INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

e-ISSN : 2583-1062

Impact Factor : 5.725

www.ijprems.com
editor@ijprems.com

Vol. 03, Issue 09, September 2023, pp : 76-79

TAMIL CHARACTER RECOGNITION USING DEEP LEARNING

Dr. T. Vijayanandh[1], Mr. K. Subramanian[2], Azhagan S[3], Naveen Kumar M[4], Jegan M[5]

[1,2,3,4,5]National Engineering College, Kovilpatti. Anna University, Chennai, India.

ABSTRACT

One of the old languages, Tamil, is primarily spoken in southern India, Sri Lanka and Malaysia. Tamil has a vast and complex character set, making it particularly challenging to recognize a non-digitalized character. Today, digitalization is crucial for maintaining historical records. We attempt to around the challenge of paper preservation by digitalizing those Tamil handwritten characters that need to be preserved, like land papers, etc. This project aims to train a convolution neural network algorithm to recognize patterns, print those patterns, and preserve a handwritten set of Tamil characters as input in an image format. Tamil has 256 letters, most of which are nearly identical, and the majority of the characters only differ slightly at the end. As a result, it can be challenging to identify a specific character and requires well-trained datasets to do so.

Keywords -convolution neural network;

1. INTRODUCTION

Image Processing is an important method for character classification and recognition but character classification is a very difficult part especially in the Tamil language because of its character similarity, structure, and shape. Character classification is the process of locating Tamil characters in a data collection that has been originally learned and input by many users. Lack of consistent databases for training and testing is one of the main barriers to handwritten character recognition. It deals with an important issue called character classification and this is more challenging because of the similarities between them. Character classification is something that identifies the characters in Tamil that are trained initially and it can identify the characters written by different users, character Recognition can be online or offline. In an Online character recognition system, the representation of two-dimensional coordinates of successive points are done.

It is the automatic conversion of text into a digital form. Using offline character recognition, textual text is transformed into letter codes. Since there are many hidden layers in a deep neural network the parameters for the training are very huge. To prevent overfitting, we require a large set of examples. Convolution Neural Network is a special type of neural network used effectively for image recognition and classification.

2. LITERATURE REVIEW

Res Net is a pre-trained deep learning neural network which can exhibit various sizes of deep layers. ResNet-18 consists of 18 deep convolutional neural network layers. The two proposed architectures are an ensemble of a typical ResNet-18 and a modified ResNet-18. Satisfactory results were obtained using all models. The best-attained accuracy, 98.30%, was obtained using a typical ResNet-18 mode. Another approach is based on MSER Architecture. It is a well-known method for detecting regions in images The results depend on the quality of the image, however, in most cases, it recognized most of the text presented in the images. Even in images such as logos with small letters, it is possible to automatically detect and recognize the present text. With the bare minimum of features (horizontal lines, vertical lines, circles, and arcs), an ANN-based classifier used for classification is tested. The Unicode is the target. The goal of OCR is to identify printed Tamil documents. The input material is read via a pre-processing step before being feature extracted, text recognition, and display in a picture box. Characters that have been optically processed are recognised using optical character recognition. Due to differences in ambient conditions, accurately deciphering text from real-world images is a difficult task, even when employing the greatest open-source OCR engine.

3. EASY OCR

Easy OCR is a template-matching algorithm-based font-dependent printed character reader. It can read any type of brief text that is printed on labels or directly on parts, including part numbers, serial numbers, expiration dates, manufacturing dates, lot codes, etc. Easy OCR needs to be trained to recognise the typeface. It can learn how to read any possible character from examples of photos. As a result, the recognition is incredibly quick, accurate, and versatile. During the training phase, an interactive application is used to display character samples so that the library can learn them and store them in a font file. Additionally, Easy OCR includes three pre-learned standard font files for the OCR-A, OCR-B, and Semi fonts. The recognition model is a CRNN (Convolutional Recurrent Neural Network). It is composed of 3 main components: feature extraction Resnet and VGG, sequence labeling LSTM (Long short-term memory), and decoding CTC (Connectionist temporal classification).

Using a combination of machine learning and image processing techniques, Easy OCR recognises characters and converts them into text that can be read by computers. It works with a wide range of text formats and styles, including

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

e-ISSN : 2583-1062

Impact Factor : 5.725

www.ijprems.com
editor@ijprems.com

Vol. 03, Issue 09, September 2023, pp : 76-79

handwritten text, and gives a high level of accuracy. Easy OCR's language recognition tool uses machine learning and statistical analysis to precisely identify the language of text contained in an image.

4. METHODOLOGY

A typical image processing pipeline is represented by a flow graph that includes a number of crucial steps, each of which contributes to the process of turning an input image into a useful output. These stages are broken down as follows:

Input

Getting a digitised image from a real-world source is what it entails. The chosen document is authored by hand. Each step in the process may introduce random changes to the values of pixels in the image. These changes are called noise. The document is sent to a program that saves it in PNG, JPG, or JPEG format. Record pictures regularly experience the ill effects of various sorts of debasement that renders the archive picture binarization a testing assignment.

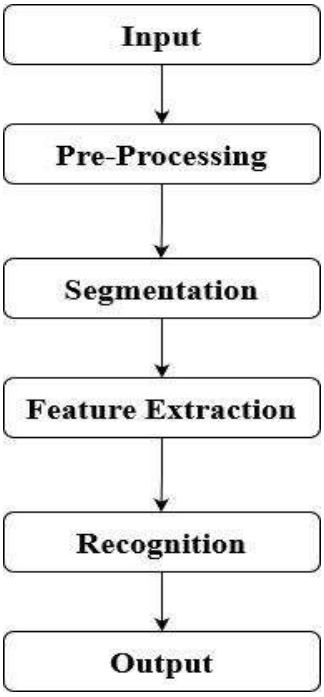

Figure 3 Methodology

Pre-processing

The initial step in processing scanned photographs is pre-processing. Three primary steps make up the reprocessing. Binarization, noise reduction and skew correction are these three. Two peak values are available. The foreground is represented by a lower peak and the white background by a high peak. To remove noise, the binarized image is first pre-processed. Noise may have built up during scanning or be a result of the document's poor quality. Before continuing with the processing, this noise needs to be eliminated. After the noise has been removed, the final image is examined for skewing. Images may be slanted to the left or right depending on their orientation. The image in this case is brightened and binarized.

5. SEGMENTATION

Clustering pixels into conspicuous image regions is the aim of image segmentation. Using segmentation, significant sections are extracted for analysis. Misrecognition or rejection results from a subpar segmentation process. Inter-line spaces are examined in the binary picture. If interline spaces are found, the image is divided into groups of paragraphs that span the gap. The paragraphs' lines are checked for horizontal space intersections with the background. The histogram of the image is used to calculate the width of the horizontal lines. The lines are then checked for intersections in vertical space by scanning them vertically. Sub-words are separated from the pre-processed image using page layout analysis and character separation. By reading through the profile starting in the first row, it is possible to identify the segmentation between text lines. A new line of text is notified if the difference in the number of black pixels between two rows is greater than a predetermined threshold. The following significant difference in the number of black pixels between two additional rows denotes the line's bottom. Similar techniques are used to separate sub-words from a line-segmented image. Then, using character width computation, the words are broken down into characters. A 64x64 window is created by scaling each character, after which a thinning algorithm is utilized to create the thinned image needed.

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

e-ISSN : 2583-1062

Impact Factor : 5.725

www.ijprems.com
editor@ijprems.com

Vol. 03, Issue 09, September 2023, pp : 76-79

6. FEATURE EXTRACTION

The next phase of segmentation is featuring extraction where each character is represented as a feature vector, which becomes its identity. The foundation of the recognition process is feature extraction. The main objective of feature extraction is to extract a set of features, such as the character's height, width, number of short and long horizontal and vertical lines, number of circles, number of horizontally and vertically oriented arcs, centroid of the image, and pixels in the character's various regions that maximise the recognition rate.

Recognition- A business solution for automating data extraction from printed or written text from a scanned document or image file and then transforming the text into a machine-readable form for use in data processing like editing or searching is optical character recognition (OCR) technology.

7. OUTPUT

The Output we get will be extracted texts from the input images. The characters are compared to a set of patterns, called a font. A character is recognized by finding the best match between a character and a pattern in the font. After the character has been located, it is normalized in size (stretched to fit in a predefined rectangle) for matching. The normalized character is compared to each normalized template in the font database and the best matches are returned.

Streamlit- An open-source Python framework called Streamlit is used to create web applications for machine learning and data science. Using Streamlit, we can quickly develop and deploy web applications. You can use Streamlit to create apps in the same manner that you create Python code. Working on the interactive cycle of coding and watching outcomes on the web app is made simple by Streamlit. It enables us to quickly develop web applications for data science and machine learning. Major Python libraries like scikit-learn, Keras, PyTorch, SymPy(latex), NumPy, pandas, Matplotlib, etc. are compatible with it.

8. RESULTS

After recognition, the Results were extracted from the provided image. In a List of Strings, each text was kept. The texts' Validation Accuracy was acknowledged as well.

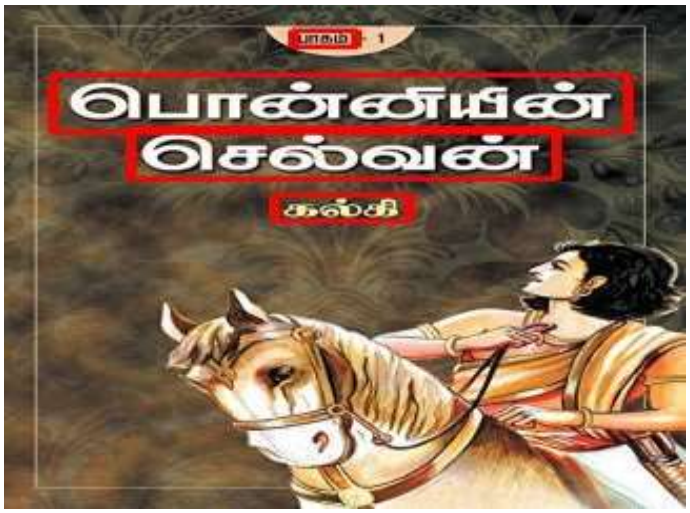

Figure 1 Test Image

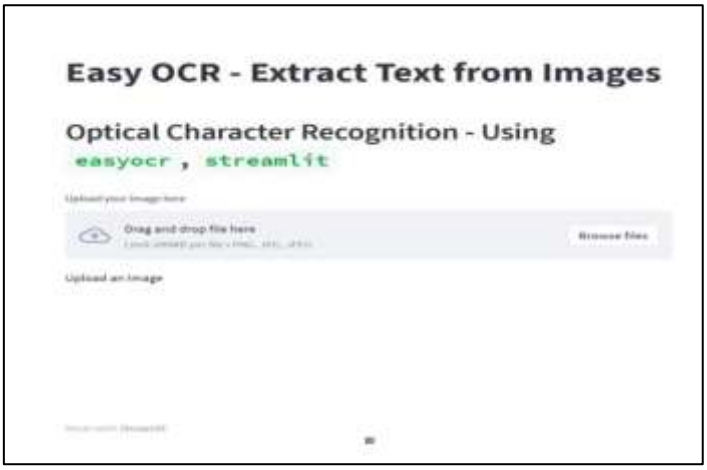

Figure 2 Result page

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

e-ISSN : 2583-1062

Impact Factor : 5.725

www.ijprems.com
editor@ijprems.com

Vol. 03, Issue 09, September 2023, pp : 76-79

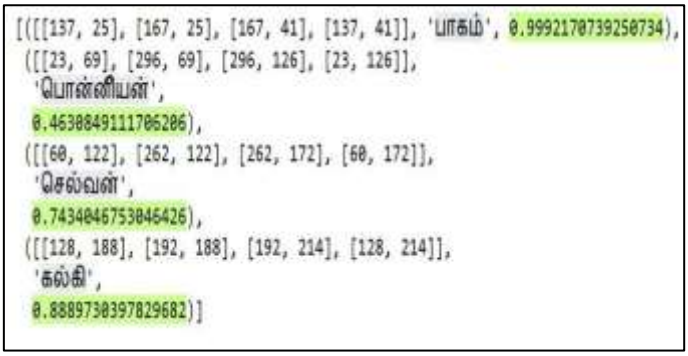

Figure 3 Result page

The Figure 2 Result page shows the final output page of streamlit webpage. There we have to give the image for classification.

9. CONCLUSION

In many ways, Easy OCR outperforms all other models. It is simple to use, only requires a few lines of code to implement, and provides accurate results for the majority of evaluated photos. It is also expanded over a wide range of languages. Without having their own language models, end users can instantly recognize and extract text from photos thanks to this. Additionally, it offers depth. coordinates for bounding boxes around identified and tokenized words, making it easy to analyze individual pieces of text.

10. FUTURE STUDIES

The full, rich dataset of handwritten Tamil-Brahmi characters will be used to train this model. The model could be used to detect Tamil-Brahmi inscriptions and scriptures which when used with enhanced image recognition algorithms could detect the inscriptions, even from rock engravings and manuscripts. We have to build up a small handwriting database ourselves, which leads to the small capacity of our experimental database. To make our experiment result more convincing, we plan to enlarge our database in future work. Additionally, it will update its features, such as language translation. so that foreigners who cannot grasp the native language may find it useful.

11. REFERENCES (APA 6TH EDITION)

[1] R, S. R., & M, M. S. (2021). Tamil Character Recognition in Palm Leaf Manuscripts. International Research Journal of Tamil, 3(2), 70-77. https://doi.org/10.34256/irjt21210

[2] Torres, P.M.B. (2017). Text recognition for objects identification in the industry. International Journal of Mechatronics and Applied Mechanics. 2017. 81-84.

[3] Kim, M.S., Cho, K.T., Kwag, H.K., Kim, J.H. (2004). Segmentation of Handwritten Characters for Digitalizing Korean Historical Documents. In: Marinai, S., Dengel, A.R. (eds) Document Analysis Systems VI. DAS 2004. Lecture Notes in Computer Science, vol 3163. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-28640-0_11

[4] Alyahya, Haifa & Al-Salman, Abdul Malik & Ben Ismail, Mohamed Maher. (2020). Deep ensemble neural networks for recognizing isolated Arabic handwritten characters. ACCENTS Transactions on Image Processing and Computer Vision. 6. 2455-4707. 10.19101/TIPCV.2020.618051.

[5] N. Shaffi and F. Hajamohideen, "uTHCD: A New Benchmarking for Tamil Handwritten OCR," in IEEE Access, vol. 9, pp. 101469-101493, 2021, Doi: 10.1109/ACCESS.2021.3096823.

[6] Can YS, Kabadayı ME. CNN-Based Page Segmentation and Object Classification for Counting Population in Ottoman Archival Documentation. Journal of Imaging. 2020; 6(5):32.https://doi.org/10.3390/jimaging6050032