocratisé, on pourrait alors se poser la question concernant la survivance de la technique OCR par rapport au nombre croissant de documents directement produits sous forme numérique. Un élément de réponse tient dans les techniques développées autour de l'OCR qui peuvent être utilisées à d'autres fins. Par ailleurs, bien que les documents actuels et futurs soient pratiquement toujours produits en numérique, toute la masse de documents produite jusqu'à nos jours contient suffisamment de challenge pour alimenter de nouvelles fonctionnalités d'OCR. Le futur des OCR permet de mieux exploiter numériquement les documents du passé.

Merci pour votre attention