Sinhala Handwritten Character Recognition Using Convolution Neural Networks

W.V.S.K.Wasalthilake*and T.Kartheeswaran

Department of Physical Science, Vavuniya Campus of the University of Jaffna, Sri Lanka
*subodhiwasalthilakeskw@gmail.com

Abstract

Automating handwritten character recognition is still new, as Sri Lanka is the only country that uses Sinhala as the national language from all over the world. The alphabet of the Sinhala language includes 60 characters and they are somewhat complex than the other languages. There are nearly 25-30 researches have been done from 1990 towards Sinhala handwritten character recognition. But there is no accurate handwritten character recognizer for the Sinhala language. Therefore, a model using used the Convolutional Neural Networks to train and classify the Sinhala handwritten characters has been proposed. The training accuracy of the CNN method is 95 % and the testing accuracy is 85.71%. This is the highest accuracy obtained for 55 characters from 1990 when comparing with primitive methods.

Keywords: character recognition, handwritten, Sinhala, convolution neural networks

Introduction

Sinhala is a unique and national language spoken only in Sri Lanka from all over the world. It is used by nearly 16 million of Sinhalese who is one of the major ethnic group of the country and also it used as a second language of another 3 million people who live in Sri Lanka (Languages, 2015). Both the alphabet and Language have changed considerably several times. The modern Sinhala alphabet consists of a total of 60 letters and it can be classified into different sets as shown in Figure 1.

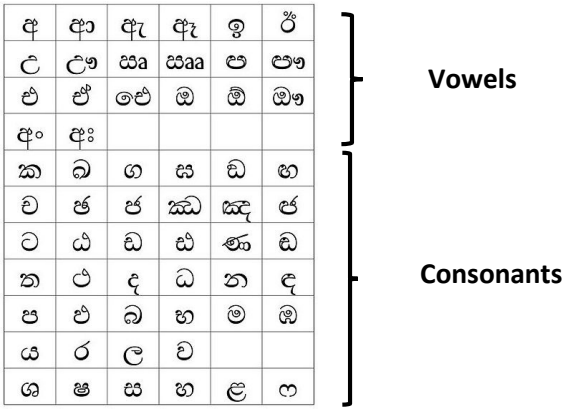


Figure 1: Modern Sinhala Alphabet

Sinhala characters are mostly circular shaped, lack horizontal or vertical lines, and are normally written from left to right in horizontal lines. All the characters in Sinhala scripts have a unique shape and some characters look similar only with noticeable minor differences between them. There are more than 15 modifiers are in the Sinhala language as shown in Table 1. Every single

letter in the alphabet combines with those modifiers and will make more than 1000 letters which can pronounce different ways. Further, there are no changes between simple and block capital letters like the English language.

Table 1: Modifiers

◌ා	◌ැ	◌ෑ	◌ි	◌ී	◌ු	◌ූ
◌ෘ	◌ෲ	◌ෟ	◌ෳ	◌ෙ	◌ේ	

Sinhalese used handwriting since ancient times as a major source of communication. Because of the less computer literacy, less access to devices, and technologies still people use handwritten text to make documentation without using printed documents. There are so many similar looking characters in Sinhala. Therefore, the recognition of individual Sinhala character is a difficult and complex process, due to misclassification of almost similar shape features of a similar group of characters. The high accurate Sinhala handwritten recognition system is still a deficiency in Sri Lanka even though there were some researches carried out on this topic since the 1990s. This research aims to find an efficient method for Sinhala handwritten character recognition with higher accuracy for the most common 55 characters among a total of 60 letters in the Sinhala alphabet.

Background/ Literature Review

Sinhala handwritten characters are extremely inconsistent when comparing to typewritten or printed characters. There are many reasons for those inconsistencies such as the usual shape of the letter, size of the character, speed of writing, and age of the writer, literacy, and personal writing styles, etc. Because of these reasons, handwritten scripts vary from person to person and even sometimes with the same writer. As Sinhala letters are more complex and with few dissimilarities in some letters, handwritten character recognition becomes a more interesting and challenging area of research. There were some researches carried out based on this topic from 1990. They were used many approaches such as Neural Networks(Rohana K. Rajapakse and Seneviratne, 1995), Hidden Markov Model(Hewavitharana *et al.*, 2015), Projection file methods(Nilaweera, Premeratne, and Sonnadara, 2014), part based matching techniques(Silva and Kariyawasam, 2014), Contour tracing method(Silva, Jayasundere and Kariyawasam, 2016) and some were similar as the above methods. All those researchers used their datasets with different image processing techniques which is suitable for their investigations. All those existing researches had many drawbacks such as limited characters, complex algorithms, less accuracy, etc. This research encompasses a set of objectives to overcome those problems, such as;

- Proposing an efficient method with higher accuracy than the current systems for 55 characters among a total of 60.
- Examine the possibility of using image processing and related techniques for character recognition.
- Identifying the different handwritten scripts with different age groups.
- Preparing dataset without age limitations (samples with 13 to 75 age limit).

Methodology

Data Collection

The process of data collection included some important steps as shown in Figure 2.

Figure 2: Data collection

Labelling characters

Characters had been labelled according to the order of English alphabetical characters such as a1, a2, a3, b1, b2, and so on as shown in Table 2.

Table 2: Sample for Character labeling

Letter	අ	ඇ	ඈ	ඉ	ඊ	උ	ඌ	එ	ඒ	ඓ	ඔ	ඕ
Label	a1	a2	a3	a4	b1	b2	b3	c1	c2	d1	d2	d3

Method

CNN is a deep learning technique which needs less pre-processing than the primitive methods. The process included some basic steps as shown in Figure 3.

Figure 3: CNN Method

The performance of CNN is based on the values of parameters. Therefore, we have created the model with optimum parameter values by tunning and testing the model several times.

Usually, the pooling layers are applying after the Convolution layer. But in this research, we followed a trial and error method and found that when inserting a pooling layer after the input layer, accuracy changes by 4%. Therefore, we used a max-pooling layer after the input layer. The final CNN architecture is shown in Figure 4.

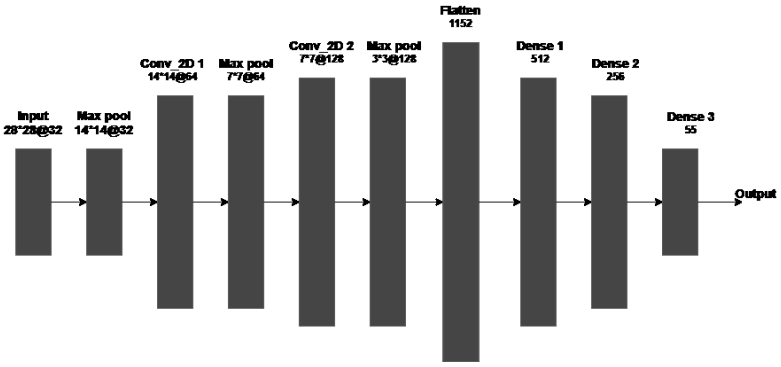


Figure 4: The CNN model with tuned parameters

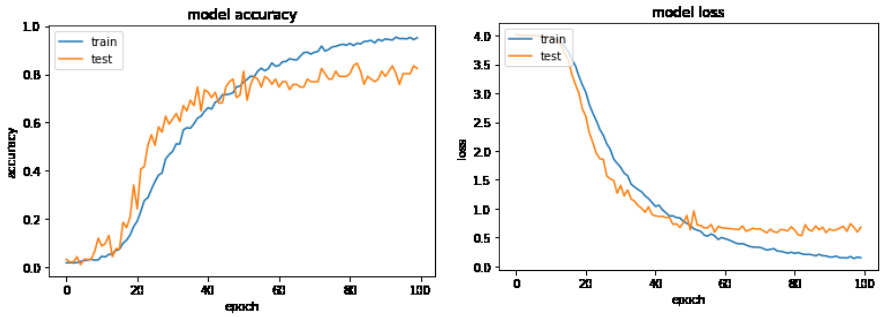
Results and Discussion

Refinement

The performance of a Convolutional Neural Network based on many factors such as number of layers of the network, number of neurons in each layer, Amount of data that are used to train and test the model image characters, tuning parameters such as optimizers, regularizers, learning rate, number of epochs etc. The following are some of the values used for the final model is tested with maximum accuracy for 95 percent of training and 5% of testing after resizing the images into 28x28 with 0.01 learning rate, SGD optimizer, and 80 epochs. However, the data augmentation, cross validation,and some more fine tuning parameterswere applied but still, the changes in the accuracy were not significant.

Results

Accuracy values are changed from time to time when tuning parameters. Finally, we have obtained 95 % training accuracy and 85.71% testing accuracy. The testing curves were given in Figure 5(a and b).



(a) (b)

Figure 5: (a) Model accuracy, (b) Model loss

Confusion Matrix is used to check the performance of the neural network classification as shown in Figure 6.

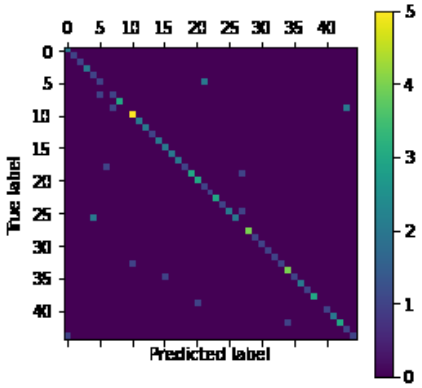


Figure 6: Confusion matrix for the model

Conclusion

Even though nearly 25 researches have been carried out on the same topic, many of them didn't get the expected results and they had many limitations in the past. The main aim of this research is to overcome those limitations. Finally, we have achieved 95% training accuracy and 85.71% testing accuracy for 55 characters among a total of 60 characters by using a small dataset of 1815 characters and with the least effort in pre-processing. It's very important to mention that the only research which is carried out using CNN for ancient Sinhala inscription letters in 2017 (Karunarathne *et al.*, 2017); achieved 92% accuracy for pre-processed images and 64 % for 9 images without pre-pressing while we are obtained 85.71% testing accuracy for a larger set as 55 real character images of modern Sinhala alphabet without using any pre-processing techniques.

References

Hewavitharana, S. *et al.* (2015) 'Off-line Sinhala Handwriting Recognition using Hidden Markov Models Sri Lanka Institute of Information Technology, Colombo, Sri Lanka', (November).

Karunanayaka, M., Kodikara, N. and Wimalaratne, G. (2004) 'Off-Line Sinhala Handwriting Recognition with an Application for Postal City Name Recognition', *Icter.Org*, (35). Available at: http://www.icter.org/conference/sites/default/files/icter/IITC-2004p4.pdf.

Karunarathne, K. G. N. D. *et al.* (2017) 'Recognizing ancient Sinhala Inscription Characters using Neural Network Technologies', (1).

Languages, 17 Minute (no date) *The scripts of the world: The Sinhalese alphabet*. Available at: https://www.17-minute-languages.com/en/blog/learn-more-about-the-sinhalese-script/.

Nilaweera, N. P. T. I., Premeratne, H. L. and Sonnadara, D. U. J. (2014) 'Comparison of Projection and Wavelet Based Techniques In Recognition of Sinhala Handwritten Scripts', (September 2007). Available at:https://s3.amazonaws.com/academia.edu.documents/30429910/2007_CSSL_wavelet.pdf

Rohana K. Rajapakse, a. R. W., and Seneviratne, E. K. (1995) 'a Neural Network Based Character Recognition System for Sinhala Script', *Vasa*, (January 1995). Available at: http://medcontent.metapress.com/index/A65RM03P4874243N.pdf.

Silva, C. and Jayasundere, N. (2018) 'Isolated Sinhala handwritten character recognition using part Based Matching Techniques', (April).

Silva, C. and Kariyawasam, C. (2014) 'Segmenting Sinhala Handwritten Characters', *International Journal of Conceptions on Computing and Information Technology*, 2(4), pp. 2345–9808. Available at: https://www.researchgate.net/profile/Chamari_Silva/publication/270280869_Segmenting_Sinhala_Handwritten_Characters/links/54a65ec10cf256bf8bb4f232.pdf.

Silva, C. M., Jayasundere, N. D. and Kariyawasam, C. (2016) 'Contour tracing for isolated Sinhala handwritten character recognition', *15th International Conference on Advances in ICT for Emerging Regions, ICTer 2015 - Conference Proceedings*. IEEE, pp. 25–31. doi: 10.1109/ICTER.2015.7377662.