Vers une segmentation robuste du tracé dans les manuscrits anciens

Colin Brisson, EPHE-PSL

La segmentation du tracé constitue une étape cruciale dans le processus de transcription automatique : toute ligne non reconnue entraîne une lacune dans l'édition numérique, tandis qu'une mauvaise segmentation peut entraîner des erreurs de transcription. Dans le cadre du projet Read_Chinese, nous avons récemment entrepris la transcription du fonds Pelliot chinois de la Bibliothèque nationale de France (BnF), une collection de près de 7000 manuscrits médiévaux chinois acquis par le sinologue Paul Pelliot en 1908 à Dunhuang dans le nord-est de la Chine. Nous disposons déjà de modèles de reconnaissance robustes entraînés grâce au moteur d'OCR kraken [1]. Cependant, la qualité de la segmentation ne permet pas encore d'obtenir des transcriptions satisfaisantes. Il s'agit d'un problème récurrent également rencontré par d'autres projets et qui nous a conduits à rechercher des alternatives.

La prédiction de polygones délimitant les contours du tracé présente plusieurs avantages par rapport à la méthode de prédiction de la base d'écriture (baseline) employée par kraken [2]. Cette approche, adoptée par les logiciels dhSegment [3] et Doc-UFCN [4], consiste à utiliser un modèle de segmentation, généralement basé sur une architecture de type U-Net, pour produire une fonction attribuant à chaque pixel une probabilité d'appartenance à une ligne d'écriture. Les polygones sont ensuite vectorisés en utilisant la méthode des composants connexes. L'élimination de l'algorithme de seam carving utilisé pour la polygonisation des bases d'écriture permet d'assurer l'intégrité des polygones et réduit considérablement le temps nécessaire pour la segmentation. De plus, l'architecture U-Net est particulièrement efficace pour la segmentation d'objets de petite taille. Cependant, les implémentations actuelles présentent certaines limitations, notamment une vulnérabilité au phénomène de fusion des lignes (line merging), où les lignes sont confondues lorsque les polygones se chevauchent. De plus, les fonctions de coût actuellement utilisées sont sensibles au déséquilibre entre les classes, ce qui conduit le modèle à se concentrer sur les classes prédominantes lors de l'apprentissage.

Durant ma présentation, je souhaite partager le travail entrepris ces derniers mois pour améliorer la méthode de prédiction des polygones et parvenir à obtenir une segmentation fiable et précise du tracé dans les manuscrits anciens. Je montrerai comment l'utilisation d'une fonction de coût de type Dice généralisée (Generalized Dice Loss) permet au modèle d'apprendre à segmenter avec une même précision différents types de lignes. Je présenterai également comment il est possible de segmenter des lignes d'écritures qui se chevauchent. J'illustrerai mon propos à l'aide d'exemples issus du fonds Pelliot chinois de la BnF, ainsi que de manuscrits composés dans d'autres écritures que le chinois.

Bibliographie :

[1] Brisson, C., Constant, F., & Bui, M. (2023). Chinese Historical documents Automatic Transcription (CHAT) models (Version v0) [Modèle]. Zenodo. https://doi.org/10.5281/zenodo.8383732

[2] Kiessling, B. (2020). A Modular Region and Text Line Layout Analysis System. in : 17th International Conference on Frontiers in Handwriting Recognition, ICFHR 2020, Dortmund, Allemagne, 8-10 September 2020 (pp. 313–318). IEEE. https://doi.org/10.1109/ICFHR2020.2020.00064

[3] Oliveira, S. A., Seguin, B. & Kaplan, F. (2018). dhsegment: A generic deep-learning approach for document segmentation. in : 2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR). IEEE (pp. 7–12).

[4] Boillet, M., Kermorvant, C. & Paquet, T. (2022). Robust text line detection in historical documents: learning and evaluation methods. International Journal on Document Analysis and Recognition (IJDAR), 25, 95–114. https://doi.org/10.1007/s10032-022-00395-7