DOI 10.18699/SBB-2022-650

Deep learning for the development of an OCR for old Tibetan books

Brodt K.[1, 2]*, Rinchinov O.[3], Bazarov A.[3], Okunev A.[2, 4]

[1] *Université de Montréal, Montréal, Canada*

[2] *Novosibirsk State University, Novosibirsk, Russia*

[3] *Institute for Mongolian, Buddhist and Tibetan Studies, SB RAS, Ulan-Ude, Russia*

[4] *MTS AI LLC, Novosibirsk, Russia*

* *cyrill.brodt@gmail.com*

Key words: deep learning, Tibetan xylographs, optical character recognition (OCR)

Motivation and Goals: The funds of the Center of Oriental Manuscripts and Xylographs of the Institute for Mongolian, Buddhist and Tibetan Studies of the Siberian Branch of the Russian Academy of Sciences have the richest Buddhist heritage collections of manuscripts and woodblock printings, numbering about 100 thousand books in Tibetan and Mongolian. The manual processing of these writings and their translation into a machine-readable form is tedious and time-consuming [1]. Automation of optical character recognition or OCR allow us to save cultural treasures for the future generation.

Methods and Algorithms: We use computer vision and deep learning advances to apply the OCR for old Tibetan books. Experts in Tibetan studies at the Institute for Mongolian, Buddhist and Tibetan Studies of the Siberian Branch of the Russian Academy of Sciences annotated a dataset of scanned rare woodblock printings, namely Jone (Chone) Kangyur of early 18th century. These works have been chosen because they represent the woodcutting and printing quality typical for most Tibetan books. The dataset includes 500 images of book pages. Each image has text annotations.

We split the dataset into 420 training images and 30 testing images. The RGB image width is 2048 pixels and the average height is 650 pixels. Each image contains 8 text lines on average. Each line is framed with bounding box, and the corresponding text annotation is transliterated in Latin characters. The text length per line is 300 characters on average. First, we detect rectangular bounding boxes with text in image. After we crop out the boxes from image and extract the text using OCR model.

Results: We evaluate the performance of the baseline model on validation dataset, consisting of 30 high-resolution images each containing 8 text lines. We use standard metrics for OCR. The model reaches: character-by-character recall (char_recall): 0.9495, character-by-character precision (char_precision): 0.9539, normalized edit distance (1-N.E.D): 0.9397.

Conclusion: We have created a full-featured system for OCR of the Tibetan script. We implement a stream decoding the scanned text in Tibetan. The results of the project were included in the report of the President of the Russian Academy of Sciences Academician A.M. Sergeev, presented to the President of the Russian Federation V.V. Putin.

Acknowledgements: We acknowledge the financial support of MTS AI LLC. We also thank the Presidium of the SB RAS for the idea and organizational support.

References

1. https://escholarship.org/uc/item/6d5781k5.