CDD Ingénieur R&D - Computer Vision Machine Learning
Tours – CESR / LIFAT

Contexte

Ce poste s'inscrit dans le cadre du projet ANR TypoRef dirigé par Remi Jimenes (CESR Tours). Conçu dans une approche interdisciplinaire, le projet vise à inventorier, décrire et étudier les matériaux typographiques français de la Renaissance en développant une plateforme et des outils d'analyse et d'exploration de corpus de livres numérisés.

La numérisation des livres anciens a débuté au milieu des années 1990. Depuis cette date, des équipes de chercheurs alliant informaticiens et spécialistes du patrimoine écrit ont collaboré pour développer des outils dédiés. Longtemps, l'analyse et la reconnaissance d'images de documents ont été considérées comme des tâches complexes décomposées en 2 sous-tâches indépendantes : l'analyse de la mise en page et la reconnaissance du texte. L'analyse de la mise en page vise à détecter tous les composants de l'image tels que les blocs de texte, les tableaux, les images, les graphiques, ... La reconnaissance de texte est consacrée à la reconnaissance des caractères à l'intérieur des blocs de texte.

Dans le cadre d'une collaboration établie en 2003, le CESR et le LIFAT ont développé les logiciels Agora and Retro. Les développements ont été prolongés en 2011 et 2012 par deux bourses successives "Award in digital Humanities " financées par Google. Ces outils d'analyse et d'indexation de la mise en page adaptés aux images des livres imprimés de la Renaissance ont permis d'extraire et d'indexer automatiquement quelque 13 500 gravures, lettrines, bandeaux et fleurons à partir de 600 fac-similés des bibliothèques virtuelles humanistes du BVH.

Agora et Retro ont déjà plus de 15 ans et sont basés sur des technologies aujourd'hui dépassées au vu des développements de ces dernières années. Récemment, des architectures profondes ont amélioré les performances obtenues pour les deux sous-tâches. Sur la base de la très forte expérience accumulée au cours des vingt dernières années, le projet TypoReF entend permettre le développement de nouveaux outils d'analyse de la mise en page, d'identification et d'indexation des matériaux typographiques adaptés aux livres imprimés anciens, en utilisant les technologies les plus puissantes du moment, notamment celles basées sur l'apprentissage profond.

Missions à réaliser

L'objectif est la mise en place d'une plateforme web de labellisation de contenus. Cette labellisation concerne à la fois la collecte de métadonnées produites par un utilisateur expert et la production de métadonnées résultant d'une analyse des images/pixels. Cette double indexation (humain/machine) doit permettre de rechercher, comparer, regrouper des formes, et ainsi de mettre en évidence des liens difficilement perceptibles.

La plateforme nécessite de développer les fonctionnalités suivantes :

- Segmentation sémantique des images de documents à l'aide d'architectures d'apprentissage profond (pour remplacer l'ancienne Agora) : localisation et caractérisation d'éléments de contenu spécifiques à l'intérieur des pages numérisées d'un livre imprimé ancien (blocs de texte, décor gravé...) avec différents niveaux de détail (au sein d'un bloc de texte, on distinguera ainsi des lignes, des caractères) et caractérisation de ces éléments (lettrage, bandeau, marque, caractères, etc.) Chaque élément sera associé à un ensemble de métadonnées permettant de le décrire et de retrouver précisément son origine.
- Le clustering des éléments de contenu extraits (pour remplacer l'ancien Rétro) consiste à faire des comparaisons entre les éléments, principalement des ornements gravés (comme les lettrages ou les bandeaux), pour proposer des correspondances entre des formes similaires. La plateforme utilisera différents types d'algorithmes d'apprentissage automatique, notamment l'apprentissage dit " non supervisé ".
- Importation/exportation des données depuis/vers la base de données selon des standards tels que ALTO ou IIIF .

Compétences

- Développement logiciel et ingénierie, Python, HTML, CSS, JS
- Apprentissage automatique, vision par ordinateur
- Capacité à travailler en équipe, esprit curieux et rigoureux.

Comment postuler ?

- Envoyez votre CV et votre lettre de motivation
- Contact : Thierry Brouard, Jean-Yves Ramel --> prenom.nom AT univ-tours.fr
- Lieu : LIFAT et CESR - Tours France - http://lifat.univ-tours.fr - https://cesr.univ-tours.fr/
- Poste : (environ) contrat de 12 mois - durée et salaire selon l'expérience. Temps plein
- Dates du contrat : 1er septembre 2023 - 31 août 2024

Engineer R&D - Computer Vision - Machine Learning
Tours – CESR / LIFAT

Context

This position is part of the ANR TypoRef project directed by Remi Jimenes (CESR Tours). Conceived in an interdisciplinary approach, the project aims at inventorying, describing and studying French typographic materials of the Renaissance by developing tools for the analysis and exploration of digitized book corpora and by setting up a dedicated digital platform.

The digitization of old books began in the mid-1990s. Since then, teams of researchers combining computer scientists and specialists in written heritage have collaborated to develop tools dedicated to the exploitation of these digital corpora. For many years, document image analysis and recognition were considered complex tasks decomposed into two independent sub-tasks: layout analysis and text recognition. Layout analysis aims at detecting all image component such as text blocks, tables, images, graphics, signatures... Text recognition is devoted to the recognition of characters inside text blocks (paragraphs).

In the framework of a collaboration established in 2003, the CESR and the LIFAT have developed the Agora and Retro software. The developments were extended in 2011 and 2012 by two successive "Digital Humanities" grants funded by Google. These layout analysis and indexing tools adapted to the images of printed books of the Renaissance have made it possible to automatically extract and index some 13,500 engravings, lettering, headbands and fleurons from 600 facsimiles of the BVH Humanist Virtual Libraries.

Agora and Retro are already more than 15 years old and based on technologies that are now outdated given the developments of recent years. Recently, deep architectures have improved the state-of-the-art performance for both sub-tasks. Based on the very strong experience accumulated during the last twenty years, the TypoReF project intends to allow the development of new tools for the analysis of layout, identification and indexing of typographic materials adapted to the ancient, printed book, using the most powerful technologies of the moment, in particular those based on deep learning.

Missions to achieve

The goal is **the implementation of a web-based content labeling platform**. This labeling concerns both the collection of metadata produced by an expert user and the production of metadata resulting from an analysis of the pixels, their layout, and the relationships between the extracted elements of content. This double indexing (human/machine) must allow to search, compare, and group shapes, and thus to highlight links that would otherwise be difficult to perceive.

The platform requires developing the following functionalities:

- **Semantic segmentation of document images using deep learning architectures** (to replace old Agora). It consists of the localization and characterization of specific elements of content inside the digitized pages of an old printed book (text blocks, engraved decoration, etc.) with different levels of detail (within a text block, we will thus distinguish lines and characters). Characterization of these elements is also needed: lettering, bands, marks, characters, etc. A set of associated metadata allows one to characterize and retrieve precisely the location of each component.
- **Clustering of the extracted elements of content** (to replace old Retro) consists in making comparisons between elements, mainly engraved ornaments (such as lettering or headbands), to propose matches between similar forms. The platform will use different types of **machine learning algorithms, including the so-called "unsupervised" learning.**
- **import/export data from/to the internal database according to standards such as ALTO or IIIF** .

Skills

- Software programing and software engineering in Python and web (HTML, CSS, and JS)
- Machine Learning, Computer vision
- Ability to work in a team, curious and rigorous spirit

How to apply?

Send CV and motivation letter

- **Contact**: Thierry Brouard, Jean-Yves Ramel --> prenom.nom AT univ-tours.fr
- **Location**: LIFAT and CESR –Tours France - http://lifat.univ-tours.fr - https://cesr.univ-tours.fr/
- **Position**: (around) 12-month contract - duration and salary according to experience.
- **Dates of the contract**: Sept 1st 2023 – August 31st 2024