Vol. 5 No. 3 January 2018 ISSN: 2321-788X UGC Approval No: 43960 Impact Factor: 2.114

TAMIL CHARACTER RECOGNITION USING AI BASED ENGINE

Mrs. L. SURIYA KALA, MCA., M.Phil.,
Research Scholar, Mother Teresa Women's University
Kodaikanal, Tamil Nadu, India

Abstract

The present archive scanners for the PC accompany the product that plays out an undertaking of character recognition. Optical Character Recognition (OCR) is a sort of report picture investigation where examined advanced picture that contains either machine printed or manually written content contribution to an OCR programming engine and making an interpretation of it into an editable machine decipherable computerized content organization. One test for perceiving Tamil is that there are a great many characters for a framework to perceive and numerous remarkable characters just seem a few times in preparing. In this paper proposed to perceive Tamil character recognition utilizing Artificial Intelligence utilizing Back Propagation Algorithm.

Keywords: *Optical Character Recognition, Artificial Intelligence, Handwriting Recognition, back Propagation Algorithm, Artificial Neural Network.*

1. Introduction

Presently a day there are numerous new philosophies required for the expanding needs in recently developing regions, with this approach there are numerous systems are available for the character recognition of impression Devanagari, Bengali, Tamil, China and so on. Be that as it may, next to no exploration is for printed material. During the most recent four decades, the field of character recognition has been accepting noteworthy consideration, from investigate laborers in various trains, for example, change of transcribed of printed report to an editable delicate configuration, recognition of postal locations for computerized postal framework, information and word preparing, information obtaining in bank checks, handling of documented institutional records. A large portion of the work done in the field of character recognition is restricted to Roman, English, Urdu, Chinese/Japanese dialects. A large portion of the character recognition systems is issue arranged. Systems are contrived for the recognition of a specific content contingent on the nature and multifaceted nature of the character. Extensively, the highlights can be physical, topological, numerical or measurable in nature. This methodology utilized for recognition can be extensively grouped into auxiliary, factual and half breed Structural procedures utilize some subjective estimations as highlights. Factual systems utilize some quantitative estimation. In the crossbreed approach, these two procedures are joined at proper stage first portrayal of characters and using them for recognition.

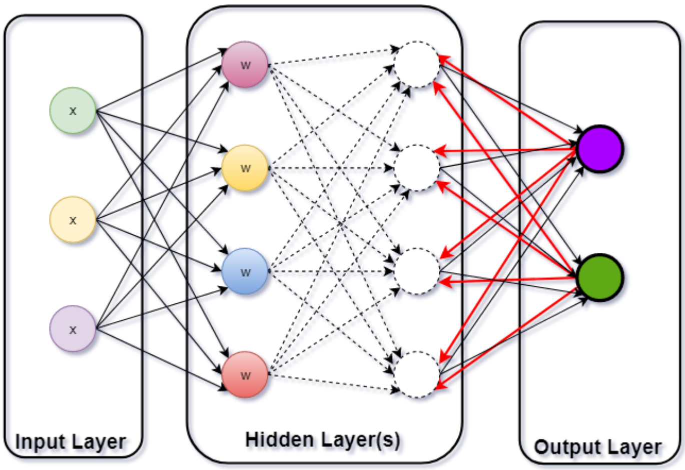

Figure 1 Proposed Algorithm Back Propagation

The issue of perceiving handwriting, recorded with a digitizer, as a period arrangement of pen facilitates is known as on-line character recognition. But it can't be applied to archives printed or composed on papers. Character recognition is a sub-field of example recognition in which pictures of characters from a book picture are perceived and because of recognition separate character codes are returned. Optical Character Recognition (OCR) is a very well-examined issue in the immense region of example recognition. Its beginnings can be found as right on time as 1870 when a picture transmission framework was concocted which utilized a variety of photocells to perceive designs. Until the center of the twentieth century, OCR was basically created as a guide to the outwardly disabled. The character recognition programming breaks the picture into sub-pictures, each containing a solitary character. The sub-pictures are then deciphered from a picture position into a paired configuration, where every 0 and 1 speaks to an individual pixel of the sub picture. The parallel information is then encouraged into a neural network that has been prepared to make the relationship between the character picture information and a numeric worth that compares to the character. The yield from the neural network is then converted into an ASCII message and spared as a document. Recognition of characters is an exceptionally mind-boggling issue. Optical Character Recognition (OCR) manages machine recognition of characters present in an info picture acquired utilizing filtering activity. It alludes to the procedure by which checked pictures are electronically handled and changed over to editable content. An Artificial Neural Network as the backend to take care of the recognition issue. Neural Network utilized for preparing of neural network. Neural networks have been utilized in a large variety of territories to undertake a large scope of issues. Dissimilar to human minds that can recognize and remember the characters like letters or digits; PCs treat them as paired illustrations. In the wake of preparing the network with the back-propagation learning algorithm, high recognition exactness can be accomplished. Recognition of printed characters is itself a difficult issue since there is a variety of a similar character because of progress of text styles or presentation of various sorts of clamors. The distinction in textual style and sizes makes recognition task troublesome if pre-handling, include extraction and recognition are not strong.

2. Literature Survey

[1] Kishna, N. P. T., & Francis, S. (2017) proposed on the recognition of transcribed Malayalam (a South Indian Language) characters. Along these lines, cursive Malayalam characters can be supposed by the Hidden Markov Model (HMM). The proposed strategy utilized the recognition of cursive transcribed Malayalam characters utilizing the Hidden Markov Model (HMM). The calculation utilized here stays away from mistakes brought about by commotion in the examined picture by applying a middle channel. Additionally, Artificial Neural Network (ANN) obtains better characterization and gives the best coordinating class for input. The examples utilized are of elevated ability to decrease the multifaceted nature in the recognition procedure.

Merits:

- Written by hand character recognition with high exactness and productive technique to perceive the cursive letters are remembered for the proposed framework.

Demerits:

- HMMs regularly have an enormous number of unstructured parameters.
- First request HMMs are constrained by their first-request Markov property.

[2] Prameela, N., Anjusha, P., & Karthik, R. (2017) proposes an OCR framework for Telugu archives the character pictures to get the component vector esteems put 3*3 lattices for each character and assess relating centroid for all the nine zones. Along these lines draw the flat and vertical projection heavenly attendant to the closest pixel of the picture. Thus further these resultant qualities are considered as the key component vector for the proposed recognition framework. The proposed framework is shape and textual style subordinate and requires pre-handling and highlight extraction. It might be seen that the two states of characters look like each other with contrast in the locale at the base and at the top.

Merits

- The proposed strategy is used for both support vector machine (SVM) and Quadratic discriminate Classifier (QDA) has been autonomously used as the classifier.

Demerits

- The fundamental weakness of the SVM calculation is that it has a few key parameters that should be set effectively to accomplish the best grouping outcomes for some random issue.

[3] Daniel Povey, Chun Chieh Chang, David Etter, Leibny Paola Garcia Perera, Sanjeev Khudanpur, Ashish Arora proposed a Decomposition expands the recognition of exceptional characters by breaking characters into littler graphemes that are shared overall characters. The utilization of character decay strategies is utilized to break characters into littler constituent graphemes. Cangjie is utilized for Chinese character disintegration and Korean Jamo is utilized for Korean character decay. Character disintegration lessens the size of the Neural Network models and enables preparing guides to be shared crosswise over remarkable characters with similar graphemes.

Merits

- A CNNTDNN neural network model utilizing character deterioration has altogether fewer parameters than the benchmark while likewise improving character mistake rate.
- The Chinese deterioration gives a slight improvement to the character mistake rate (10.44% versus 9.87%) while utilizing a model with less parameter (24.9M versus 16.7M).

Demerits

- The boss burden of optical character recognition checking is the possibility to bring blunders into an examined report.
- No OCR examining framework is reliable, and low-quality archives can make enough mistakes to require long and tedious editing.

[4] Li, Q., An, W., Zhou, A., & Ma, L. (2016) proposed a disconnected written by hand Chinese character recognition device has been created dependent on the Tesseract open source OCR motor. The instrument chiefly contributes on the accompanying two points: First, a manually written Chinese character highlights library is produced, which is free of a particular client's composing style; second, by preprocessing the information picture and changing the Tesseract motor, numerous applicant recognition results are yield dependent on weight positioning. At long last, the Tesseract motor is changed in accordance with yield various recognition results.

Merits

- The disconnected manually written Chinese character recognition dependent on the Tesseract motor is plausible at a specific degree.

Demerits

- In OCR the whole report should have been looked at cautiously and afterward physically adjusted.
- Not worth accomplishing for limited quantities of content.

[5] Mathur, A., Pathare, A., Sharma, P., & Oak, S. The motivation behind conveying the yield in the type of voice/discourse is to serve the data that is available on the record to the outwardly impeded. Man-made intelligence-based perusing framework utilizing OCR is an artificial knowledge perusing framework created utilizing an advanced cell camera joined with OCR. This application identifies the content utilizing the camera and sweeps the content and afterward changes over it into computerized content which is perceived by the framework and presentations the deciphered content and gives discourse yield. This framework is an OCR perusing framework that utilizes camera application present in your advanced cells joined with OCR (Optical Character Recognition). This framework will assist you with taking an image or sweep the archive present with client utilizing the telephone's camera, the picture will be checked and the application will peruse the content written in English language and convert the yield in discourse group.

Merits

- The whole working of Language Translator, alongside the least prerequisites expected to execute it. Subsequently, the outwardly disabled individual can without much of a stretch utilize this AI-based Reading framework as an amicable basic application in all around the world.

Demerits

- OCR isn't 100% precise, there are probably going to be a few mix-ups made during the procedure.
- If the first report is low quality or the penmanship hard to peruse, more mix-ups will be happens.

3. Proposed Algorithm

3.1 Back propagation algorithm

Back propagation (backward propagation) is a significant scientific apparatus for improving the precision of expectations in information mining and machine learning. Artificial neural networks use back propagation as a learning algorithm to figure a slope plunge regarding loads. It is the technique for adjusting loads of a neural net dependent on the mistake rate got in the past age (i.e., emphasis). Appropriate tuning of the loads enables you to lessen blunder rates and to make the model solid by expanding its speculation. The back propagation algorithm requires a numerical depiction of the characters. Learning is realized using the back-propagation calculation with the learning rate. The inclination is determined, after each emphasis and contrasted and edge slope esteem. On the off chance that inclination is more noteworthy than the limit esteem, at that point it performs next emphasis. The group steepest plunge preparing capacity is prepared. The loads and predispositions are refreshed toward the negative slope of the presentation work.

Pseudo Code of Back Propagation Algorithm

```
Initialize Weights;
While not Stop-Criterion do
    For all i,j do
    $w_{ij} = w_{ij} - \eta\frac{\partial E}{\partial w_{ij}}$
    End For
End While
```

4. Experimental Results

Classification Ratio

Table 1 Comparison Table of Classification Ratio

Genetic Algorithm	Self Organizing Map (SOM) Algorithm	Proposed Back Propagation Algorithm
33	41	50
38	53	65
47	66	74
59	72	83
68	80	95

The comparison table of Classification Ratio shows the different values of Genetic Algorithm, Self Organizing Map Algorithm and proposed Back Propagation Algorithm. When Comparing the Genetic Algorithm, Self Organizing Map (SOM) Algorithm and proposed Back Propagation Algorithm, the proposed Back Propagation Algorithm provides the better results. The Genetic Algorithm value starts from 33 to 68, Self Organizing Map (SOM) Algorithm values starts from 41 to 80 and the proposed Back Propagation Algorithm values starts from 50 to 95. The proposed Back propagation Algorithm provides the great results.

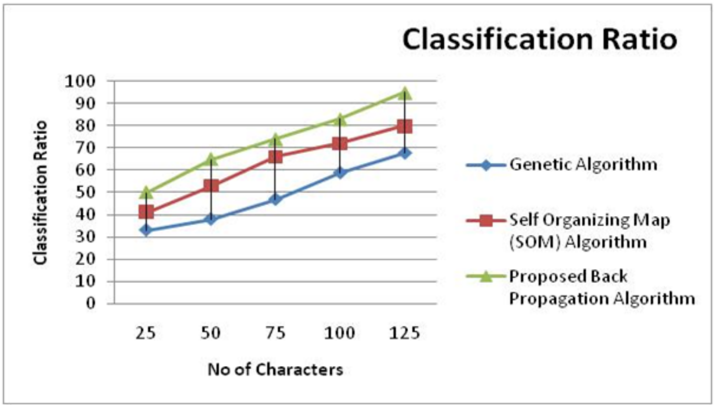

Figure 2 Comparison Chart of Classification Ratio

The Comparison Chart of Classification Ratio demonstrates the different values of Genetic Algorithm, Self Organizing Map (SOM) Algorithm and proposed Back Propagation Algorithm. In this Chart shows the No of Characters in X axis and the Classification Ratio in Y axis. The Genetic Algorithm value starts from 33 to 68, Self Organizing Map (SOM) Algorithm values starts from 41 to 80 and the proposed Back Propagation Algorithm values starts from 50 to 95. The proposed Back propagation Algorithm provides the great results compared than other two algorithms.

Identification Ratio

Table 2: Comparison table of Identification Ratio

Genetic Algorithm	Self Organizing Map (SOM) Algorithm	Proposed Back Propagation Algorithm
44.2	51	57.3
49.5	63	67.9
54.4	70	76.8
62.8	81	89.2
73.1	89	96.6

The comparison table of Identification Ratio shows the different values of Genetic Algorithm, Self Organizing Map (SOM) Algorithm and proposed Back Propagation Algorithm. When Comparing the Genetic Algorithm, Self Organizing Map (SOM) Algorithm and proposed Back Propagation Algorithm, the proposed Back Propagation Algorithm provides the better results. The Genetic Algorithm value starts from 44.2 to 73.1, Self Organizing Map (SOM) Algorithm values starts from 51 to 89 and the proposed Back Propagation Algorithm values starts from 57.3 to 96.6. The proposed Back propagation Algorithm provides the great results.

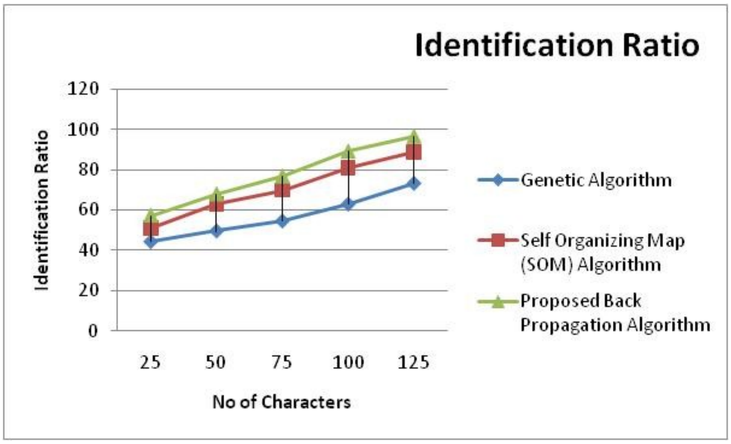

Figure 3 Comparison Chart of Identification Ratio

The Comparison Chart of Identification Ratio demonstrates the different values of Genetic Algorithm, Self Organizing Map (SOM) Algorithm and proposed Back Propagation Algorithm. In this Chart shows the No of Characters in X axis and the Identification Ratio in Y axis. The Genetic Algorithm value starts from 44.2 to 73.1, Self Organizing Map (SOM) Algorithm values starts from 51 to 89 and the proposed Back Propagation Algorithm values starts from 57.3 to 96.6. The proposed Back propagation Algorithm provides the great results compared than other two algorithms.

Accuracy Ratio

Table 3 Comparison table of Accuracy Ratio

Genetic Algorithm	Self Organizing Map (SOM) Algorithm	Proposed Back Propagation Algorithm
18	19.5	26.3
26	34.9	40.8
30	41.8	55.5
48	52.4	68.7
57	69.6	74.2

The comparison table of Accuracy Ratio shows the different values of Genetic Algorithm, Self Organizing Map (SOM) Algorithm and proposed Back Propagation Algorithm. When Comparing the Genetic Algorithm, Self Organizing Map (SOM) Algorithm and proposed Back Propagation Algorithm, the proposed Back Propagation Algorithm provides the better results. The Genetic Algorithm value starts from 18 to 57, Self Organizing Map (SOM) Algorithm values starts from 19.5 to 69.6 and the proposed Back Propagation Algorithm values starts from 26.3 to 74.2. The proposed Back propagation Algorithm provides the great results.

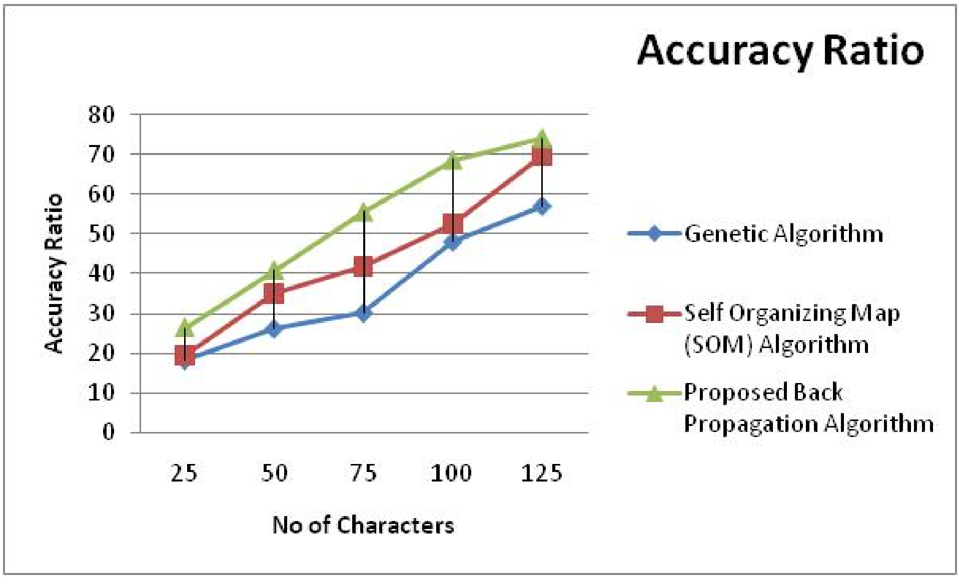

Figure 4 Comparison Chart of Accuracy Ratio

The Comparison Chart of Accuracy Ratio demonstrates the different values of Genetic Algorithm, Self Organizing Map (SOM) Algorithm and proposed Back Propagation Algorithm. In this Chart shows the No of Characters in X axis and the Accuracy Ratio in Y axis. The Genetic Algorithm value starts from 18 to 57, Self Organizing Map (SOM) Algorithm values starts from 19.5 to 69.6 and the proposed Back Propagation Algorithm values starts from 26.3 to 74.2. The proposed Back propagation Algorithm provides the great results compared than other two algorithms.

Conclusion

The proposed Back propagation algorithm is utilized for character recognition of Tamil Script. We examined another portrayal of Tamil Character Recognition, and algorithm effectively orders written by hand and furthermore for Printed Tamil characters. Progressively compelling and productive component location methods will make the framework all the more dominant. There are still some more issues in recognition. OCR is planned for perceiving printed reports. The information archive is perused preprocessed, include extricated and perceived and the perceived content is shown in an image box. Keeping up and getting the substance from and to the books is exceptionally troublesome. The algorithm is utilized to present for Tamil character recognition. Back propagation is quick, straightforward and simple to program. It has no parameters to tune separated from the quantities of information.

References

1. Kishna, N. P. T., & Francis, S. (2017). Intelligent tool for Malayalam cursive handwritten character recognition using artificial neural network and Hidden Markov Model. 2017 International Conference on Inventive Computing and Informatics (ICICI). doi:10.1109/icici.2017.8365201.
2. Prameela, N., Anjusha, P., & Karthik, R. (2017). Off-line Telugu handwritten characters recognition using optical character recognition. 2017 International Conference of Electronics, Communication and Aerospace Technology (ICECA). doi:10.1109/iceca.2017.8212801.
3. Daniel Povey, Chun Chieh Chang, David Etter, Leibny Paola Garcia Perera, Sanjeev Khudanpur, Ashish Arora," Optical Character Recognition with Chinese and Korean Character Decomposition", ©IEEE International Conference on Document Analysis and Recognition Workshops (ICDARW).
4. Li, Q., An, W., Zhou, A., & Ma, L. (2016). Recognition of Offline Handwritten Chinese Characters Using the Tesseract Open Source OCR Engine. 2016 8th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC). doi:10.1109/ihmsc.2016.239.
5. Mathur, A., Pathare, A., Sharma, P., & Oak, S. AI based Reading System for Blind using OCR. 3rd International Conference on Electronics, Communication and Aerospace Technology (ICECA). doi:10.1109/iceca.8822226.
6. P. Voigtlaender, P. Doetsch and H. Ney, "Handwriting Recognition with Large Multidimensional Long Short-Term Memory Recurrent Neural Networks," 2016 15th International Conference on Frontiers in Handwriting Recognition (ICFHR), Shenzhen, 2016, pp. 228-233.
7. S. Rawls, H. Cao, J. Mathai and P. Natarajan, "How To Efficiently Increase Resolution in Neural OCR Models,"IEEE 2nd International Workshop on Arabic and Derived Script Analysis and Recognition (ASAR), London, pp. 140-144.
8. T. Bluche and R. Messina, "Faster Segmentation-Free Handwritten Chinese Text Recognition with Character Decompositions," 2016 15th International Conference on Frontiers in Handwriting Recognition (ICFHR), Shenzhen, 2016, pp. 530-535.
9. Y. Wu, F. Yin, Z. Chen and C. Liu, "Handwritten Chinese Text Recognition Using Separable Multi-Dimensional Recurrent Neural Network," 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), Kyoto, 2017, pp. 79-84.
10. D. Povey, V. Peddinti, D. Galvez, P. Ghahremani, V. Manohar, X. Na, Y. Wang and S. Khudanpur, "Purely Sequence-Trained Neural Networks for ASR Based on Lattice-Free MMI." 2016 INTERSPEECH.
11. Jisha Gopinath, Aravind S, Pooja Chandran, Saranya S S, "Text to Speech Conversion System using OCR", International Journal of Emerging Technology and Advanced Engineering , Volume 5, Issue 1, January 2015.
12. Aaron James S, Sanjana S, Monisha M, "OCR based automatic book reader for the visually impaired using Raspberry PI", Vol. 4, Issue 7, January 2016.
13. Sujata Atul Oak, Dr. Amarsinh Vidhate, “Improved Duplicate Address Detection For Fast Handover Mobile IPv6", International Conference on Computing Communication, Control and Automation, IEEE section, (ICCUBEA- 2016), August 2016.

14. Sujata Oak, “Video Piracy Detection using Invisible watermarking", International Journal of Research in Science and Engineering (IJRISE), Vol. 3, Issue 3, June 2017.
15. Sujata Oak, “Emotion Based Music Player", International Journal of AdvancedResearch in Computer and Communication on Engineering (IJARCCE), Vol. 6, Issue 4, April 2017.