20th International Conference on Advances in ICT for Emerging Regions (ICTer 2020): 298 – 299

# Deep Learning Based Sinhala Optical Character Recognition (OCR)

Isuri Anuradha[#1], Chamila Liyanage[#2], Harsha Wijayawardhana[*3], Ruvan Weerasinghe[#4]

[#]*Language Technology Research Laboratory, University of Colombo School of Computing, Colombo, Sri Lanka*

[1]isa@ucsc.cmb.ac.lk, [2]cml@ucsc.cmb.ac.lk, [4]arw@ucsc.cmb.ac.lk

[*]*Theekshana R&D Company, Malalasekera Mw, Colombo 007, Sri Lanka.*

[3]wijayawardhana@gmail.com

2020 20th International Conference on Advances in ICT for Emerging Regions (ICTer) | 978-1-7281-8655-9/20/$31.00 ©2020 IEEE | DOI: 10.1109/ICTer51097.2020.9325428

***Keywords—Sinhala OCR, Optical Character Recognition, Tesseract, Deep learning.***

## I. Introduction

With the advancement of computer technology during the last few years, researchers have integrated machine learning and deep learning techniques to analyse the textual representations on digital documents. As a result of that, people have tended to integrate Optical Character Recognition (OCR) technology to recognize printed texts into machine operable text for different character sets. Sinhala as an abugida script has its own writing system which is used to write Sinhala and Pali languages. With the complexities of the Sinhala script, it makes hard to develop an OCR system. When considering recent literature, most research groups try to reduce the complex nature of the Sinhala script with the support of computer science and Neural networks [1], [2]. Tesseract is an open-source, deep-learning based OCR engine developed by Google [3]. Despite decades of research on the engineering aspects, our attempt was taken to improve the accuracy of Sinhala character recognition using deep learning mechanisms.

## II. Related Works

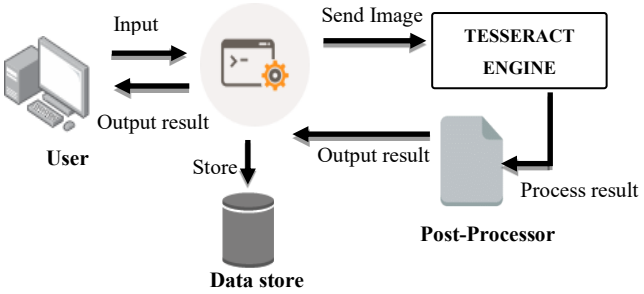
Despite decades of research on the engineering aspects, Sinhala character recognition problem is remaining as a challenging issue in the OCR field. When considering the past few years, some studies have been conducted to identify widely using font types in Sri Lanka [4]. When considering OCR for the Sinhala language, initially the K-Nearest Neighbour (KNN) algorithm-based Sinhala OCR was developed by Language Technology Research Laboratory University of Colombo School of Computing [4].

For the following study, commercially used font types have been employed by varying font sizes to obtain 94% of average accuracy. In addition, Software Development Unit of University of Colombo School of Computing has trained a Sinhala OCR model using Tesseract 3[4]. This system shows relatively good results only for the high-resolution images. Further, Manisha et al. [5] has also tried to combine the Tesseract OCR engine with the Sinhala characters and mentions 97% of accuracy. However, the performance has not been well documented.

## III. System Overview

The proposed system is based on Tesseract 4.0 OCR engine with the Graphical user interface. Mainly system consists with five main subcomponents: The user, API, Tesseract engine, post-processor and a data store. The user: User can access through a web browser and upload document (jpg, jpeg, png, pdf), API: Restful API is responsible for handling user requests, Tesseract engine: Tesseract engine process the image and recognize the image using deep learning techniques, Post-processor: Some characters are not recognized through the tesseract engine. Postprocessor identifies those characters and mapped with linguistic rules and provide accurate output, data store: stores the uploaded document to the system.Prior to that UCSC 10 million word corpus [6] and 400K distinct word list [7] were used to create the deep learning-based LSTM based model combining with different UNICODE font types. We selected 6 popular UNICODE Sinhala fonts available: iskolapotha, Malithi Web, LKLug, Noto Sans, Bhashita Complex and Dinamina. The pre-processing steps for noise removal, adaptive thresholding, page layout analysis and component analysis were performed by the Tesseract 4.0 OCR engine.

Moreover, a web application was developed using Python, HTML and CSS which fitted with the Tesseract environment and hosted in a server which enhance accessibility to the OCR system.




## IV. Results and Conclusion

A comprehensive evaluation was carried out under the categories of complementary Sinhala books, old Sinhala books and old Sinhala newspapers. Considering the Sinhala newspaper category, most of the training images are not in a human-readable format. According to the results of our system, the model trained with font iskolapotha gave an accuracy of 87.63% in contemporary Sinhala books. In the Sinhala old book category, models developed using fonts Malithi Web, LKLug and combined font models using Noto Sans, LKLug and Malithi Web gave accuracies of 87.07%,

978-1-7281-8655-9/20/$31.00 ©2020 IEEE

87.15% and 87.52%. Meanwhile, in the old Sinhala newspaper, category 67.02% of accuracy was obtained from the model developed by font iskolapotha.

Analyzing linguistics rules and mapping them with computer science is quite challenging for low resource languages like Sinhala and Tamil. As future improvements, we will work on identifying touching and conjoining letters which are frequently occurred in Sinhala and Pali writing systems. Moreover, we plan to integrate n-gram or word embedding based post-processing techniques to enhance the accuracy of the proposed system.

## REFERENCES

[1] Rimas, Mohamad, Rohana Priyantha Thilakumara, and PriyarangaKoswatta. "Optical character recognition for Sinhala language." 2013 IEEE Global Humanitarian Technology J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[2] Premachandra, H. Waruna H., et al. "Artificial neural network based sinhala character recognition." International Conference on Computer Vision and Graphics. Springer, Cham, 2016.

[3] "Tesseract documentation", Tesseract OCR, 2020. [Online]. Available: https://tesseract-ocr.github.io/. [Accessed: 27- Nov- 2020].

[4] A. R. Weerasinghe, D. L. Herath, and N. P. K. Medagoda, “A nearest-neighbor based algorithm for printed sinhala character recognition,” Innov. a Knowl. Econ., p. 11, 2006.

[5] U. Manisha and S. R. Liyanage, “Sinhala Character Recognition using Tesseract OCR,” 2018.

[6] "Downloads | Language Technology Research Lab", Ltrl.ucsc.lk, 2020. [Online]. Available: http://ltrl.ucsc.lk/download-3/. [Accessed: 27- Nov- 2020].

[7] "Downloads | Language Technology Research Lab", Ltrl.ucsc.lk, 2020. [Online]. Available: http://ltrl.ucsc.lk/download-3/. [Accessed: 27- Nov- 2020].