International Journal of Advanced Computational Engineering and Networking, ISSN(p): 2320-2106, ISSN(e): 2321-2063
Volume-5, Issue-12, Dec.-2017, http://iraj.in

LEECH STEP PATH FINDING ALGORITHM

[1]DAWPADEE B. KIRIELLA, [2]LAKSHMAN JAYARATNE

[1,2]University of Colombo School of Computing, 35, Reid Avenue, Colombo 07, Sri Lanka
E-mail: [1]kiriella.dawpadee@gmail.com, [2]klj@ucsc.cmb.ac.lk

Abstract - Sinhala characters inherit a variety of shapes especially different types of curves by nature. Therefore Sinhala character recognition cannot reach the maximum accuracy with the existing techniques and algorithms used in Sinhala character feature extraction within current intelligent character recognition (ICR), optical character recognition (OCR) and optical music recognition (OMR) engines. Existing OCR engines even with their adaptive classifier have not been able to train for Sinhala characters as it gets maximum accuracy with the output in the process of character recognition. In order to solve this problem, we propose a novel algorithm, named as Leech Step Path Finding Algorithm (LSPFA) which can be mainly used to extract/ recognize character features (of segmented/ isolated characters) in the adaptive classifier of ICR, OCR and OMR engines. Even though our study principally targets Sinhala characters, LSPFA can be further used for any other set of characters/ symbols of the alphabet of any other language.

Keywords - Character Feature Extraction; Sinhala Characters; Intelligent Character Recognition; ICR; Optical Character Recognition; OCR; Optical Music Recognition; OMR; Adaptive Classifier

I. INTRODUCTION

Optical character recognition or optical character reader (OCR) provides a full alphanumeric recognition of natural language characters on digital scripts [1]. Intelligent character recognition (ICR) is an enhanced version of OCR and its main significance is catering not only for printed characters but also handwritten characters [2]. Optical music recognition (OMR) engines are used in machine reading of sheet music (scripts with music notations) and converting them into the expected machine readable format [3]. Here, the set of music notations consists of natural language characters as well as a pre-defined set of symbols. In all the above character recognition engines, the overall accuracy of the character recognition process highly depends on the accuracy of their technique/ algorithm of character feature extraction/ recognition. Each character of a particular natural language denotes its unique feature/s by nature. Only if the computer machine can recognize this feature/s accurately, the relevant character is exactly recognized. If an algorithm can be used to accurately recognize and extract the character features of the set of characters from the alphabet of a particular language and through that, the character recognition is involved in both printed and handwritten characters, that algorithm can be used in ICR engines in the context of that particular language. Moreover, it can be used in the process of character/ symbol recognition in OCR and OMR engines used for the same language. It may need some modification according to the application which uses the OCR or OMR. But, for the same context (language), the core idea of the algorithm does not require to be changed.

1.1. The Problem

Though a number of research studies as well as commercial products exist for English and some other languages to recognize its characters/ character features with higher accuracy level [1], no ICR engine for Sinhala characters has been invented with maximum accuracy yet. The research studies on OCR for Sinhala printed characters [4] and off-line Sinhala handwriting recognition [5] which carried out up to now are highly appreciated. But they also are not with maximum accuracy with 100%.

The majority of Sri Lankans (nearly 80%) use Sinhala characters in various applications (i.e. academics, industrial, and normal day-to-day activities). Therefore, there is a critical and higher requirement of solving the above problem. In order to come up with an accurate ICR for Sinhala characters, inventing an algorithm which can be used to precisely recognize Sinhala character features is vital.

Comparing with the characters of other languages; especially English alphabet (Because many studies of the area have been carried out in the context of English. Further, the alphabetical characters of English are used in some other languages too), Sinhala characters have more nonlinear curving shapes which make the character recognition process more difficult. Because of the above-mentioned fact, the technique/ algorithm used in adaptive OCRs (in its adaptive classifier) such as Tesseract OCR engine [6] cannot be used in Sinhala context (for Sinhala characters) with 100% accuracy.

1.2. Our Contributions

Following are the main contributions of our study.

- Proposing a novel algorithm (LSPFA) which can be used to recognize character featuresnot only in OCR engines but also in OMR and ICR engines to solve the identified problem
- Analysis of the feasibility of implementing such a solution, its effectiveness and the challenges in deploying the solution within the target applications

- A survey to identify the most critical problems which are in the ICR process for Sinhala characters
- Literature survey on related available studies and commercially available products in order to address the problem in other contexts (other natural languages in ICR, OCR and set of music notations in OMR)

II. THE LEECH STEP PATH FINDING ALGORITHM

Ideally when a digital script with characters is input to an ICR engine the following process is carried out in the system with three main phases as follows.

- Phase 1: Preprocessing and segmentation
- Phase 2: Character feature extraction and recognition
- Phase 3: Presenting the output in required format (the output format may vary according to the application software which uses the ICR engine)

We propose the algorithm called LSPFA for the above-mentioned phase 2. Since this is mainly focused on Sinhala characters and its feature recognition, it is necessary to know the main features of Sinhala characters.

Mainly, three different types of features are recognized in Sinhala characters. They are;

- Straight lines
- Curves
- Rapid turns

Fig. 1. depicts a character; "ඔ" of Sinhala alphabet which consists of all the above features within one character. Therefore, this is used to demonstrate the algorithm in detail afterwards.

Fig.1. Sinhala character: ඔ

2.1. Main steps of the algorithm

- Step1: Get Character segment and decide mask radius (Fig. 2.)

Fig.2. Step 1.

- Step 2: Start from the top most left grayed pixel (Fig. 3.)

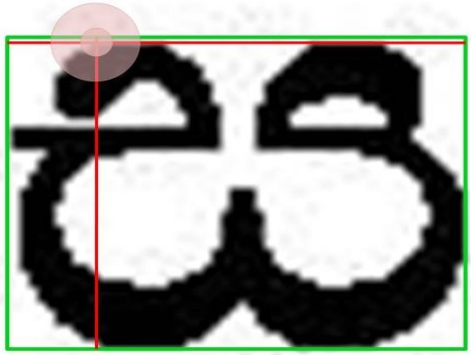

Fig.3. Step 2.

- Step 3: Consider the clock-wise immediate path (Fig. 4.)

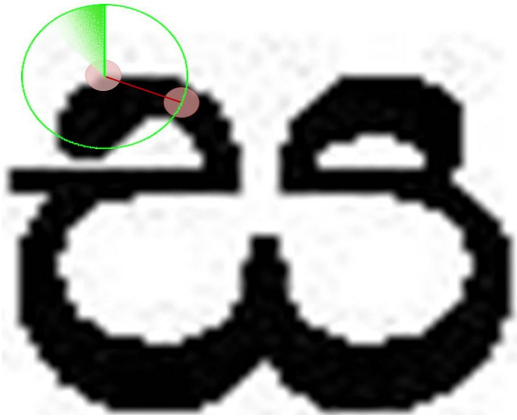

Fig.4. Step 3.

- Step 4: Find the next node (Fig. 5.)

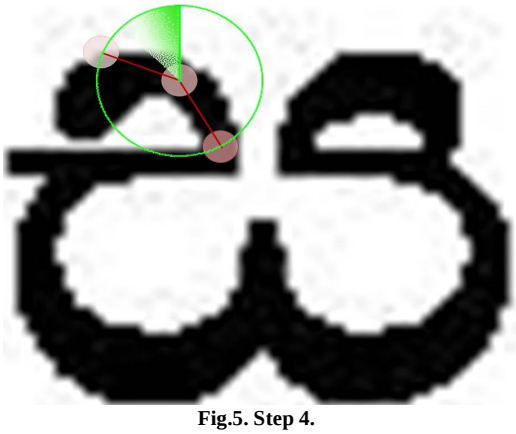

Fig.5. Step 4.

- Step 5: Find the next node by ignoring the previous node (Fig. 6.)

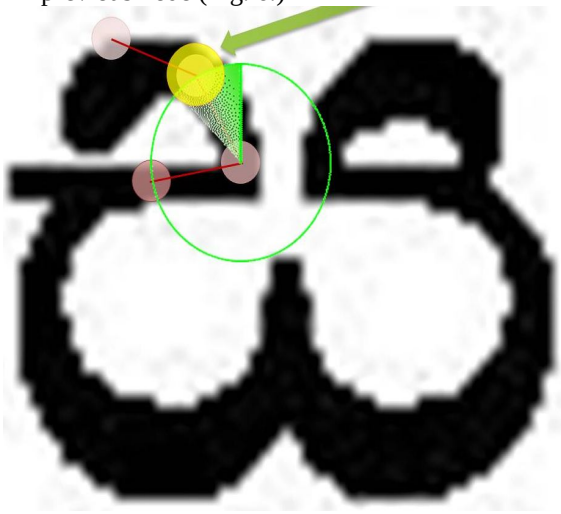

Fig.6. Step 5.

- Step 6: Consider the path with continuous gray values between center node and candidate node (Fig. 7.)

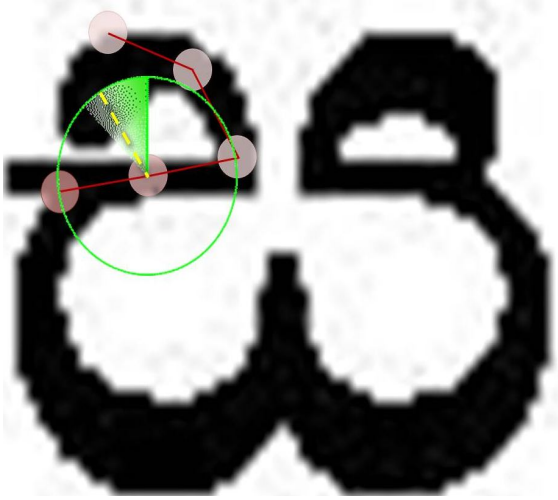

Fig.7. Step 6.

- Step 7: Clone bot in junctions (Fig. 8.)

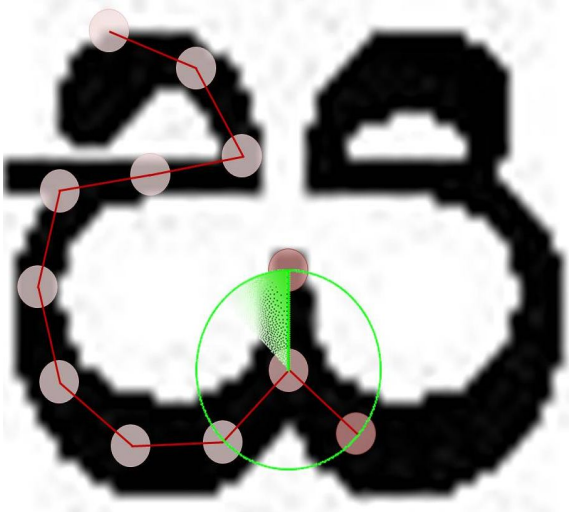

Fig.8. Step 7.

- Step 8: Stop if continuous path not found (Fig. 9.)

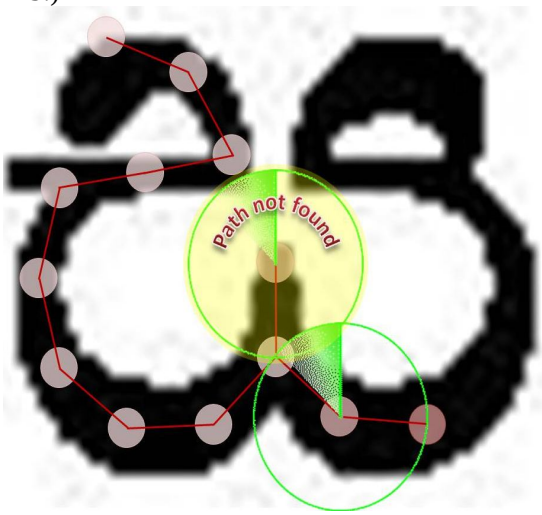

Fig.9. Step 8.

- Step 9: Stop on face to face meeting, if search zone overlap percent exceed therecommended threshold (Fig. 10.)

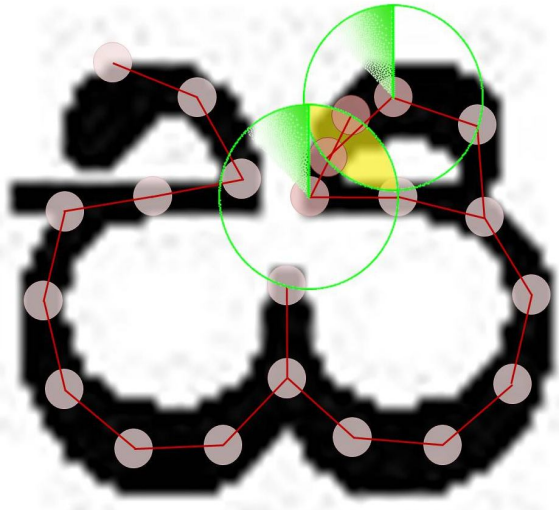

Fig.10. Step 9.

2.2. Possible Preliminary Outputs of the algorithm

- Detected straight lines of the character (Fig. 11.)

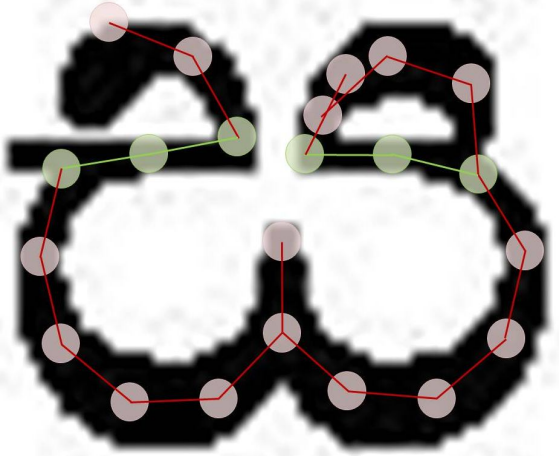

Fig. 11. Detected straight lines (depicted in green colour)

- Detected curves of the character (Fig. 12.)

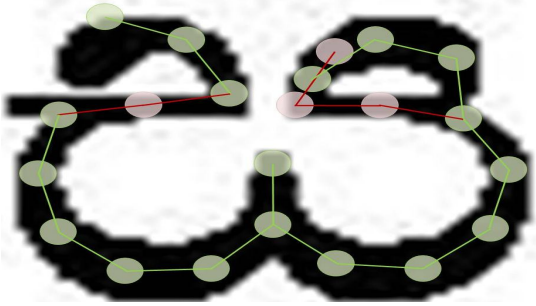

Fig. 22. Detected curves (depicted in green colour)

- Detected rapid turns of the character (Fig. 13.)

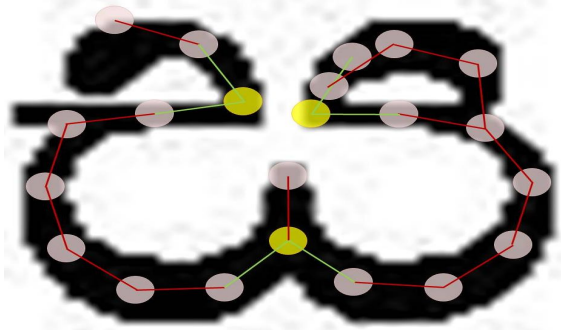

Fig. 33. Detected rapid turns (depicted in yellow colour)

III. THE ROLE OF THE PROPOSED ALGORITHM IN INTELLIGENT CHARACTER RECOGNITION PROCESS

Fig. 14. depicts the typical architecture of an ICR processing system. At the phase of segmented/isolated character recognition, all the atomic symbols/characters should be classified using an adaptive classifier rather than a static classifier.
It has been demonstrated [7] that OCR engines can benefit from the use of an adaptive classifier. Since the static classifier has to be good at generalizing to any type of character, its ability to discriminate between different characters is weakened. A more character-sensitive adaptive classifier that is trained by the output of the static classifier is therefore commonly [8] used to obtain greater discrimination within each document, where the number of character types is limited.

LSPFA is there to make the adaptive classifier more accurately with exact correct recognition of the characters.

IV. RELATED WORK

Following statement which is under one, among the three principal issues that need to be considered in the character feature recognition and extraction approach pointed out by Shridar and Kimura [1] proves the importance of accurate character recognition in OCR and ICR. 'Erroneous recognition of characters extracted from the word image can lead to incorrect word recognition [1]'. According to Trier et al. [9], selection of a feature extraction method is probably the single most important factor in achieving high recognition performance in character recognition systems.

By moving more towards ICR approaches, the following studies on Sinhala handwritten character recognition can be taken into account. Hewawitharana et al. [10] presented an off-line Sinhala handwriting recognition using Hidden Markov Models which could be used in ICR for Sinhala and they saidthat the best accuracy of their algorithm is 92.1%. Rajapakse et al. [11] proposed a neural network based character recognition system and their training algorithm is stated below.

"The objective of training the network is to adjust the weights so that the application of a set of inputs (input vectors) produces the desired outputs (output vectors). Training a back propagation network involves each input vector being paired with a target vector representing the desired output; together they are called a training pair….. Before starting the training process, all of the weights are initialized to small random numbers. Training the back propagation network requires the following steps:

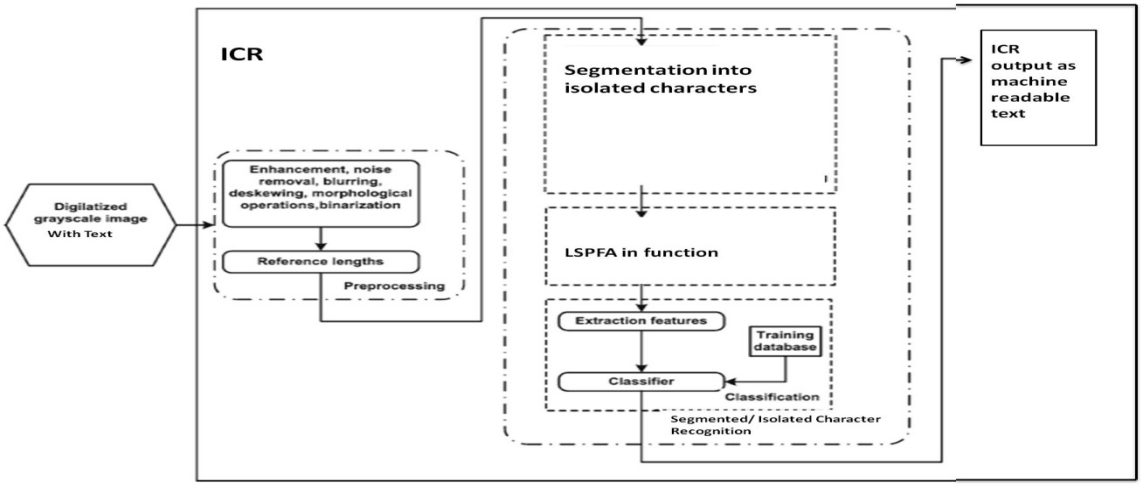

Fig. 44. The Typical architecture of ICR

- **Step 1:** Select a training pair (next pair) from the training data set and apply the input vector to the network input
- **Step 2**: Calculate the output of the network, i.e. to each neuron NET=ΣXiWi must be calculated and then the activation function must be applied on the result (F(NET))
- **Step 3:** Calculate the error between the network output and the desired output (TARGET – OUT)
- **Step 4:** Adjust the weights of the network in a way that minimizes the ERROR
- **Step 5:** Repeat step 1 through 4 for each vector in the training set until no training pair produces an ERROR larger than a pre-decided acceptance level." [11]

Pujari et al. [12] have presented an enhanced dynamic neural network (DNN) model in their publication: an adaptive character recognizer for Telugu scripts using multiresolution analysis and associative memory. They [13] further have carried out their research work towards an intelligent character recognizer for Telugu scripts using multiresolution analysis and associative memory.

In their future work, Premaratne et al. [4] stated:"Since the Sinhala language and the script have been evolved more independently in the island of Sri Lanka, without a close connection to other South Asian languages in the Indian sub-continent, some of its characteristic features need to be handled distinctly."[4]. Further, they have suggested in their future work to enhance their proposed OCR which can coverall the features of Sinhala characters in character feature extraction/ recognition.

In the research paper of [14], Deodhare et al. have discussed the form image registration technique and the image masking and image improvement techniques implemented in their system as part of the character image extraction process. According to them, those techniques support in preparing the input character image (with English characters) for the neural network-based classifiers and go a long way in improving overall system accuracy. Further, they say that "Although these algorithms have been discussed with reference to our ICR system they are generic in their applicability and may find use in other scenarios as well." [14]. Considering our scenario, we cannot use their feature extraction/ recognition technique as it is since it is not strong enough to recognize Sinhala characters which are with a variety of shapes(especially curves and rapid turns) comparing English.

Ranasinghe et al. [15] have presented an experimental setup which proposes an initial OMR engine which uses Sinhala character recognition for their study on adaptive music score trainer for visually impaired in Sri Lanka. In the context of Sinhala characters, the OMR of [16] works for the characters"ඝ, ඊ, ග, ම, ඵ, ධ, ණි". Further, they have enhanced thier work: Swarālōka [17] with a novel character recognition approach named as "Profile-based approach" which is more suitable for Sinhala characters. The scope of above studies covers only the seven Sinhala characters used in eastern music scripts.

CONCLUSIONS AND FURTHER WORK

In our past research work, we have presented the idea of using an OMR with any type of character as scripting symbols [18]. There, we took our first step towards inventing an adaptive classifier. As aresult of researching for an adaptive OMR for eastern music, we needed to study more on accurate recognition of Sinhala characters. There, we have come up with the LSPFA which has been presented in this paper.

As the authors, we believe that the output of our study will be a strong contribution to the research area of character recognition not only for Sinhala characters but also for characters in any other natural language since in our algorithm we have covered all the possible character features which can exist with any type of character.

Further, not only the LSPFA can be used in ICR, OCR or OMR engines but also the applications such as singing synthesizers [19] can utilize it in order to enhance their overall performance.

Moreover, as future work, we will try to enhance the LSPFA towards which can facilitate in a mobile platform. It also will be more important in free and open ICRs as well as will enhance the commercial value of the application which uses the LSPFA.

ACKNOWLEDGMENTS

The research is partially granted by the National Centre for Advanced Studies in Humanities and Social Sciences (NCAS), Sri Lanka. The authors would like to thank all the lecturers of the University of Colombo School of Computing (UCSC), University of Colombo, Sri Lanka who have kindly supported throughout this research work which is carried out at the UCSC. Kavindu Ranasinghe; senior software engineer is gratefully acknowledged for his helpful comments on the content of this paper.

REFERENCES

[1] H. Bunke and P. S. P. Wang, editors. Handbook of character recognition and document image analysis. World scientific; 1997 May 2.

[2] G. A. Baraghimian and G. E. Host, inventors; Hughes Aircraft Company, assignee. Apparatus and method of fusing the outputs of multiple intelligent character recognition (ICR) systems to reduce error rate. United States patent US 5,970,171. 1999 Oct 19.

[3] Rebelo, I. Fujinaga, F. Paszkiewicz, A. R. Marcal, C. Guedes and J. S. Cardoso, Optical music recognition: state-of-the-art and open issues. International Journal of Multimedia Information Retrieval. 2012 Oct 1;1(3):173-90.

[4] H. L. Premaratne and J. Bigun, Recognition of printed Sinhala characters using linear symmetry. InThe 5th Asian Conference on Computer Vision 2002 Jan (pp. 23-25).

[5] S. Hewavitharana, H. C. Fernando and N. D. Kodikara, Off-Line Sinhala Handwriting Recognition Using Hidden Markov Models. InICVGIP 2002.

[6] R. Smith, "An overview of the Tesseract OCR engine." (2007).

[7] G. Nagy and Y. Xu, "Automatic Prototype Extraction for Adaptive OCR", Proc. of the 4th Int. Conf. on Document Analysis and Recognition, IEEE, Aug 1997, pp 278-282.

[8] Marosi, "Industrial OCR approaches: architecture, algorithms and adaptation techniques", Document Recognition and Retrieval XIV, SPIE Jan 2007, 6500-01.

[9] Ø. D. Trier and A. K. Jain, Taxt T. Feature extraction methods for character recognition-a survey. Pattern recognition. 1996 Apr 1;29(4):641-62.

[10] S. Hewavitharana, H. C. Fernando and N. D. Kodikara, Off-Line Sinhala Handwriting Recognition Using Hidden Markov Models. InICVGIP 2002.

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

[12] K. Pujari, C. D. Naidu and B. C. Jinaga, An adaptive character recogniser for Telugu scripts using multiresolution analysis and associative memory. In3rd Indian Conference on Computer Vision, Graphics and Image Processing ICVGIP 2002.

[13] K. Pujari, C. D. Naidu, M. S. Rao and B. C. Jinaga, An intelligent character recognizer for Telugu scripts using multiresolution analysis and associative memory. Image and Vision Computing. 2004 Dec 1;22(14):1221-7.

[14] D. Deodhare, N. R. Suri and R. Amit, Preprocessing and Image Enhancement Algorithms for a Form-based Intelligent Character Recognition System. IJCSA. 2005;2(2):131-44.

[15] Ranasinghe, S. Kumari, D. Kiriella and L. Jayaratne, Adaptive music score trainer for visually impaired in Sri Lanka. In24th Australasian Conference on Information Systems (ACIS) 2013 (pp. 1-11). RMIT University.

[16] C. Ranasinghe, S. C. Kumari, D. B. Kiriella and L. Jayaratne, "Computational approach to train on music notations for visually impaired in Sri Lanka," Proc. Seventh Annual International Conference on Computer Games, Multimedia and Allied Technology (CGAT 2014), Singapore, 2014, p. 34-43.

[17] D. B. Kiriella, K. C. Ranasinghe, S. C. Kumari and K. L. Jayaratne, "Music Training Interface for Visually Impaired through a Novel Approach to Optical Music Recognition." Journal on Computing (JoC) 3.4 (2014)

[18] D. B. Kiriella, K. C. Ranasinghe and L. Jayaratne, "Music training interface for visually impaired with a universally applicable OMR engine," Proceedings of 8th International Research Conference of KDU (KDU-IRC 8), The General Sir John Kotelawala Defence University, Sri Lanka, August 2015, pp. 121-127.

[19] Ranasinghe and L. Jayaratne, "Analysis on Using Synthesized Singing Techniques in Assistive Interfaces for Visually Impaired to Study Music," GSTF International Journal on Computing (JoC), vol. 4, no. 4, pp. 37-45, April 2016.

★★★