Estimating the Effects of Text Genre, Image Resolution and Algorithmic Complexity needed for Sinhala Optical Character Recognition

Isuri Anuradha[#1], Chamila Liyanage[2], Ruvan Weerasinghe[3]

Abstract— **While optical character recognition for Latin based scripts have seen near human quality performance, the accuracy for the rounded scripts of South Asia still lags behind. Work on Sinhala OCR has mainly reported on performance on constrained classes of font faces and so been inconclusive. This paper provides a comprehensive series of experiments using conventional machine learning as well as deep learning on texts and font faces of diverse types and in diverse resolutions, in order to present a realistic estimation of the complexity of recognizing the rounded script of Sinhala. While texts of both old and contemporary books can be recognized with over 87% accuracy, those in old newspapers are much harder to recognize owing to poor print quality and resolution.**

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

I. Introduction

Optical Character Recognition (OCR) technology is designed to recognize printed texts into machine operable text. OCR is a collection of multiple steps such as scanning, pre-processing, segmentation, feature extraction, classification, recognition and post-processing. In recent literature, many OCR systems have been developed for recognizing Latin characters [1]. With the advancement of Natural Language Processing during the past few years, researchers have integrated machine learning/deep learning techniques for analysing the textual representations on digital documents. Template Matching, Neural Network (NN) and Recurrent Neural Network (RNN) are popular and widely used algorithms for character recognition. These technologies are better when applied for the other character sets, since large volumes of data are available in print media for many languages. The proposed Sinhala OCR is discussed in this paper with special focus on the text genre, image resolution and algorithmic complexities needed for training an OCR system for the Sinhala character set.

As the state of the art OCR technology, currently Tesseract is used in the training of OCR systems for many character sets. Further, Tesseract has moved from machine learning to deep learning with LSTM architecture and provides relatively better recognition competence [2]. However, algorithmic complexity is not enough for training an OCR model, as text genre and image quality affect training a more accurate OCR model. Since large volumes of available data are in print media and they have been printed before the computer era, the documents have been printed using different techniques such as offset printing and screen printing. Therefore, common type-faces used in the history of printing should also be trained to train the model to get such text recognizable. Further, types and sizes of the fonts and size of the training text is also significant. In this paper we discuss the OCR system developed for Sinhala by estimating the effects of text genre, image resolution and algorithmic complexity.

The rest of this paper is structured as follows: Section II gives a brief overview of the related work in this area. Section III discusses some properties and characteristics of the Sinhala script as it is significant to review the complexities with regard to the particular script. Algorithmic complicacy adopted to OCR is discussed in section IV. Further, section V gives the motivation and rationale for the experimental set and systematic description on training data, word lists, and training regime adopted to develop the Sinhala OCR. Section VI presents experimental results on the OCR methods, and we also give an analysis of their performance comparison. Finally, the paper is concluded with a discussion of future works.

II. Related Works

Despite decades of research on the engineering aspects, the problem of Sinhala character recognition remains as a challenging issue in the OCR field. When the past few years are considered, some studies have been conducted to identify widely used font types in Sri Lanka [3]. When considering OCR for the Sinhala language, initially the K-Nearest Neighbour (KNN) algorithm-based Sinhala OCR was developed by the Language Technology Research Laboratory, University of Colombo School of Computing [3]. For the following study, commercially used font types have been employed by varying font sizes to obtain 94% of average accuracy.

Considering literature, Neural Network based Sinhala OCR systems have been developed in recent years [4], [5], [6]. In 2013, the Sri Lanka Institute of Information Technology conducted a research based on applying neural networks for Sinhala optical character recognition [4]. In this study they have only focused on 36 characters in the alphabet. Another Sinhala OCR application integrating neural networks was developed by a local research group [5]. These studies mainly focused on the character level accuracies and not on word accuracies.

Correspondence: I. Anuradha [#1] (E-mail isa@ucsc.cmb.ac.lk)
Received: 29-12-2020 Revised:31-05-2021 Accepted: 11-06-2021

I. Anuradha[#1] C Liyanage[2], R. Weerasinghe[3], are form University of Colombo School of Computing. (isa@ucsc.cmb.ac.lk)

This paper is an extended version of the paper "*Deep Learning Based Sinhala Optical Character Recognition (OCR)*" presented at the ICTer Conference (2020)
DOI: http://doi.org/10.4038/icter.v14i3.7231

© 2021 International Journal on Advances in ICT for Emerging Regions

In addition, the Software Development Unit of University of Colombo School of Computing has trained a Sinhala OCR model using Tesseract 3 [7]. This system shows relatively good results only for the high-resolution images. Also, Language Technology and Research laboratory at University of Colombo is experimenting on the integration process of machine learning concepts to Sinhala OCR applications [8]. Further, Manisha et al. [9] 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. It's well-known that Indic languages have many complexities and variations of characters which makes OCR systems hard to develop. But in the past few years, multiple studies have been conducted integrating Tesseract OCR engine for character recognition using different low resource languages such as Tamil [10], Hindi [11], Bengali [12] and Urdu [13].

III. Sinhala Script

The Sinhala script is an *abugida* or *alphasyllabary* script in which consonant-vowel sequences are written as units and thereby it is called a segmental writing system. The script has evolved from the *Brahmi* script. The letters in Sinhala are circular-shaped and are written from left to right [14]. The Sinhala script is used primarily to write the Sinhala language, which is one of the official languages of Sri Lanka spoken by about 16 million people in the country. In addition, it is also used in Sri Lanka for writing Pali, the canonical language of Theravada Buddhism, and sometimes Sanskrit, the Old Indo-Aryan language [15].

There are 20 vowels and 41 consonants in the Sinhala script. Since Sinhala is a segmental writing system, vowels take two representations as independent vowels: occur in the initial position of a word (infrequently occur in the middle of a word: E.g. නුවරඑළිය, ජාඇල) and dependent vowels also known as vowel modifiers: occur after a consonant. Figure 1 and 2 illustrate the vowels with their modifiers and consonants in Sinhala script respectively.

Vowels and modifiers included for the training data

ආ	◌ා	උ	◌ු	ඒ	◌ේ
ඇ	◌ැ	ඌ	◌ූ	ඓ	◌ෛ
ඈ	◌ෑ	ඍ	◌ෘ	ඔ	◌ො
ඉ	◌ි	ඎ	◌ෲ	ඕ	◌ෝ
ඊ	◌ී	එ	◌ෙ	ඖ	◌ෞ
			◌ං		◌ඃ

Vowels and modifiers not included for the training data

ඏ	◌ෟ	ඐ	◌ෳ

Fig. 1 Vowel characters and modifiers in Sinhala script

From among the vowel modifiers in figure 1, ◌ං (*anusvara*) and ◌ඃ (*visarga*) are two specific modifiers. They occur not only with consonants but also with vowels. E.g. අං, ඉං, උං, අඃ, ඕඃ.

Consonants included for the training data

ක	ඛ	ග	ඝ		ඟ	ළ	හ
ච	ඡ	ජ	ඣ	ඤ	ඥ	ය	ශ
ට	ඨ	ඩ	ඪ	ණ	ඬ	ර	ෂ
ත	ථ	ද	ධ	න	ඳ	ල	ස
ප	ඵ	බ	භ	ම	ඹ	ව	ෆ

Consonants not included for the training data

ඞ	ඦ

Fig. 2 Consonant characters in Sinhala script

Two vowels: ඏ, ඐ and their corresponding vowel modifiers in figure 1 and ඦ in figure 2 were not included for the training data as they do not occur in old or contemporary Sinhala books. However, ඞ in figure 2 occurs in a limited number of words in old Sinhala books. It was not included because the shape of the particular character would cause misrecognition with similar characters in Sinhala script.

Sinhala consonants imply the inherent vowel /a/ (අ) when they are occur with no modifiers. Absence of the inherent vowel is marked by adding a symbol called *hal lakuna* or *halkirima* to the top of the particular consonant. E.g. ක්, ච්. Further, *hal lakuna* also occurs with two vowels and their modifiers. It has two shapes as illustrated in figure 3.

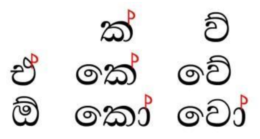


Fig. 3 Two different shapes of *hal lakuna* with vowels and consonants

As a segmental writing system, vowel modifiers appear above, below or to the right or left of the basic consonant. From all the consonant-vowel sequences in Sinhala script, ඍ is a special character as it appears as a separate symbol to represent ළ+උ sequence. As an example, following figure 4 illustrates all the consonant-vowel sequences for consonant 'ක'.

කා	කැ	කෑ	කි	කී	කු
කූ	කෘ	කෲ	කෙ	කේ	කෛ
කො	කෝ	කෞ	කං	කඃ	

Fig. 4 Consonant 'ක' with all the vowel modifiers

There are three consonant modifiers which occur in the Sinhala script, known as *rakaranshaya* (්‍ර), *yanshaya* (්‍ය) and *rephaya* (ර්‍). Among them *rakaranshaya* represents 'ර' (ra) and *yanshaya* represents 'ය' (ya) when they appear after a consonant (from which the inherent vowel has been removed). However, as symbols, *rakaranshaya* appears below (e.g. ක්‍රම, ආශ්‍රය, වක්‍ර) and *yanshaya* to the right (ව්‍යසන, සත්‍ය,

සංඛ්‍යාව) of the basic consonant. Further, *rephaya* is also used to denote 'ර්' when it occurs before a consonant and the symbol appears on top of the basic consonant (e.g. ධර්ම, සර්ව, තර්ක). Using *rephaya* is an alternative rule in the Sinhala writing system while *rakarakshaya* and *yanshaya* are essential. All the vowel modifiers surround the consonant-*rakaranshaya* (e.g. ක්‍රෝ), consonant-*yanshaya* (e.g. ක්‍යෝ) or *rephaya*-consonant (e.g. ර්කෝ) units. Figure 5 illustrates how vowel modifiers occur with *rakaranshaya*.

ක්‍ර	ක්‍රා	ක්‍රැ	ක්‍රෑ	ක්‍රි	ක්‍රී
ක්‍රැ	ක්‍රෑ	ක්‍රෙ	ක්‍රේ	ක්‍රෛ	ක්‍රො
ක්‍රෝ	ක්‍රෞ	ක්‍රං	ක්‍රඃ		

Fig. 5 Consonant ‘ක’ with *rakaranshaya* and the vowel modifiers

One other significant characteristic in Sinhala writing system is using compound consonants. This frequently occurred in old Sinhala books. However, in contemporary Sinhala this writing system is infrequent and therefore only the first set of compound consonants in figure 6 (which are rarely occurred in contemporary Sinhala books) have been concerned for the training data in this research.

Compound consonants rarely occurred in contemporary Sinhala books	
ක්‍ව	ක + ZWJ + ව
ක්‍ෂ	ක + ZWJ + ෂ
ග්‍ධ	ග + ZWJ + ධ
ත්‍ථ	ත + ZWJ + ථ
ත්‍ව	ත + ZWJ + ව
න්‍ථ	න + ZWJ + ථ
න්‍ද	න + ZWJ + ද
න්‍ධ	න + ZWJ + ධ
න්‍ව	න + ZWJ + ව
ද්‍ය	ද + ZWJ + ය
Compound consonants occurred in old books	
ඬ්‍ව	ඬ + ZWJ + ව
ට්‍ඨ	ට + ZWJ + ඨ
ද්‍ධ	ද + ZWJ + ධ
ද්‍ව	ද + ZWJ + ව

Fig. 6 Compound consonants in Sinhala script

IV. System Overview

In our study, different Sinhala text genres were given different accuracy results. From a variety of genres, explanation and descriptive writings, narrative writings, and news reportage were selected for our purpose. When selecting documents, we considered a variety of documents and Unicode font types from different printing eras. When image resolutions were considered, low image resolutions may affect not only quality but also speed degradation of overall OCR performance, since uncertainty in character pictures produce more recognition variants. In the *Tesseract* engine also, high resolution images were able to give high accuracy by identifying all the punctuations, modifiers and complex letters. In the Tesseract engine, image processing is a combination of several steps such as rescaling, Binarization, Dilation / Erosion, and etc.

For the training process, we adapted and experimented on both *Tesseract 3.0* (Legacy version) and *Tesseract 4.0* (Deep learning) OCR engines as a tool. *Tesseract* has a standard level of accuracy in its engine. It’s necessary to have a library file in the OCR engine called ‘*traineddata*’ which works on Sinhala inputs. This file is a concatenation of multiple files. According to the accuracy and richness in the library file, the OCR engine can work to its full potential. Sinhala language is complicated and has various types of letters including vowels, consonants, compound characters and other special types. Therefore, for *Tesseract 3.0*, we developed a large character set for Sinhala. It is important to mention that, for *Tesseract 3.0* we need to uniquely identify each and every character. Sometimes due to the complexity of the character set, the OCR may not always detect a character correctly even if the character is included in the training files.

The preparation of data and the training process adopted for developing the Sinhala OCR model for both tesseract 3.0 and 4.0 versions are described in the following subsections.

V. Training Process

The preparation of data and the training process adopted for developing the Sinhala OCR model for both T*esseract* 3.0 and 4.0 versions are described in the following subsections.

- Preparation Tesseract version 3.0

A. *Setting up the OCR Engine*

We installed the *Tesseract version 3.0* in the Windows Operating System. Since there is no user interface of *Tesseract 3.0*, we used several commands in the command line to launch the application.

B. *Preparing training data*

The process followed by preparing training data is described below.

1) Preparing OCR alphabet for Sinhala: Initially the OCR alphabet for Sinhala was defined to collect text data. The OCR alphabet is distinct from the character alphabet which includes the basic units of training data as follows.

- All vowels: e.g. අ ඉ උ එ ඒ ඔ ඕ
- All consonants e.g. ක ව ඟ ඳ ඝ
- Consonants with touching modifiers: e.g. කි කු චි
- Consonants with *hal lakuna*: e.g. ක් ච්
- Compound consonants: e.g. ක්‍ර ව්‍ර ක්‍ෂ ත්‍ය
- Compound consonants with touching modifiers: e.g. ක්‍රි ව්‍රි ක්‍ෂි ක්‍ෂු ත්‍යී ක්‍රු
- Non-touching modifiers: e.g. ෙ ා ෘ ෲ
- Punctuation marks: e.g. ! ? ( ) /

2) Text data collection: Text data for preparing training images were collected from the UCSC 10M Sinhala corpus and these were extracted as per the OCR alphabet that we defined. We selected 1050 words from extracted data for the preparation of training images. figure 7 shows a sample of text data.

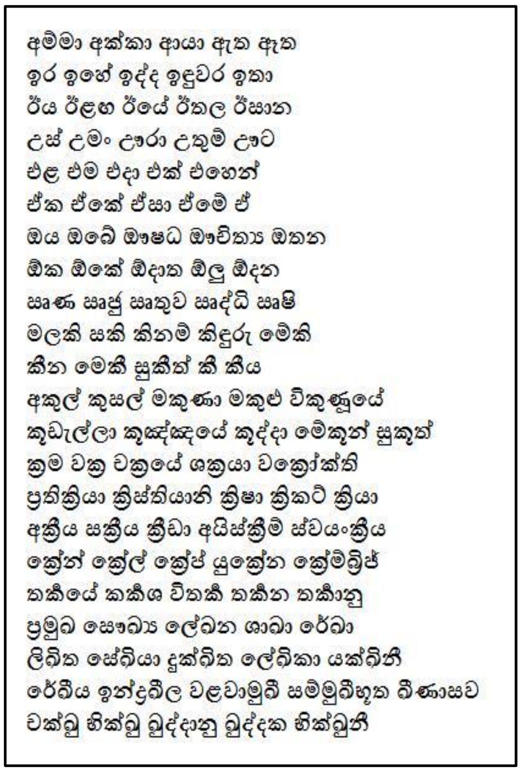

අම්මා අක්කා ආයා ඇත ඇත
ඉර ඉහේ ඉද්ද ඉඳුවර ඉතා
ඊය ඊළඟ ඊයේ ඊතල ඊසාන
උස් උමං ඌරා උකුම් ඌට
එළ එම එදා එක් එහෙන්
ඒක ඒකේ ඒසා ඒමේ ඒ
ඔය ඔබේ ඖෂධ ඖචිත්‍ය ඔතන
ඕක ඕකේ ඕදාත ඕලු ඕදන
ඍණ ඍජු ඍතුව ඍද්ධි ඍෂි
මලකි සකි කිනම් කිඳුරු මෙකි
කීන මෙකී සුකීත් කී කීය
අකුල් කුසල් මකුණා මකුළු විකුණූයේ
කුඩැල්ලා කුඤ්ඤයේ කුද්දා මෙකුන් සුකුත්
ක්‍රම වක්‍ර චක්‍රයේ ශක්‍රයා වක්‍රෝක්ති
ප්‍රතික්‍රියා ක්‍රිස්තියානි ක්‍රිෂා ක්‍රිකට් ක්‍රියා
අක්‍රීය සක්‍රීය ක්‍රීඩා අයිස්ක්‍රීම් ස්වයංක්‍රීය
ක්‍රේන් ක්‍රේල් ක්‍රේප් යුක්‍රේන ක්‍රේම්බ්‍රිජ්
තර්කයේ කර්කශ විතර්ක තර්කන තර්කානු
ප්‍රමුඛ සෞඛ්‍ය ලේඛන ශාඛා රේඛා
ලිඛිත සේඛියා දුක්ඛිත ලේඛිකා යක්ඛිනී
රේඛීය ඉන්ද්‍රඛීල වළවාමුඛී සම්මුඛීභූත ඛීණාසව
චක්ඛු භික්ඛු ඛුද්දානු ඛුද්දක භික්ඛුනී

Fig. 7 Sample of text data

3) Preparing training Image: Training images were prepared for fonts widely used in Sinhala typing. The criteria followed to create text images are as follows.

Colour: *grayscale*
Font size of the text: *12, 14, 16*
DPI: *300*
Fonts used: *Iskoola Pota, FMAbhaya, Malithi Web, BhashithaScreen, DinaminaUniWeb*

Based on the above criteria we prepared two sets of training images. The first set consisted of computer-generated images (screenshots). As an iterative process of training, the second set of training images were prepared with scanned images for the same text data. Figure 8 shows a sample of such training images.

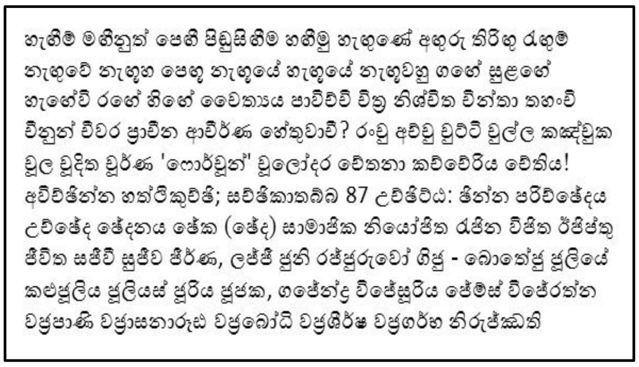

හැඟීම් මඟීනුත් පෙඟී පිඬුසිඟීම හඟීමු හැඟුණේ අඟුරු තිරිඟු රැඟුම්
නැඟුවේ නැඟූහ පෙඟූ නැඟූයේ හැඟූයේ නැඟූවහු ගඟේ සුළඟේ
හැඟේවි රඟේ හිඟේ චෛත්‍යය පාච්චිච් චිත්‍ර නිශ්චිත චින්තා තහංචි
චිනුන් චිවර ප්‍රාචීන ආචීර්ණ හේතුවාචී? රංචු අච්චු චුට්ටි චුල්ල කඤ්චුක
චූල චූදිත චූර්ණ 'පොර්චූන්' චූලෝදර චේතනා කච්චේරිය චේතිය!
අච්ඡින්න හත්ථිකුච්ඡි; සච්ඡිකාතබ්බ 87 උච්ඡිට්ඨ: ඡින්න පරිච්ඡේදය
උච්ඡේද ඡේදනය ඡේක (ඡේද) සාමාජික නියෝජිත රැජින විජිත ඊජිප්තු
ජීවිත සජීවී සුජීව ජීර්ණ, ලජ්ජී ජුනි රජ්ජුරුවෝ ගිජු - බොතේජු ජූලියේ
කළුජූලිය ජූලියස් ජූරිය ජූජක, ගජේන්ද්‍ර විජේසූරිය ජේම්ස් විජේරත්න
වජ්‍රපාණි වජ්‍රාසනාරූඪ වජ්‍රබෝධි වජ්‍රශීර්ෂ වජ්‍රගර්භ නිරුජ්ඣති

Fig. 8 Sample of training images

4) Character Segmentation: In *Tesseract 3.0*, character segmentation is performed using the process of creating box files. A 'box file' is a text file, which contains the necessary information of the training images. The coordinate values of the characters in the training images along with corresponding Unicode characters are stored in these box files. The segmentation of the characters in the training images was done as per the OCR alphabet we defined. Each training image should have a box file in which the number of boxes must tally with the number of training character segments in the image. Figures 9 and 10 show an image with segmented characters and a sample of box information.

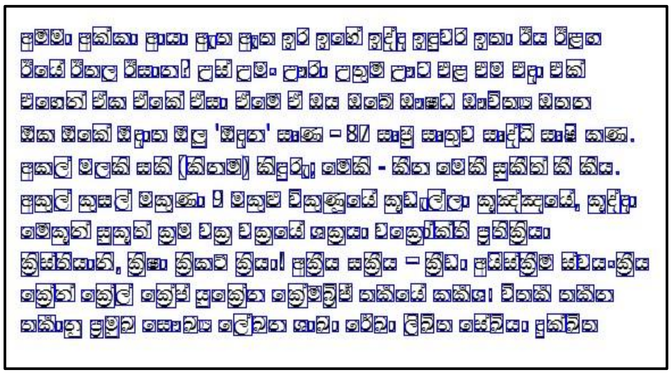
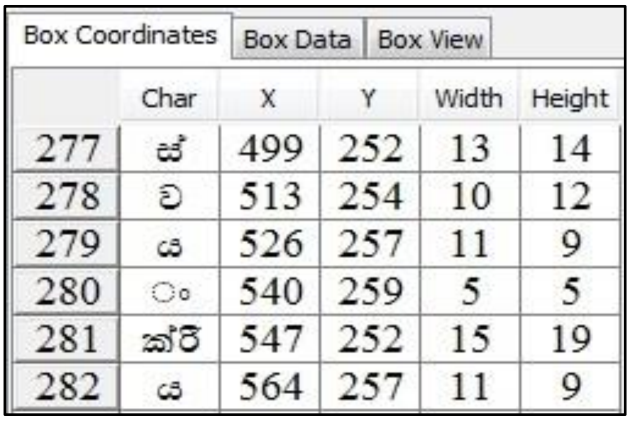


Fig. 9 Character segmentation



	Char	X	Y	Width	Height
277	ස්	499	252	13	14
278	ව	513	254	10	12
279	ය	526	257	11	9
280	◌ං	540	259	5	5
281	ක්රි	547	252	15	19
282	ය	564	257	11	9

Fig. 10 Sample of box information

C. Training the model

The training was performed as an iterative process until better results were obtained. Firstly, the training models were done for individual data sets of computer-generated images for given font types and sizes. Secondly, we combined the training data sets for multiple fonts and multiple sizes and trained the models. Thirdly, the training was performed using the scanned images and trained multiple models for the given font types and sizes. Finally, all the data sets of computer-generated images and scanned images were combined in several ways and trained multiple models.

- *Preparation Tesseract 4.0 version*

A. Setting up the OCR Engine

For setting up the *Tesseract 4.0* version we selected Ubuntu environment. Since *Tesseract 4.0* deals with deep learning techniques such as Long Short-Term Memory (LSTM), the Ubuntu operating system provides full compatibility for OCR engines. And all the tasks were carried out in the terminal and instructions were given as commands.

B. Preparation of training data sets

Training data plays an important role in *Tesseract version 4.0*. With the integration of deep learning techniques, more training data will result in good outcomes. For our experiment, we have employed 3 datasets which are available for the Sinhala language. Further details of the *1)* UCSC 10 million Sinhala dataset, *2)* common crawler Sinhala dataset and *3)* Google dataset will be discussed in the next few lines.

1) UCSC 10 Million Word Sinhala Corpus: UCSC 10M Word Sinhala Corpus has been compiled by the Language Technology Research Laboratory - University of Colombo School of Computing (UCSC) in Sri Lanka. This text corpus contains a huge variety of Sinhala books including novels, short stories, translations, critiques written by renowned Sinhala writers, and Sinhala newspapers: Silumina, Dinamina, Lankadeepa and Lakbima. The UCSC 10 million dataset includes texts which belong to different eras in Sri Lanka. It also contains texts from various sources; the text is rich with different writings. Noise data and other textual data with different languages have been removed from this dataset in order to minimize the errors.

2) 5million+ sentences in Sinhala common crawler: In 2019, Guzman [16] presented two monolingual corpora for Sinhala. Those were a combination of 155k+ sentences of filtered Sinhala Wikipedia and 5178k+ sentences of Sinhala common crawl. Since this study considered only textual data available online, the diversity of textual representation is considerably low. Furthermore, a high noise rate exists in this dataset with other common issues like the zero width joiner problem and the combination of multiple language textual data with Sinhala textual data. And these affect the overall accuracy of the system.

3) Google dataset for Sinhala is especially built with the Tesseract. This dataset includes variety of textual representations gathered in recent years.

4) UCSC 400K distinct wordlist: This list of monolingual vocabulary was developed from the UCSC 10 million words Sinhala corpus by the Language Technology Research Lab of UCSC. The list includes 440,021 distinct entries and is available on the web.

After comparing these 3 datasets, the UCSC 10 million Sinhala dataset [17] was selected by the authors due to the enrichment of textual combinations in different eras and less noise data. UCSC 400K Distinct Word List [18] was also combined with the existing Tesseract word list.

As a special feature, the *Tesseract version 4.0* generates the tiff file and box file automatically. Additionally, image and corresponding UTF-8 text transcription are generated on lstmf file at the process of font training. Also in T*esseract 4.0* the clustering steps (mftraining, cntraining, shape clustering) are replaced with a single slow lstm training step.

මේ ´කලියුගය´ වූ කලි ´ගම් පෙරළිය´ හා ´යුගාන්තය´ එකට
ඇඳීමෙන් ඒ තුන, තුන් වැදෑරුම් එකම කතාන්තරයක් කරන්නකි.
බටහිර සභ්‍යත්වය රිංගීම නිසා මඳින් මඳ වෙනස් වන්ට වූ දකුණු
පළාතේ ගැමි පවුලක ජීවිතයත් ගැමි සමාජයත් ගම් පෙරළියෙහි
කථාවට වස්තු වී ය. එහි පසළොස් වන පරිච්ඡේදයෙහි මේ කියුම
දක්නා ලැබේ.
´´මංගලොත්සවයෙන් පසු කොළඹ ගිය පියල් ඇතැම් විට ගමට
එන්නේ සතියකට වරකි; ඇතැම් විට දෙසතියකට වරකි. ක්‍රමයෙන්
දියුණු වන වෙළඳාම උදෙසා තම කාලය වඩාත් යෙදිය යුතු වූයෙන්
පියල් කොළඹ පදිංචියට යාමට සිතුවේ ය. කලකට පෙර ඔහු මෝදර
තනවන්ට ආරම්භ කළ ගෙයි වැඩ නිම කෙළේ විවාහයට දෙමසකට
පෙර ය. ඔහු ඒ ගෙය තනවන්ට පටන් ගත්තේ කුලියට දෙනු පිණිස
මිස තමාගේ පදිංචිය පිණිස නොවේ. තමා සෑද වූ ඒ ගෙයින් ගත යුතු
ප්‍රයෝජනය පියල්ට දැන් වැටහෙයි.´´
පියල් හා නන්දා, අනුලා හා තිස්ස ද සමඟ කොළඹට සංක්‍රමණය වී
මෝදර අර ගෙයි පදිංචි වෙති. මේ ´කලියුගය´ ඔවුන්ගේත් ඔවුන්ගේ
දරුවන්ගේත් ජීවිතය වර්ණනා කරනු පිණිස පබැඳුණකි.

Fig. 11 Sample of training data for Tesseract 4.0

C. Selection of font types and sizes

Since typefaces are significant in training an OCR system, we investigated the commonly used Sinhala fonts to train the OCR model in *Tesseract 4.0*. Though there are hundreds of non-Unicode fonts available for the Sinhala script, they have no unique character code point for identification. Owing to its 16-bit encoding, UNICODE is theoretically able to support over 65,000 unique character code points [19] and we selected 9 Unicode fonts from the limited number of Sinhala Unicode fonts available. They include Unicode fonts which are most commonly used in printed and digital media [20]. The font types involved with the research is given below.

- Noto-Sans font
- LKLug font
- Malithi font
- Dinamina font
- Iskolapotha font
- BhashitaComplex font

D. Training the model

As pre-processing steps for noise removal, adaptive thresholding, page layout analysis and connected component analysis were performed by the *Tesseract* OCR engine. The following steps were followed to train the model. Initially generated training data is provided as the input to the engine and extract the generated model. Then the model was fine tuned to decrease the error rate and finally the fine-tuned model was combined with the initial trained model. We combine multiple fonts for model creation. Single font models, Double font models and Triple font models were used for analysis.

VI. Evaluation and Results

For the evaluation process, we considered both *Tesseract 3.0* and *Tesseract 4.0*. As the first phase, the *Tesseract version 3.0* was evaluated by character level. Meanwhile, the *Tesseract 4.0* was evaluated at both character and word level. The developed OCR models have been tested with 30 images selected for three different categories (10 for each category). When selecting images for testing we chose non identical images with different typefaces and different image qualities.

1) Old Sinhala newspapers: The testing data for this category was selected from archived images of Sinhala newspapers published in 1870 – 1890. The newspapers include: සත්‍යෝදය (*sathyodaya*), සත්‍යාලංකාරය (*sathyalankaraya*) and දිනපතා ප්‍රවෘත්ති (*dinapatha prawruththi*). All the images are in 200 DPI.

2) Old Sinhala books: Testing images for this category were selected from old Sinhala books which are printed on Letterpress printing. The old books selected include: බුත්සරණ (*buthsarana*), පූජාවලිය (*pujawaliya*) and සද්ධර්මරත්නාවලිය (*saddharmarathnawaliya*). The images in this category are in 72 DPI.

3) Contemporary Sinhala books: The books printed with computerized fonts were selected for this category. 10 images of randomly selected pages from 10 books were taken and they were scanned for 300 DPI.

To calculate the accuracy of the systems we compare the common and different characters between original and OCR document.

$$\text{Accuracy (\%)} = \frac{\text{Count of common characters}}{\text{Count of (common characters + different characters)}} \text{ X } 100\%$$

A. Evaluation of the models from Tesseract 3.0

The evaluation of *Tesseract version 3.0* was conducted only for the third category of testing images for two reasons. Firstly, the results for the other two categories were not at satisfying level and secondly, we gave our main priority for the evaluation of *Tesseract version 4.0* Therefore, we selected the most accurate model (*Scanned-iskolapotha* model) out of 18 multiple models created by varying different font types and sizes. Original data of the testing samples consist of 2592 words and 16380 characters. Testing results are illustrated in table i.

TABLE I
TESSERACT 3.0 RESULTS OF OCR DOCUMENTS

Font type	**Recognized character count**	**Recognized word count**	**Accuracy of the system**
Iskolapotha	16962	2507	36.89%

B. Evaluation of the models from Tesseract 4.0

The models generated from *Tesseract 4.0* OCR engine were evaluated for the three categories of testing samples explained above. From the generated models, all the models of individual fonts and three selected Combined Models (CM) were evaluated. The same set of testing images were used in the evaluation process.

For the first category of evaluation, we selected 10 images from old newspapers and they consist of 1557 words and 9821 characters. Some of the texts in these images are even hard to read by a human. The results for the first category of images are shown in table ii.

TABLE II
CATEGORY 01 RESULTS OF OCR DOCUMENTS

Font type	**Recognized character count**	**Recognized word count**	**Accuracy of the system**
Noto-Sans	10142	1462	61.43%
LK-LUG	10031	1441	61.66%
Malithi	10094	1516	65.51%
Iskolapotha	10067	1458	67.02%
Dinamina	9897	1426	59.83%
Bashitha	10056	1451	61.96%
Noto-LKLug (CM)	10071	1458	61.51%
Malithi-Lug (CM)	10003	1449	63.30%
Noto-Lug-Malithi (CM)	10035	1445	64.40%

There were 3032 words and 18074 of characters in the images of category 2, the old books printed in letterpress printing era. The results obtained are illustrated in table iii.

TABLE III
CATEGORY 02 RESULTS OF OCR DOCUMENTS

Font type	Recognized character count	Recognized word count	Accuracy of the system
Noto-Sans	18623	3019	85.13%
LK-LUG	18584	3012	87.15%
Malithi	18774	3039	87.07%
iskolapotha	18688	3022	85.28%
Dinamina	18428	2807	84.97%
Bashita	18387	2983	87.53%
Noto-LKLug (CM)	18627	3023	86.06%
Malithi-Lug (CM)	18704	3017	85.69%
Noto-Lug-Malithi (CM)	18461	3010	87.52%

Third category of 10 images captured from contemporary Sinhala books and they consist 2592 words and 16151 characters. The results are denoted in table iv.

TABLE IV
CATEGORY 03 RESULTS OF OCR DOCUMENTS

Font type	Recognized character count	Recognized word count	Accuracy of the system
Noto-Sans	16476	2620	85.91%
LK-LUG	16306	2615	84.91%
Malithi	16530	2626	86.14%
iskolapotha	16391	2635	87.63%
Dinamina	16104	2459	85.49%
Bashita	16178	2591	87.49%
Noto-LKLug (CM)	16335	2627	84.83%
Malithi-Lug (CM)	16338	2796	85.49%
Noto-Lug-Malithi(CM)	16259	2613	86.59%

When three categories of input image sets were considered, old newspapers have a low accuracy rate in character recognition due to the high existence of noises and low quality of images. In the old Sinhala book category *Malithi*, *LKLug* and the combined model of *Noto Sans*, *LKLug* and *Malithi* were given best accuracy when recognizing characters. The model trained with the font I*skolapotha* obtained the highest accuracy rate in 3rd category of contemporary Sinhala books. When analysing OCR outputs, some identifiable improvement can be done in the recognition process of the system.

Moreover, as another part of this research we randomly selected 5 images from the contemporary Sinhala books and converted for low DPI count (from 300px to 96px). Thereafter, we chose our best 3 models in the contemporary Sinhala books category and evaluated the performance. A total count of 1482 words and 8735 characters were there in the selected 5 image data.

TABLE V
COMPARISON WITH THE LOW DPI LEVEL

Font type	Recognized character count	Recognized word count	Accuracy of the system
iskolapotha	8983	1481	87.88%
Malithi	9067	1479	86.09%
Noto-Lug-Malithi (CM)	8895	1473	87.40%

As a special note, after reducing DPI count, some character combinations which were not recognized correctly but samples with 300 DPI were able to recognized. For instance *yanshaya* was not recognized well in previous efforts but with this modification *yanshaya* was recognized well. Some clearly identifiable errors in the recognition process of our OCR system has been briefly noted below.

- Confusing with similar shaped individual characters. (e.g. හ - හ - ග - භ - ශ, ච - ච - ට, ඹ - ඹ - ම, එ - එ - ඵ, ඬ - ඞ, ද - ඳ, ය - ස - ඝ, ජ - ඡ, බ - ඛ, ඨ - ඪ, ත - න) Errors of this type frequently occur in 1st and 2nd categories of testing images. When the images are not clear, the tiny variations of the characters are difficult to be captured.
- Inability to recognize *hal lakuna*. (e.g. ක් - ක, කෙ - කේ, ච - ච්, චෙ - චේ)
- Confusing with vowel modifiers and *hal lakuna* in the same character or in different characters. (e.g. චි - චී - ච - චි - චී - ච, හි - හී - හි - හී, මි - මී - ම, මු - මූ, සි - සී)
- Misidentification of *rakaranshaya* with *papilla*, the vowel modifier for 'u'. (e.g. ප්‍ර - පු, ව්‍ර - වු, ම්‍ර - මු)
- Inability to recognize touching letters. (e.g. සස - ස්ස, දව - ද්ව, ටඨ - ට්ඨ, ණණ - ණ්ණ, කබ - ක්බ) Using touching letters was a writing style in old Sinhala. And in Pali it is a rule, as Pali does not have a sign like *hal lakuna* to show the absence of the inherent vowel. This writing style can occur in some testing images of category 1 and 2. Since

contemporary writing does not follow this style, the training data is not rich with these sequences. This has resulted in not recognizing touching letters.

- Inability to recognize compound consonants. The compound consonants given in figure 6 hardly occur in contemporary Sinhala and therefore they are not well recognized.

Some English characters were also in the testing images of all three categories. As we focused on developing a better recognition model for Sinhala characters, we did not include enough English text data in the training process, this resulted in some errors in recognition and affected the overall accuracy of the system. However, the above limitations will be considered in the next stage as a future enhancement.

VII. Conclusion and Future Works

In this paper we presented a process of developing an Optical Character Recognition system for Sinhala. In this research we identified the characteristics of Sinhala script along with properties of writing style in Sinhala scripts. The training process of the OCR model was initiated with *Tesseract 3.0* and later moved to *Tesseract 4.0*, as it was the state of the art of deep learning.

The evaluation was carried out by comparing the results with the different types of Sinhala fonts and adapting systems to recognize varieties of test data gathered from different sources. Although we tested some samples with the model built from a Sinhala common crawl dataset, overall accuracy is less compared to others and unable to identify characters.

According to the results, our system model trained with font *iskolapotha* gave accuracy of 87.63% in contemporary Sinhala books. In the Sinhala old book category, models developed using fonts *Malithi*, *LKLug* and combined font models using *Noto Sans*, *LKLug* and *Malithi* gave accuracies of 87.07%, 87.15% and 87.52% respectively. Meanwhile in the old Sinhala newspaper category 67.02% of accuracy was obtained from the model developed with font *iskolapotha*.

Developing OCR systems for low resource languages needed a considerable amount of effort from both linguistics and computer science domain areas. Analysing linguistics rules and mapping them with computer science is quite challenging for low resources languages like Sinhala and Tamil. In this stage of the research, we focused highly, only on the recognition of the Sinhala script. As mentioned in the above sections, the Sinhala script is also used to write “Pali” and “Sanskrit” languages in Sri Lanka. As a future enhancement we will work on identifying touching and conjoining letters which occur frequently in Pali. We also plan to apply some n-gram or word embedding based post-processing techniques to enhance the accuracy. Also in real world OCR can be categorized as one of sequence learning tasks. Therefore, it is necessary to predict the sequence of labels from noisy, unsegmented input data. As future work, we plan to combine connectionist temporal classification (CTC) with deep learning algorithms to train the Recurrent Neural Network (RNN) to label unsegmented sequences directly. Moreover, neural net compressions and conventional neural machine translations for Sinhala OCR will be studied in the future.

Acknowledgment

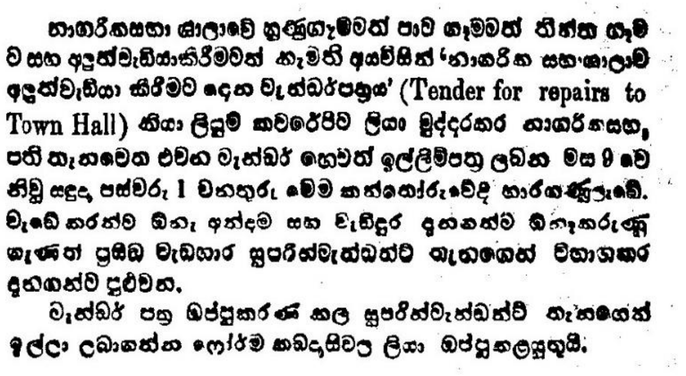
This work was carried out as a part of a project funded by Theekshana - Research and Development Company. We acknowledge Mrs. Dinuji Godigamuwa for assisting the work and thank all the members of Language Technology Research Laboratory of the University of Colombo School of Computing, Sri Lanka, who helped in various ways to make this work bear fruit.

References

[1] R. Weerasinghe, A. Wasala, D. Herath, and V. Welgama, “Nlp applications of sinhala: Tts & ocr,” in Proceedings of the Third International Joint Conference on Natural Language Processing: Volume-II, 2008.
[2] Smith, R. (2007) ‘An Overview of the Tesseract OCR Engine’, in Ninth International Conference on Document Analysis and Recognition (ICDAR 2007), pp. 629–633. doi: 10.1109/ICDAR.2007.4376991.
[3] 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.
[4] M. Rimas, R. P. Thilakumara, and P. Koswatta, “Optical character recognition for Sinhala language,” in 2013 IEEE Global Humanitarian Technology Conference: South Asia Satellite (GHTC-SAS), 2013, pp. 149–153.
[5] S. Ajward, N. Jayasundara, S. Madushika, and R. Ragel, “Converting printed Sinhala documents to formatted editable text,” in 2010 Fifth International Conference on Information and Automation for Sustainability, 2010, pp. 138–143.
[6] H. W. H. Premachandra, C. Premachandra, T. Kimura, and H. Kawanaka, “Artificial neural network based sinhala character recognition,” in International Conference on Computer Vision and Graphics, 2016, pp. 594–603.
[7] [Online]. Available: http://192.248.22.122/ocrsinhala/
[8] I. Anuradha, C. Liyanage, H. Wijayawardhana, and R. Weerasinghe, “Deep Learning Based Sinhala Optical Character Recognition (OCR),” in 2020 20th International Conference on Advances in ICT for Emerging Regions (ICTer), 2020, pp. 298–299.
[9] U. Manisha and S. R. Liyanage, “Sinhala Character Recognition using Tesseract OCR,” 2018.
[10] C. Liyanage, T. Nadungodage, and R. Weerasinghe, “Developing a commercial grade Tamil OCR for recognizing font and size independent text,” in 2015 Fifteenth International Conference on Advances in ICT for Emerging Regions (ICTer), 2015, pp. 130–134.
[11] N. Mishra, C. Patvardhan, C. V. Lakshmi, and S. Singh, “Shirorekha chopping integrated tesseract ocr engine for enhanced hindi language recognition,” Int. J. Comput. Appl., vol. 39, no. 6, pp. 19–23, 2012.
[12] M. T. Chowdhury, M. S. Islam, B. H. Bipul, and M. K. Rhaman, “Implementation of an Optical Character Reader (OCR) for Bengali language,” in 2015 International Conference on Data and Software Engineering (ICoDSE), 2015, pp. 126–131.
[13] S. Hussain, A. Niazi, U. Anjum, F. Irfan, and others, “Adapting Tesseract for complex scripts: an example for Urdu Nastalique,” in 2014 11th IAPR International Workshop on Document Analysis Systems, 2014, pp. 191–195.
[14] R. M. Joshi and C. McBride, Handbook of Literacy in Akshara Orthography, vol. 17. Springer, 2019.
[15] J. W. Gair and W. S. Karunatilaka, “Literary Sinhala Inflected Forms: A Synopsis with a Transliteration Guide to Sinhala Script.,” 1976.
[16] Guzmán, F., Chen, P. J., Ott, M., Pino, J., Lample, G., Koehn, P., ... & Ranzato, M. (2019). Two new evaluation datasets for low-resource machine translation: Nepali-english and sinhala-english. arXiv 2019. arXiv preprint arXiv:1902.01382.
[17] [Online]. Available: http://ltrl.ucsc.lk/tools-and-resourses/
[18] [Online]. Available: http://ltrl.ucsc.lk/tools-and-resourses/
[19] V. K. Samaranayake, S. T. Nandasara, J. B. Disanayaka, A. R. Weerasinghe, and H. Wijayawardhana, “An introduction to UNICODE for Sinhala characters,” Univ. Colombo Sch. Comput., 2003.
[20] R. Subasinghe, S. Eramudugolla, S. Samarawickrama and G. Dias, "Atomic Vs Anatomic Features of Sinhala Fonts", 10th Typography Meeting, 2019.

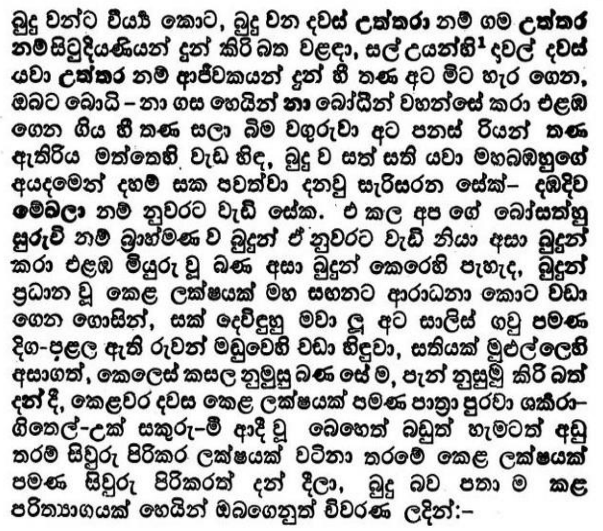
APPENDIX A

Following include the images of three categories used for testing each OCR model.



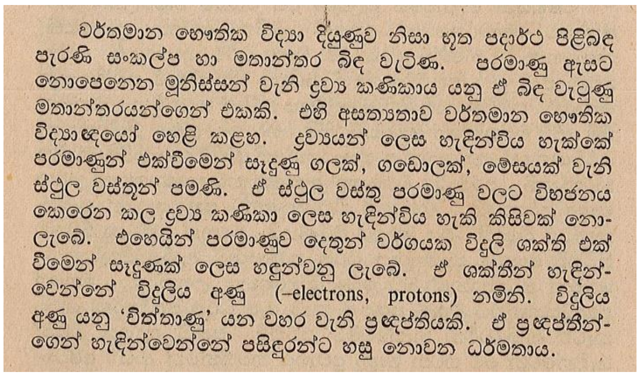
නාගරිකසභා ශාලාවේ හුණුගෑම්මත් පාට ගෑම්මත් තීන්ත ගෑම
ට සහ අලුත්වැඩියාකිරීමටත් කැමති අයවිසින් 'නාගරික සභාශාලාව
අලුත්වැඩියා කිරීමට දෙන ටැන්ඩර්පත්‍රය' (Tender for repairs to
Town Hall) කියා ලිපුම් කවරේපිට ලියා මුද්‍රදරකර නාගරිකසභා,
පති තැනවෙත එවන ටැන්ඩර් හෙවත් ඉල්ලීම්පත්‍ර ලබන මස 9 වෙ
නිවූ සඳුද, පස්වරු 1 වනතුරු මෙම කන්තෝරුවේදී භාරගණු ලැබේ.
වැඩේ කරන්ට ඕනෑ අන්දම සහ විස්තර දැනගන්ට ඕනෑකරුණු
ගැණත් ප්‍රසිඩ වැඩගාර සුපරින්ටැන්ඩන්ට් තැනගෙන් විභාගකර
දැනගන්ට පුළුවන.
ටැන්ඩර් පත්‍ර ඔප්පුකරණ කල සුපරින්ටැන්ඩන්ට් තැනගෙන්
ඉල්ලා ලබාගන්න ෆෝරම් කඩදාසිවල ලියා ඔප්පුකලයුතුයි.

Fig A.1. A sample image of category 1 (old Sinhala newspapers)



බුදු වන්ට වීයෑ කොට, බුදු වන දවස් උත්තරා නම් ගම උත්තර
නම් සිටුදියණියන් දුන් කිරි බත වළඳා, සල් උයන්හි[1] දවල් දවස්
යවා උත්තර නම් ආජීවකයන් දුන් හී තණ අට මිට හැර ගෙන,
ඔබට බොධි–නා ගස හෙයින් නා බෝධීන් වහන්සේ කරා එළඹ
ගෙන ගිය හී තණ සලා බිම වගුරුවා අට පනස් රියන් තණ
ඇතිරිය මත්තෙහි වැඩ හිඳ, බුදු ව සත් සති යවා මහබඹහුගේ
අයදමෙන් දහම් සක පවත්වා දනවු සැරිසරන සේක්– දඹදිව
මෙම්බලා නම් නුවරට වැඩි සේක. එ කල අප ගේ බෝසත්හු
සුරුචි නම් බ්‍රාහ්මණ ව බුදුන් ඒ නුවරට වැඩි නියා අසා බුදුන්
කරා එළඹ මියුරු වූ බණ අසා බුදුන් කෙරෙහි පැහැද, බුදුන්
ප්‍රධාන වූ කෙළ ලක්ෂයක් මහ සඟනට ආරාධනා කොට වඩා
ගෙන ගොසින්, සක් දෙව්දුහු මවා ලූ අට සාලිස් ගවු පමණ
දිග-පුළල ඇති රුවන් මඩුවෙහි වඩා හිඳුවා, සතියක් මුළුල්ලෙහි
අසාගත්, කෙලෙස් කසල නුමුසු බණ සේ ම, පැන් නුසුමු කිරි බත්
දන් දී, කෙළවර දවස කෙළ ලක්ෂයක් පමණ පාත්‍රා පුරවා ශකීරා–
ගිතෙල්-උක් සකුරු-මී ආදී වූ බෙහෙත් බඩුත් හැමටත් අඩු
තරම් සිවුරු පිරිකර ලක්ෂයක් වටිනා තරමේ කෙළ ලක්ෂයක්
පමණ සිවුරු පිරිකරත් දන් දීලා, බුදු බව පතා ම කළ
පරිත්‍යාගයක් හෙයින් ඔබගෙනුත් විවරණ ලදින්:–

Fig A.2. A sample image of category 2 (old Sinhala books)

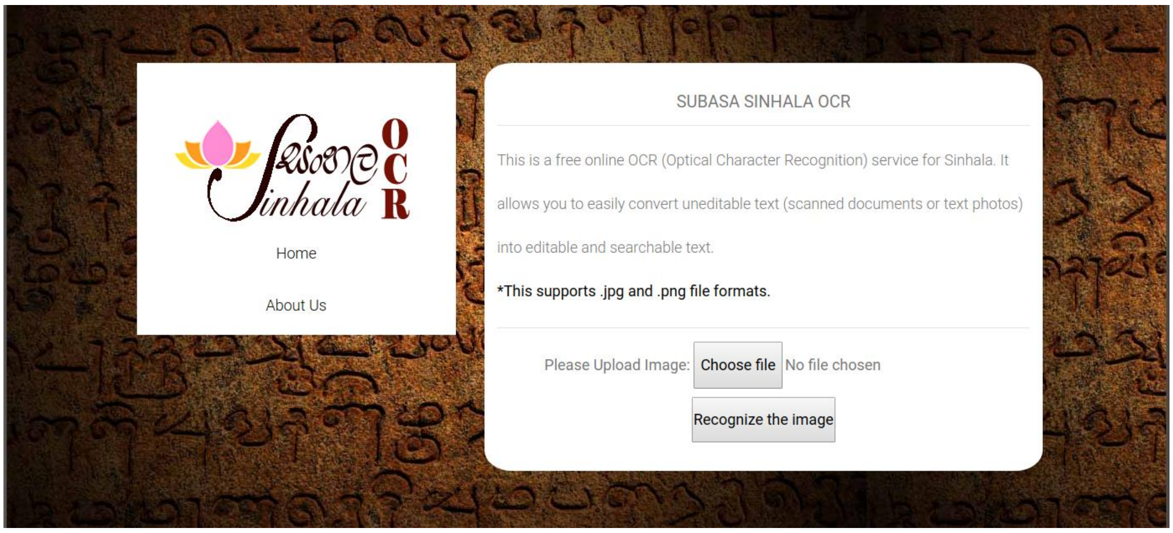


වර්තමාන භෞතික විද්‍යා දියුණුව නිසා භූත පදාර්ථ පිළිබඳ
පැරණි සංකල්ප හා මතාන්තර බිඳ වැටිණ. පරමාණු ඇසට
නොපෙනෙන මූනිස්සන් වැනි ද්‍රව්‍ය කණිකාය යනු ඒ බිඳ වැටුණු
මතාන්තරයන්ගෙන් එකකි. එහි අසත්‍යතාව වර්තමාන භෞතික
විද්‍යාඥයෝ හෙළි කළහ. ද්‍රව්‍යයන් ලෙස හැඳින්විය හැක්කේ
පරමාණූන් එක්වීමෙන් සෑදුණු ගලක්, ගඩොලක්, මේසයක් වැනි
ස්ථූල වස්තූන් පමණි. ඒ ස්ථූල වස්තු පරමාණු වලට විභජනය
කොරෙන කල ද්‍රව්‍ය කණිකා ලෙස හැඳින්විය හැකි කිසිවක් නො-
ලැබේ. එහෙයින් පරමාණුව දෙතුන් වර්ගයක විදුලි ශක්ති එක්
වීමෙන් සෑදුණක් ලෙස හඳුන්වනු ලැබේ. ඒ ශක්තීන් හැඳින්-
වෙන්නේ විදුලිය අණු (–electrons, protons) නමිනි. විදුලිය
අණු යනු 'විත්තාණු' යන වහර වැනි ප්‍රඥප්තියකි. ඒ ප්‍රඥප්තීන්-
ගෙන් හැඳින්වෙන්නේ පසිඳුරන්ට හසු නොවන ධර්මතාය.

Fig A.3. A sample image of category 1 (contemporary Sinhala books)

APPENDIX B

Interface of developed OCR system is shown in Figure B.



SUBASA SINHALA OCR

This is a free online OCR (Optical Character Recognition) service for Sinhala. It allows you to easily convert uneditable text (scanned documents or text photos) into editable and searchable text.

*This supports .jpg and .png file formats.

Please Upload Image: Choose file No file chosen

Recognize the image

Fig B. Online Application developed for the Sinhala OCR