Paper ID: IRCUWU2022-407

Recognition of Sinhala machine-printed text for postal address interpretation and postal automation

A.M.P.R.B. Arawa[1,✉], T. Thevathayarajh[1], Y. Mehendran[2] and D.B.B. Senanayake[1]

[1]*Department of Science and Technology, Uva Wellassa University, Badulla, Sri Lanka*
[2]*Department of Computer Science and Informatics, Uva Wellassa University, Badulla, Sri Lanka*

✉*sct17004@std.uwu.ac.lk; +94782636121*

While other Sri Lankan sectors are automating, the Sri Lankan postal system still uses manual intervention for mail sorting and processing. It takes more time to sort the mail according to the postal codes in the central nail exchange, even with the staff having a lot of experience and with the high number of employees while working overtime. The Sinhala language is used by the majority of Sri Lankans in their daily lives. On the other hand, less research has been done on Sinhala letter identification. Several systems have been established for this purpose in other languages including English. However, these types of systems are not available much in Sinhala due to the complexity of the language. Still, the findings have not been highlighted except in the above-mentioned research. Optical Character Recognition (OCR) and image processing technologies were used in the proposed system to recognize Sinhala printed addresses. The Google Tesseract was utilized to produce better optimal results faster and more accurately. Training, testing, and validation were done for the images taken from the printed postal envelopes. The model was trained and tested using the image data obtained under various criteria. Out of 15 Sinhala fonts, this system had an accuracy of 86.67%. A particular type of format was used to write the given addresses. This system can be expanded to include other formats in the future to automate the postal address classification system completely.

Keywords: Image processing; Machine learning; Postal address sorting; Sinhala OCR