Extraire un tableur depuis une image

Introduction générale

La reconnaissance optique des caractères (abrégé OCR, pour *optical caractère recognition*) consiste à convertir des fichiers images (documents scannés, fichiers PDF, photos numérisées, etc.) représentants des textes dactylographiques en fichier en format texte. Cette technique est centrale dans plusieurs applications, car elle facilite non seulement le stockage, mais aussi le traitement automatique des données qui auparavant était difficilement accessible.

Dans les dernières années, les algorithmes d'OCR se sont fortement améliorés en adoptant des techniques d'apprentissage profond, comme les réseaux de neurones récurrents. Plusieurs logiciels offrent d'excellentes performances dans une multitude de situations (images déformées, flous, mises en page multiples, langues et polices différentes) et offrent plusieurs fonctionnalités, comme : reconnaissance d'une police d'écriture, localisation spatiale du texte, répartition du texte en différents blocs, extraction de tableaux.

Jusqu'à récemment, le fonctionnement de systèmes OCR efficaces était mal connu car il était protégé par le secret industriel et réservé aux logiciels propriétaires. La publication en open source de systèmes performants (en particulier `Tesseract` en 2006) a quelque peu changé cette situation, avec des très bonnes réussites dans une grande variété de situations (polices différentes, supports multilangues etc.). Pourtant, les fonctionnalités offertes par les logiciels open source sont relativement basiques, permettant de transformer un document scanné au format `txt` ou `rtf`. En particulier, à notre connaissance, il n'existe pas un logiciel libre permettant d'extraire automatiquement des tableaux depuis des images.

Le but de ce projet est d'enrichir un logiciel libre de reconnaissance optique avec la fonctionnalité d'extraction d'un tableur en format `csv` depuis un fichier en format image (`png`, `tiff`, `jpeg` etc.) représentante ce tableau.

Objectifs

- Collecte et génération automatique d'un ensemble d'images de tableaux de différents formats.
- Installation et utilisation d'un algorithme d'OCR existant sur une grande variété d'images représentante des tableaux dactylographiques.
- Enrichir un algorithme d'OCR existant pour extraire des tableaux de format `csv`.
- Étudier les performances des modèles neuronaux d'un OCR en spécifiant leur apprentissage sur des ensembles d'images représentant des tableaux.

Testabilité

On donnera des instructions précises pour tester notre fonctionnalité sur des images en format `png`.

On fournira des exemples de tests et les résultats obtenus, en détaillant les exemples où notre programme est moins performant.

Calendrier

- Novembre : étude de l'existant.
  - étude des différents logiciels libres d'OCR, en particulier `Tesseract`;
  - prise en main d'une bibliothèque de *computer vision*, comme `OpenCV`;
  - collection et création d'un ensemble de donnes d'images de tableaux avec différentes mises en pages.
- Dicembre-Janvier :
  - expérimentation du logiciel d'OCR choisit et études des différents paramètres et techniques de prétraitement d'image pour améliorer les performances dans le cas spécifique d'images de tableaux.
- Fevrier-Mars:
  - mise en œuvre d'un algorithme qui générer un fichier `csv` à partir des informations extrait d'une image par le logiciel OCR chosit ;
  - étude de la performance de cet algorithme et analyses dès ses points faibles.
- Avril-Mai :
  - étude de la possibilité d'améliorer l'apprentissage de l'OCR grâce à des images spécifiques des tableaux.

Références

- OCR libre Tesseract, actuellement maintenu par google: https://github.com/tesseract-ocr/tessdoc
- Wrapper Python pour Tesseract: https://pypi.org/project/pytesseract/
- Librerie Python pour Computer Vision: https://github.com/opencv/opencv-python