State of Handwriting Recognition of modern Sinhala Script

Chamari M. Silva, N. D. Jayasundere and C. Kariyawasam
Department of Electrical and Information Engineering,
Faculty of Engineering, University of Ruhuna, Sri Lanka.

***Abstract*- Sinhala is a language used by Sinhalese the major ethnic group in Sri Lanka. Most of the textual data gathered in Sri Lanka is in Sinhala language. Converting the data collected on handwritten papers in to digital format enables the automatic data retrieval and also many other advantages including editing, searching, distribution over a network. The manual conversion of the paper documents to electronic format is the available solution which requires an enormous amount of human labour. Using a computer to create an electronic document in Sinhala is extra difficult. If the process of conversion can be automated using handwriting recognition, a significant cost can be saved. Considering the importance of the task, the research effort taken for Sinhala handwriting recognition is not adequate. This paper has examined the current state of handwriting recognition in Sinhala Language and the problem areas that requires attention.**

I. INTRODUCTION

A massive amount of data is collected and stored in day today activities of a human. Some of the example places of data collecting include banks, hospitals, schools, universities, and even places like play grounds and roads. This data comes in a wide variety of forms such as text, image, audio and video. Understanding and making sense of this massive and diverse collection of data requires some automated procedures.

Textual data can be printed or handwritten. Handwriting recognition is the basis of automated data retrieval from handwritten textural data. Handwriting recognition is a one of the challenging research areas coming under pattern recognition in image processing [1]. Further, handwriting recognition can be considered as a development of machines with human intelligence. Therefore handwriting recognition is also a field of research in artificial intelligence and machine vision.

Electronic document processing and managing software are used extensively because of their massive benefits such as the ability of editing, copying, pasting, and saving. The electronic documents provided by these software are in a format which is understandable by computers. Further, when the documents are in electronic format they can be distributed through a network, the content can be searched for keywords regardless of the length of the document, the important contents can be highlighted, the changes can be tracked, comments can be added and many more operations can be carried out. When the documents are on handwritten papers, some of these tasks can be performed with the cost of time and human labour but some operations are impossible.

In Sri Lanka, the main medium of education is Sinhala, even though several higher educational institutions like universities and few schools operate in English medium. Therefore, an average person in Sri Lanka is proficient in Sinhala writing and lack the proficiency in English language and English writing. Consequently the textual data which is gathered in Sri Lanka is in Sinhala language. Consequently the automation process of retrieving textual data from paper documents in Sri Lanka need to be done in Sinhala. Using a computer to create an electronic document in Sinhala is a task which requires a considerable effort even for an expert of word processing tools. That makes the effort of Sinhala handwriting recognition even more appreciated.

II. SINHALA LANGUAGE

Sinhala is the language used by Sinhalese people, the major ethnic group Sri Lanka. Sinhala belongs to the Indo-Aryan language family. It differs considerably from other Indo-Aryan languages due to language contact with the neighbouring Dravidian languages Tamil and Malayalam [2]. Since Sri Lanka is located close to India, the language, as well as the scripts of Sri Lanka are greatly influenced by India. The Ancient

Indian Brahmi script is considered as the beginning of Sinhala writing system [3].

Sinhala characters are circular in shape and straight lines are almost nonexistent in the character set. The reason was the use of palm leaves for scripting. The circular shape was preferred since the dried palm leaves tend to split along the veins in straight line writing. Sinhala letters are written from left to right. Written Sinhala also has the cursive variety, but there's no such notion of block capital letters as it does in English.

Sinhala script is categorized as a segmental writing system and therefore consonant–vowel sequences are identified as a unit. The vowels can be categorized in to two categories: independent and modifier. An independent vowel does not attached to a consonant. A modifier vowel is always attached to a consonant or pure vowel. In the second case the vowel is denoted by one or more strokes positioned around the consonant. Depending on the vowel, the modifier can attach above, below, following or preceding the consonant. A pure vowel is generally used only at the beginning of a word, and has a distinct symbol [4]. Figure 1 shows sample letters for consonant letters and pure vowels and Figure 2 shows sample letters for modifiers and their associations. For the same vowel, the modifier symbol can change according to the consonant used. Moreover there can be more than one modifier applied to a single consonant as indicated in the Figure 3.

Figure 1. (a) consonant letters and (b) pure vowel letters of Sinhala alphabet

Figure 2. Modifier symbols placed before, after, above and below respectively.

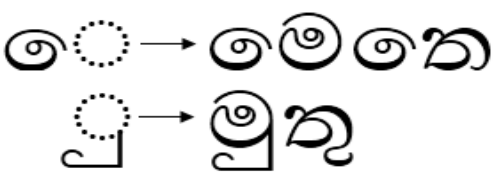

Figure 3. The application of two modifiers for two different letters are shown. The first modifier for vowel 'e' is applied to the two letters without any difference. But the second modifier for vowel 'u' is applied to the same two letters in different symbols.

Figure 4. Application of more than one modifier on a single consonant

At some incidences pairs of consonants may be combined and they are said to form a conjunct letter, e.g., ක්ෂ (ksha). The forming of these conjunct letters is not compulsory, and has become less common in modern use.

III. LITERATURE REVIEW

Handwritten data can be structured or unstructured. Data is retrieved in structured handwriting if a preformatted form is filled to collect data. Example situations include hospital bed head tickets, bank slips, birth certificate or identity card application forms and marriage registration forms. The generation of a structured document is easy as well as the analysis is a lot less complex compared to an unstructured document. Since the set of possible words used is limited for a particular entry in a structured handwriting, it is less complex to identify compared to unstructured handwriting.

Further handwriting can be constrained or unconstrained. Constrained refers to handwriting which performed under a set of predefined writing constraints. As an example, writing all characters so that they are not touching each other. Identification of constrained handwriting is less complex compared to unconstrained handwriting.

In machine reading, usually flat-bed scanners are used. The input of a handwriting recognition system is an image $I(x, y)$ where the intensity I may be scalar, as in gray-scale images, or vectorial, as in RGB color images. Compared to machine vision, human vision does not function like a camera or a scanning device. Cameras and scanning devices are designed to record images with an even resolution over the whole visual field. The eyes, on the contrary, view a limited part of the visual field and are sampling the surrounding world in a series of saccades (eye jumps), and fixations [5]. The human system has the advantage of not

having to compute image transformations over the whole image. In machine reading, there exists a problem of layout analysis and segmentation which is completely different from the problems which are encountered and solved by a human reader [5].

As an average, human reading speed varies from 160-200 words per minute and speed of character recognition can exceed 1500 words per minute. Even though the human is proved to be the better reader, the machine still provides an attractive alternative due to the speed of processing and the massive amounts of data that can be processed [5]. There are only a numbered set of attempts [6] [7] [8] [9] that has been made for Sinhala handwriting recognition and it is still an unsolved problem. In Sinhala writing system, a consonant is indicated by a letter. Vowels are written as modifiers to the consonant letters. The writing system shows major differences from English, which makes the techniques used for English handwriting recognition less applicable for Sinhala handwriting recognition. Further, the large character set, the complexity of characters and the existence of similar character patterns improves the difficulty compared to English.

Hindi Language uses vowels as modifiers in its script Devanagari. In Devanagari script, most of the characters have a horizontal line at the upper part. The letters in Devanagari script occupy the middle zone and upper and lower parts are the modifiers (Figure 5). But in Sinhala script there are letters as well as modifiers which occupy the upper and lower zones (Figure 6). Therefore zoning techniques used to differentiate between characters and modifiers in Devanagari cannot be used for Sinhala. Likewise the unique nature of the characters in each script, the recognition techniques cannot be applied directly from each other and Sinhala needs its own set of techniques in character recognition.

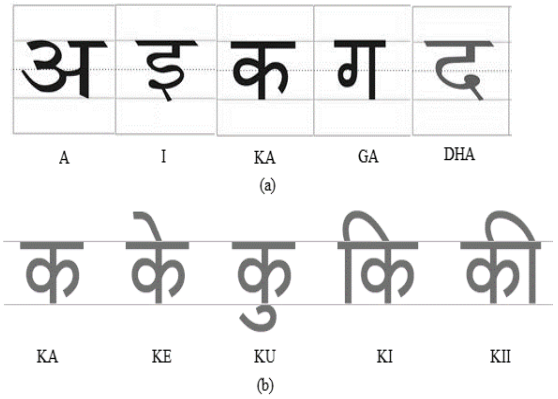

Figure 5. (a) Devanagari consonants *[10]* (b) vowel mark occurring above, below, to the left, and to the right of a Devanagari consonant *[11]*

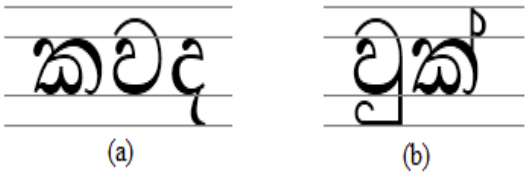

Figure 6. (a) Sinhala letters occupying core zone, upper area and lower area (b) modifiers occupying the upper area and lower area

Recognition becomes more complicated if the process should not depend on the writer. The writing styles of the handwritten characters vary from one writer to another. Figure 7 shows a sample of two characters from 7 different writers.

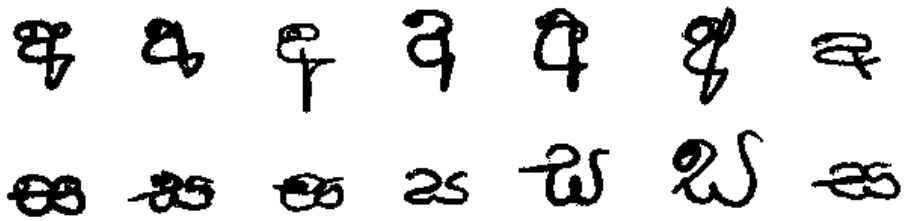

Figure 7. Variations of writing styles present in handwritten characters

Number of distinct characters is a key factor in a character recognition system. All the letters in the Sinhala alphabet are not used in modern Sinhala writing. There are 37 characters and 15 modifier symbols which are used in modern Sinhala writing. Figure 8 shows the 37 letters and 15 modifiers. Out of the 37 characters the first 6 are pure vowels. Among them අ, උ, එ and ඔ are used with character modifiers to generate few other vowels. They are ආ, ඇ, ඈ, ඌ, ඒ, ඓ, ඕ and ඖ. There are some character and modifier combinations which do not occur. ෙළ, ණෘ, ඵෘ, න්ර, ෙර, එං and ල්ර are some examples.

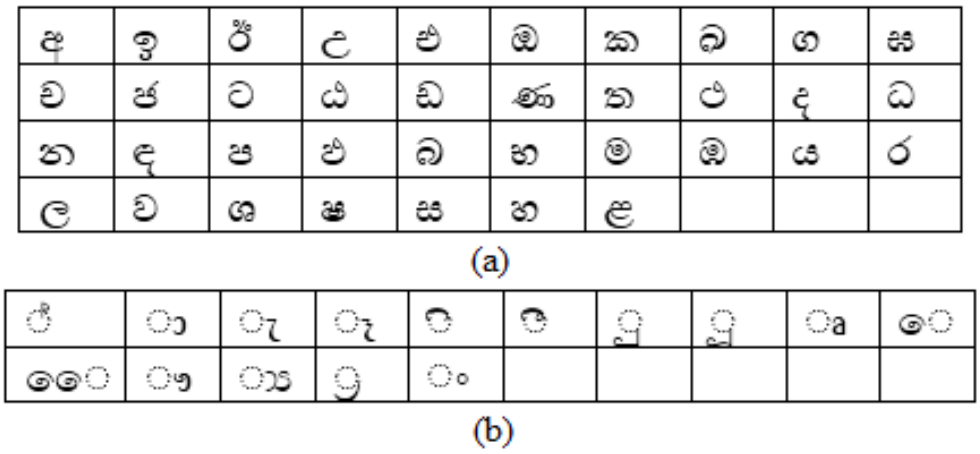

Figure 8. (a) letters used in modern Sinhala script (b) character modifiers used in modern Sinhala script

Character segmentation is the process of decomposing an image with characters in to a set of sub images of individual characters. Given a text line image X, let $P = \{x_1, x_2, \ldots, x_n\}$ be a certain partition of X (n is the number of partitioned units), $R = \{c_1, c_2, \ldots, c_n\}$ be the recognition result of P, the purpose of character segmentation is to find the optimal P to maximize the confidence of R [12].

Character segmentation is a pre-processing step of handwritten character recognition. Segmentation is a crucial step in handwriting recognition [13]. Segmentation becomes harder if there are touching characters present, which is a common scenario for unconstrained handwritten text. Correct segmentation of characters plays a major role in the accuracy of the recognition algorithm.

Among the effort made for Sinhala handwriting recognition, in [6] the system was trained to respond only to 36 of the consonant letters of the Sinhala language. In [7] the modifier symbols also has been considered. But some of the modifier symbols and consonant symbols which are frequently required in writing has been neglected. In [8] the details about the testing dataset, particularly whether it is from a single writer or multiple writers and whether it is constrained or unconstrained is not available. A research in the specific domain of postal city name recognition is available in [9]. Considering the work done, a detailed and complete research in the area of Sinhala handwriting recognition is a gap that needs to be filled.

IV. PROBLEMS

Considering the importance of the task, the research effort taken for Sinhala handwriting recognition is not adequate. The problem areas of Sinhala handwriting recognition that requires attention are as follows.

1. Lack of the knowledge of the best features to represent the characters which need to be tested with a reference for Sinhala character shapes.

   A standard set of characters which can be used as the reference for Sinhala characters need to be tested for best features to represent the characters. The printed Sinhala fonts cannot be taken as a reference to the handwritten characters as there are artistic features available in those fonts (Fig. 9). Testing for reference features from a standard set is another area which has not been addressed yet.

(a) (b) (c)

Figure 9. (a) different stroke sizes present inside a single letter (b) differences of several printed letters (c) a handwritten letter

2. Lack of the recognition of all the consonants and vowels used in modern Sinhala writing.
Identification of all the consonants and vowels used in modern Sinhala writing has not been addressed in existing work.

3. Problem of misclassification of results of nearly similar character shapes.
There are nearly similar character shapes present in Sinhala alphabet and they can lead to misclassification results. Some of the similar shaped characters are given in Figure 10. After classifying in to groups, some special techniques need to be proposed in order to identify the minor changes in the characters.

Figure 10. Similar shaped characters

4. Issue of handling character modifiers and touching characters in character segmentation.
Correct segmentation of characters plays a major role in the accuracy of the recognition algorithm. Further, in Sinhala character modifiers also need to be segmented to reduce the number of pattern classes. Some modifiers are touching the character by its nature. The issue of handling character modifiers in the Sinhala character segmentation is an issue that has not been addressed yet.
Some segmentation free approaches also have been proposed for character recognition [14]. These are called holistic approaches. In the holistic methods, the idea is to recognize words as a whole. This removes the ambiguities found in segmentation, but it improves the complexity of the classification task by increasing number of classes drastically. The holistic approaches are yet to be applied on Sinhala handwriting recognition.

5. Handling of conjunct letters
Most of the conjunct letters are not use in modern Sinhala writing. But still some writers prefer using very limited number of conjunct letters in writing certain words.
Eg. : ක්ෂය. The use is very minimum that we can ignore the problem in recognition of modern script.

6. Lack of a dataset for Sinhala handwritten text aimed at general applications.
The availability of large datasets is important for the evaluation of the recognition accuracy. The accuracy depends upon the database used for the recognition. Covering a large variety of writing styles is necessary for an unbiased evaluation. When researchers use their own collected datasets the results cannot be compared easily. Standard datasets are available for digit recognition [15], and languages like English [16] and Chinese [17]. A Publicly available large dataset for

Sinhala handwritten text aimed at general applications is not yet available. Creation of such a dataset will be beneficial for future researchers.

V. CONCLUSIONS

Sinhala handwriting recognition can be applied in textually gathered data processing in Sinhala. Conversion of paper documents to electronic format in Sinhala is extra difficult. There are many more problems need to be addressed in order to convert the research in to a working application. Considering the work done, research effort taken in the area of Sinhala handwriting recognition is not adequate.

ACKNOWLEDGMENT

The authors would like to appreciate the support given by academic and non-academic staff of the faculty of Engineering, University of Ruhuna, Galle, Sri Lanka in carrying out this research work.

REFERENCES

[1] J. Pradeep, E. Srinivasan and S. Himavathi, "Diagonal based feature extraction for handwritten alphabets recognition system using neural network.," *arXiv preprint arXiv:1103.0365*, 2011.

[2] M. Hilpert, "Auxiliaries in spoken Sinhala," *Functions of Language*, vol. 13, pp. 229-253, 2006.

[3] H.L.Premaratne and J.Bigun, "A segmentation-free approach to recognise printed Sinhala script," in *International Information Technology Conference*,, 7-9 october2002.

[4] J. B. Disanayaka, අකුරු හා පිලි (Letters and Strokes), Godage, 2000.

[5] L. Schomaker, "Reading systems: An introduction to digital document processing," *Digital Document Processing. Springer London*, pp. 1-28, 2007.

[6] R. K. Rajapakse, A. R. Weerasinghe and E. K. Seneviratne, "Neural Network based character recognition system for Sinhala Script," in *Department of Statistics and Computer Science, University of Colombo*, 1995.

[7] Hewavitharana, S. and Kodikara N. D., "A Statistical Approach to Sinhala Handwriting Recognition.," in *Proc. of the International Information Technology Conference (IITC)*, Colombo, Sri Lanka, 2002.

[8] B. Jayasekara and L. Udawatta, "Non-Cursive Sinhala Handwritten Script Recognition: A Genetic Algorithm Based Alphabet Training Approach.".

[9] M. L. M. Karunanayaka, N. D. K. and G. D. S. P. Wimalaratne, "Off Line Sinhala Handwriting Recognition with an Application for Postal City Name Recognition".

[10] N. Sushma, "Hindi Alphabet and Letters Writing Practice Worksheets.," 28 November 2013. [Online]. Available: StudyVillage.com. [Accessed 2015 January 2015].

[11] "INDIC SCRIPTS," [Online]. Available: http://www.linotype.com/5835/indicscripts.html. [Accessed 29 January 2015].

[12] B. B. Chaudhuri, Digital Document Processing: Major Directions and Recent Advances., Springer.

[13] A. Choudhary, R. Rishi and S. Ahlawat, "A New Character Segmentation Approach for Off-Line Cursive Handwritten Words.," *Procedia Computer Science 17* , pp. 88-95., 2013.

[14] S. Madhvanath and V. Govindaraju, "The role of holistic paradigms in handwritten word recognition," *Pattern Analysis and Machine Intelligence, IEEE Transactions*, vol. 23.2, pp. 149-164, 2001.

[15] Y. LeCun and a. C. Cortes, "MNIST handwritten digit database.," AT&T Labs [Online], 2010. [Online]. Available: http://yann. lecun. com/exdb/mnist.

[16] A. e. a. Shivram, "IBM_UB_1: A Dual Mode Unconstrained English Handwriting Dataset," in *Document Analysis and Recognition (ICDAR), 2013 12th International Conference on. IEEE*, 2013.

[17] H. e. a. Zhang, "HCL2000-A large-scale handwritten Chinese character database for handwritten character recognition," in *Document Analysis and Recognition, 2009. ICDAR'09. 10th International Conference on. IEEE*, 2009.