ORGANISME PROPOSANT LE SUJET : Laboratoires GeF et Cédric (Cnam)

MAÎTRES DE STAGE : AUDEBERT Nicolas (Cedric), DURAND Frédéric (GeF), FOLLIN Jean-Michel (GeF), SIMONETTO Elisabeth (GeF).

DURÉE : 5 mois minimum

LIEU : Laboratoire GeF au Mans (ESGT/Cnam), séjours à prévoir au laboratoire Cédric à Paris (Cnam)

Détection par Deep Learning des numéros de parcelles issus des plans scannés du Cadastre Napoléonien

PROBLÉMATIQUE

Les plans cadastraux anciens représentent une mine d'informations sur un territoire, par exemple pour l'analyse de l'évolution du parcellaire au cours du temps en lien avec les politiques d'aménagement. Ces plans correspondent à la numérisation de feuilles au format papier de qualités très inégales selon leur année de création et les conditions de leur conservation. Une avancée considérable pour l'analyse fine de l'évolution du territoire par les historiens, géographes, urbanistes, aménagistes et sociologues, viendrait sans conteste de la construction d'une base de données multi-dates du cadastre « ancien » au cadastre « actuel ». A ce jour, il n'existe pas à notre connaissance d'outils permettant l'analyse automatique du contenu de ces planches en vue de les intégrer dans un SIG (Système d'Informations Géographiques).

Aussi, le laboratoire GeF mène depuis 2016 des travaux de recherche sur le développement d'une chaîne semi-automatique d'analyse des images du cadastre ancien appelée « GeFVectoMoCad » (pour Géoréférencement, Vectorisation et Mosaïquage du Cadastre) à partir d'outils libres, dont le langage Python. Cette chaîne comporte plusieurs étapes : 1) la vectorisation, 2) le géoréférencement et 3) le mosaïquage des planches cadastrales anciennes.

L'étape de vectorisation est cruciale pour la réussite du processus et repose actuellement sur l'emploi d'algorithmes classiques de détection de segments qui constituent ici les limites de parcelles, comme le « Line Segment Detector » et la Transformée de Hough Probabiliste. Ces segments sont ensuite convertis en polygones fermés formant les parcelles. Bien qu'efficace et intégrant des post-traitements adaptés, cette approche n'est pas exempte d'erreurs notamment avec de la sur-segmentation et de la sous-segmentation [Follin et al., 2021]. Elle requiert donc des corrections manuelles a posteriori.

L'amélioration de ces résultats peut notamment passer par la détection des numéros de parcelles. En effet chaque parcelle est associée à un numéro unique écrit à la main. Les numéros de parcelles extraits pourront être confrontés aux polygones des parcelles pour détecter les incohérences, et donc les éventuelles erreurs de segmentation, et les corriger.

Une approche par apprentissage profond ou « Deep Learning » (DL) semble adaptée à la détection des caractères manuscrits (*digits*). La reconnaissance de caractères écrits à la main est un des premiers cas traités par les réseaux de neurones convolutifs. [LeCun et al., 1989] a ainsi proposé une approche

de type DL pour la reconnaissance des chiffres pour le service postal des États-Unis. Ce sujet étant plutôt bien maîtrisé, les problématiques du stage sont :

1. La production d'un jeu de données suffisamment volumineux pour entraîner des modèles profonds de reconnaissance de caractères. Cela pourra notamment passer par la génération d'images synthétiques et l'usage de techniques d'adaptation de domaine pour rendre ces images similaires aux planches de cadastres numérisées.
2. Le choix et la mise en œuvre d''un ou plusieurs réseaux profonds de reconnaissance de chiffres manuscrits. On pourra notamment étudier des détecteurs génériques (YOLO [Redmon et al., 2015], Mask-RCNN [He et al. 2017]) mais aussi des architectures spécifiques à la reconnaissance de caractères (CharGrid-OCR [Reisswig et al., 2019], Calamari [Wick et al., 2018]).
3. Une fois les chiffres détectés, les numéros devront ensuite être reconstruits. Le parcellaire sera alors représentée sous forme de graphe avec pour sommets les parcelles et pour arêtes les relations d'adjacence. Chaque sommet sera associé éventuellement à un (ou plusieurs) numéro de parcelle. On pourra s'appuyer sur ce graphe pour détecter des incohérences (absence ou multiplicité de numéro) et réfléchir aux corrections à apporter.

CONTEXTE DE L'ÉTUDE

Des travaux récents en traitement d'images ont prouvé l'intérêt de méthodes basées sur le DL pour l'extraction de textes [Laumer et al. 2020] et de nombres manuscrits [Kusetogullari et al., 2020] sur des documents anciens.

Plusieurs jeux de données existent. MNIST (Modified National Institute of Standards and Technology) contenant des chiffres manuscrits et USPS (United-States Postal Service) qui regroupe des chiffres mais aussi des lettres et des mots se présentent sous forme d'images respectivement en noir et blanc et en niveaux de gris. ARDIS (ARkiv Digital Sweden) correspond à des chiffres écrits à la main issus de registres religieux suédois et DIDA, extension d'ARDIS, contient des chiffres manuscrits issus de documents historiques suédois. Ces deux derniers jeux de données sont sous forme d'images en couleurs et se rapprochent le plus de nos données.

Un stage de fin d'études a été mené au laboratoire GeF en 2021 et a permis de réaliser de premières expérimentations. Elles ont consisté en la génération d'un jeu de données reprenant des chiffres de ARDIS et la mise en œuvre d'un réseau adoptant une architecture Faster-RCNN. Les résultats obtenus sur nos données, bien qu'encourageants, sont perfectibles.

OBJECTIFS VISÉS

- Améliorer l'entraînement par la génération d'un jeu de données basé notamment sur DIDA (publié en juin 2021) et par la réduction de l'écart (caractéristiques telles que la colométrie, le bruit et les objets représentés) entre nos données (planches scannées du cadastre) et le jeu de données d'entraînement du modèle.
- Choisir et paramétrer une architecture plus efficace que Faster-RCNN pour la détection des digits manuscrits : YOLO et, si le temps le permet, CharGrid-OCR pourront être étudiés.
- Mettre en œuvre des méthodes (1) de reconstruction des numéros complets des parcelles à partir des chiffres et (2) de génération d'un graphe ayant pour sommets les numéros de parcelles et pour arêtes les relations de voisinage entre parcelles a priori connexes.

PROFIL

Nous recherchons pour ce stage un·e candidat·e de niveau M2 ou dernière année d'école d'ingénieur avec une formation en géomatique et/ou en apprentissage automatique. Le ou la candidat·e idéal·e a

une appétence pour la recherche et des bases en apprentissage profond. Sans être indispensable, un intérêt pour les données géographiques est un point positif pour ce stage. Une connaissance de la programmation avec Python est nécessaire. Une première expérience avec une bibliothèque d'apprentissage profond (TensorFlow ou PyTorch) est un plus.

ORGANISATION DU STAGE

Le stage se déroulera au laboratoire GeF situé au Mans, avec des visites à prévoir au laboratoire CEDRIC à Paris.

Le ou la stagiaire disposera d'un poste de travail avec tous les logiciels requis ainsi que de l'accès à un serveur de calcul GPU.

Indemnités de stage : environ 1/3 du SMIC

DOCUMENTS A FOURNIR POUR CANDIDATURE

- CV et lettre de motivation à envoyer avant le 31 octobre à :
- elisabeth.simonetto@lecnam.net ; nicolas.audebert@cnam.fr ; jean-michel.follin@lecnam.net ; frederic.durand@lecnam.net

RÉFÉRENCES

Follin, Jean-Michel, Élisabeth Simonetto, et Anthony Chalais. 2021. « Détection automatique des parcelles sur les plans napoléoniens : comparaison de deux méthodes ». *Humanités numériques*, n° 3 (mai). https://doi.org/10.4000/revuehn.1779.

He, Kaiming, Gkioxari, Georgia, Dollár, Piotr, et Ross Girshick. 2017. « Mask RCNN ». IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2980-2988. https://doi.org/10.1109/ICCV.2017.322.

Kusetogullari, Huseyin, Yavariabdi, Amir, Hall, Johan, et Niklas Lavesson. 2020. « DIGITNET: A Deep Handwritten Digit Detection and Recognition Method Using a New Historical Handwritten Digit Dataset », Big Data Research 23 (2021): 100182. https://doi.org/10.1016/j.bdr.2020.100182.

Laumer, Daniel, Hasret Gümgümcü, Magnus Heitzler, et Lorenz Hurni. 2020. « A Semi-Automatic Label Digitization Workflow for the Siegfried Map ». In *Automatic Vectorisation of Historical Maps: International Workshop Organized by the ICA Commission on Cartographic Heritage into the Digital. Budapest – 13 March, 2020*, 57-64. Department of Cartography and Geoinformatics ELTE. https://doi.org/10.21862/avhm2020.07.

LeCun, Yann, Léon Bottou, Yoshua Bengio, et Patrick Haffner. 1998. « Gradient-based learning applied to document recognition ». *Proceedings of the IEEE* 86 (11): 2278-2324. https://doi.org/10.1109/5.726791.

Redmon, Joseph, Santosh Divvala, Ross Girshick, et Ali Farhadi. 2016. « You Only Look Once: Unified, Real-Time Object Detection ». In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 779-88. https://doi.org/10.1109/CVPR.2016.91.

Reisswig, Christian, Anoop R. Katti, Marco Spinaci, et Johannes Höhne. 2019. « Chargrid-OCR: End-to-End Trainable Optical Character Recognition through Semantic Segmentation and Object Detection ». In Workshop on Document Intelligence at NeurIPS 2019. https://openreview.net/forum?id=SkxhzT5qIB.

Wick, Christoph, Christian Reul, et Frank Puppe. 2020. « Calamari − A High-Performance Tensorflow-based Deep Learning Package for Optical Character Recognition ». *Digital Humanities Quarterly* 014 (2).