www.ijprems.com
editor@ijprems.com

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

Vol. 03, Issue 05, May 2023, pp : 1358-1361

e-ISSN : 2583-1062

Impact Factor : 5.725

A SURVEY ON VARIOUS TECHNIQUES FOR HANDWRITTEN TEXT EXTRACTION

R. Swetha[1], N. Rajathi[2]

[1]M. Tech [Data Science] Department of Information Technology Kumaraguru College of Technology, Coimbatore, India

[2]Professor Department of Information Technology Kumaraguru College of Technology Coimbatore, India

ABSTRACT

An electronic device may be able to read and comprehend handwritten input from a variety of sources, including written words, photographs, and other digital touch-screen devices. Character recognition in handwriting is the term used to describe this idea. This concept is applied in a wide range of industries, such as form data entry, bank check processing, and package delivery. It is currently developing into a major issue in the field of pattern recognition, making a solution very challenging. This article looks at how deep learning methods can be used to solve detection and pattern recognition problems. The best character recognition technique will ultimately be determined by which approach achieves the highest prediction rates for categorizing the characters across a range of datasets.

Index Terms—Pattern Recognition, Character Recognition, Deep Learning.

1. INTRODUCTION

Handwriting recognition (HWR), also known as handwritten text recognition (HTR), is the ability of a computer to receive and interpret intelligible handwritten input from sources such as paper documents, photographs, touch-screens and other devices. In many different industries nowadays, pattern character recognition is used, especially when it comes to handwritten character recognition for data form entry, mail sorting, and package sending. There are numerous ways to write characters, some of which produce characters of variable sizes and forms depending on the writer's handwriting style. Because of this, the aforementioned application incorrectly classifies the characters, thereby affecting the overall process. This is why the recognition of handwriting is currently a top priority in the field of pattern recognition. In order to address this issue, the main objective is to apply multiple deep learning techniques to provide error-free handwritten character recognition.

2. LITERATURE SURVEY

In [1], it has been found that edge-based text extraction techniques and hybrid approaches using connected component (CC) and texture analysis perform best in terms of accuracy (98.53%).The EM algorithm and ML segmentation approach have the best precision (96%) and recall rate (93%) rates. The best processing time for a digital filter utilizing a Haar wavelet is 1.17 seconds. The ALCM and histogram projection-based methods have the highest accuracy (99.5%) for handwritten text document images. In[2], the authors suggested an architecture based on CNN . Easter2.0 is made up of numerous layers of 1D Convolution, Batch Normalization, and Utilize the following terms: ReLU, Dropout, Squeeze-and-Excitation Module. This work achieves state of the art algorithm(SOTA).

The paper [3] describes the architecture of the neural network and two ways of increasing the volume of training data. StackMix+blot is the name of the model used here.The described augmentations – Hand Written Blots and StackMix - further improve the quality of recognition, demonstrating the best result among the currently known handwriting recognition systems. To overcome manual filling of data and to reduce the time and effort drastically, the paper [4] proposes a system that can convert the text written in the text boxes in the forms to digital data that can be processed by a machine., Pytesseract is used which is an Optical Character Recognition (OCR) Tool. Pytesseract is used to identify the text data from printed text through the scanned image of the form.

The authors [5] proposed methodology to predict the text for three types of text. The text can be of printed text, handwritten text and it can also be cursive and the other type is of semi printed text. For printed text pytesseract is used and for handwritten text ,CRNN(convolutional recurrent neural network) is used.According to Shrawan Ram et al. [6] , who mainly focused on Using deep convolution neural networks, Devanagari character recognition is possible. Twelve of the languages spoken in India were created with Devanagari Lippi. By choosing the optimum network hyperparameters, they optimize the network .

The researchers Shailesh Acharya et al. [7] explained a dataset of images of characters used in the Devanagari script. They also recommend a deep learning architecture for character recognition. On their dataset, the suggested design had the highest test accuracy rating of 98.47%. Researchers were able to increase test accuracy by around 1% by using these techniques in Deep CNN . According to Prasad K.[8] presented a novel approach to deep neural for the offline detection of handwritten characters. Networks. Because of the huge amount of data available now, training deep

neural networks has gotten simpler. OpenCV was employed to perform image processing, and TensorFlow was utilized to train the neural network.

The depth of research proposed by Duddela Sai Prashanth et al.[9] said Because of the existence of complex characters, Authentication of penned Gurmukhi symbols is still an unsolved problem. HDCR requires a common baseline dataset that facilitates the building of deep learning models. Convolution neural networks (CNNs) are constructed using three different architectures. Using CNN, they were able to attain a classification performance of 96% for training data and 94% for untrained data. .

According to Seba Susan et al. [10]explained in their research article The Government of India's language documentation and digital archiving effort includes the digitalization of Gurmukhi articles from the intervening decades that are currently perpetuated in national museums. Premised on the hidden state activations of convolutional neural networks that are autonomously trained on segments, they suggest deep structure learning of image regions .

According to Ashim Dahal et al.[11], the innovative research on the DOC Convolutional Neural Network (CNN) architecture for accuracy and efficiency is described . The Devanagari Handwritten Character Dataset (DHCD) The model has been trained using a set of data, ensuring both efficiency and the maximum accuracy possible on the dataset According to C.N. Deshmukh et al.[12] Using RGB image datasets, researchers offer in this work a finger-point-based system for classifying and identifying signed language symbols in text. The palm-sized photographs with various sizes, backgrounds, and orientations are taken to be pre-processed by the parameters for creating an algorithm based on convolution neural networks. This method makes use of Alex net for the pre-processing requirements, augmenting 47 Devanagari script symbols in conjunction with the reference ruleset made specifically for our needs, as noted in the paper. This approach provides a remarkable recognition performance at the primary level, which implies refinement for our research shortly. In this study, They have provided specific instructions and a description of the categorization parameters that encompass when developing the algorithm on the MATLAB platform with the usage of several machine learning built-in libraries .

Mayank Tiwari et al. [13] suggested in this article applies to any company, university, hospital, etc. to complete all the information requested by the specific organization's policy. As handwriting is distinctive to each person and difficult to reproduce, researchers are taking handwritten input in this circumstance. By pairing it with sensors, they can further enhance the confidentiality of that information. Machine learning and Internet of Things (IoT) devices can be used to build an environment that is smarter, more productive, and environmentally sustainable .

[14] According to B.Sharda et al. In this article, they offer several methods for differentiating handwritten Devanagari word documents into their specific characters (or pseudo-characters). Researchers used the encoder-decoder-based convolutional neural network ShiroreakhaNet for accurate identification and delineation of the shiroreakha. After that, upper and lower modifiers are segregated using the shiroreakha structural patterns and attributes. They gathered data from a variety of domains to determine the efficacy of the solutions. The proposed approaches greatly outperformed state-of-the-art methodologies in comparison, it was concluded .

According to Sandhya Sharma et al. [15] , this research recognizes city names. The interpretation of scribbled town names is one of the prospective domains in the field of postal automation. To identify objects employing a segmentation-free method (Holistic approach). Convolutional neural networks (CNNs), are one of the deep learning architectures, and their function is clarified in the suggested study. Adam and the stochastic gradient descent (SGD) optimizer of the proposed CNN system has been trained, affirmed, and examined using batch sizes of 2, 4, and 8 with learning rates (LR) of 0.001, 0.01, and 0.1. Ten classes of 400 samples each of scribbled city names written in the Gurmukhi script are employed to train and assess the model. With a batch size of 4, an LR of 0.001, as well as an Adam optimizer, their analysis found that the CNN model has the greatest average coefficient of determination, falling in at 99.13.

According to Jossy George et al. [16] , segregating handwritten text remains challenging due to a variety of issues, including skewed and overlapping lines, the existence of touched, damaged, and degraded letters, and variances in writing styles. As a result, experts in this field are constantly attempting to create new methods for the accurate segmentation and recognition of characters. Segmentation can be used in the character recognition process at the line, word, and character levels. The initial phase in the text/character recognition process is text line segmentation. This tudy presents the line segmentation techniques for character recognition of manuscripts. With a concentration on line segmentation, the multiple levels of segmentation, comprising word and character segmentation, are studied .

The major goal of J.Shana et al. [17] is to effectively identify characters using convolutional neural networks (CNN). Following appropriate scaling, the other conventional algorithms, such as Naive Bayes and decision trees, were also fitted. Convolutional, max-pooling, dense, and dropout layers are used in the CNN model. All layers have been stimulated using ReLU, with the notable exception of the last dense layer, which tends to make use of the SoftMax activation function. Over the previous models, CNN has an advantage in that it can identify key traits without human

oversight. Pre-processing techniques used included Gray scaling, straightening, smoothing, and enlarging the user-input image for the English alphabet. The model predicts the characters from input that has already been processed. According to experimental findings, the CNN model delivered 93% more accuracy than different models .

Meghna B. Patel et al. [18], HCR simplifies the process to scan residences, postal codes, bank checks, and tax and admission forms automatically. Handwritten character identification is challenging compared to printed character recognition since it might depend heavily on the style of writing, convexity, breadth, slashes, and density of the alphabet, which happens from time to time. There has been enough research done on foreign handwritten scripts like Arabic, Chinese, and Japanese, but relatively little has been done on Indian scripts. This essay compares and contrasts the well-known Indian scripts of Tamil, Telugu, Kannada, and Gujarati. This work offers a thorough investigation depending on the dataset, procedures, and precision adopted .

Given the extreme deformation of MODI documents, MODI recognition necessitates a transform invariant technique[19]. Feature extraction techniques have been used in the past to achieve invariant handwritten character recognition, but there is still room to improve the outcomes under global transformations. Only local transform invariance is now displayed by convolution-pooling architecture and data augmentation in convolutional neural networks. The suggested classification system employed to maintain regional randomness for MODI recognition, CNN-based transfer learning, and a generic feature extractor spectrum of the aligned differential is used. The empirical investigation centered on PCA and confusion matrices are used to select the invariant property and highlight the groups that contribute to the low acknowledgment rate. The suggested classifiers are evaluated on a converted MODI dataset after being trained on a self-generated handwritten MODI character dataset. The observations confirmed that the proposed structure can still comprehend MODI handwritten characters after improvements without the necessity for data augmentation or network alteration .

The effectiveness of numerous well-known classifiers is examined in this research as it takes into consideration the real-world challenge of examination of published scripts using image data of official Indian documents[20]. For this evaluation of the performance, the AAR (average accuracy rate) and MBT (model building time) are two vital evaluative factors that are determined. 459 printed document images were used in the experiment, which used 5-fold cross-validation. The Simple Logistic model has the highest AAR of any model, at 98.9%. With minimal MBT of 0.09 s, BayesNet and Random Forest models possess median prediction performance of 96.7% and 98.2%, correspondingly .

3. CONCLUSION

The numerous methods for character identification and feature extraction for Devanagari handwritten characters are also covered in this paper. The survey's results demonstrate that Transfer Learning is the best deep learning algorithm for recognizing handwritten characters because it offers a wide range of internally optimized architectural designs that are carefully built to yield excellent results. Future versions of the models might be enhanced by making use of a huge database and adding additional Latin and Indian characters to create a more universal system that could be used with other languages or scripts.

4. REFERENCE

[1] Ghai, D., & Jain, N. (2013). Text extraction from document images-a review. International Journal of Computer Applications, 84(3).

[2] Chaudhary, K., & Bali, R. (2022). Easter2. 0: Improving convolutional models for handwritten text recognition. arXiv preprint arXiv:2205.14879.

[3] Shonenkov, A., Karachev, D., Novopoltsev, M., Potanin, M., & Dimitrov, D. (2021). StackMix and Blot Augmentations for Handwritten Text Recognition. arXiv preprint arXiv:2108.11667.

[4] González-Acuña, R. G., Chaparro-Romo, H. A., & Melendez-Montoya, I. (2021). Optical character recognition. In Optics and Artificial Vision. IOP Publishing.

[5] Bagwe, S., Shah, V., Chauhan, J., Harniya, P., Tiwari, A., Gupta, V., ... & Mehendale, N. (2020). Optical Character Recognition Using Deep Learning Techniques for Printed and Handwritten Documents. Available at SSRN 3664620.

[6] Ram, S., Gupta, S., & Agarwal, B. (2018). Devanagri character recognition model using deep convolution neural network. Journal of Statistics and Management Systems, 21(4), 593-599.

[7] Acharya, S., Pant, A. K., & Gyawali, P. K. (2015, December). Deep learning-based large-scale handwritten Devanagari character recognition. In 2015 9th International conference on software, knowledge, information management and applications (SKIMA) (pp. 1-6). IEEE.

[8] Sonawane, P. K., & Shelke, S. (2018, August). Handwritten Devanagari Character Classification using Deep Learning. In 2018 International Conference on Information, Communication, Engineering and Technology (ICICET) (pp. 1-4). IEEE.

[9] Prashanth, D. S., Mehta, R., Ramana, K., & Bhaskar, V. (2022). Handwritten Devanagari Character Recognition using modified Lenet and Alexnet convolution neural networks. Wireless Personal Communications, 122(1), 349-378.

[10] Susan, S., & Malhotra, J. (2020). Recognizing Devanagari script by deep structure learning of image quadrants. DESIDOC Journal of Library & Information Technology, 40(5), 268-271.

[11] Dahal, A., & Kattel, S. (2022). Predicting Handwritten Devanagari Characters using modified-Lenet Model Architecture.

[12] Tantarpale, M. S., & Deshmukh, C. N. (2022, August). An Approach to Pattern Recognition for Identification of Devnagari Script Based on Fingertips and Palm. In Journal of Physics: Conference Series (Vol. 2327, No. 1, p. 012032). IOP Publishing.

[13] Tripathi, O., Tiwari, M., & Gupta, A. (2022). Handwritten Text Recognition with IoT Devices. In Advanced Computational Techniques for Sustainable Computing (pp. 153-165). Chapman and Hall/CRC.

[14] Bhat, M. I., Sharada, B., Imran, M., & Obaidullah, S. (2022). Automatic Segmentation of Handwritten Devanagari Word Documents Enabling Accurate Recognition. In International Conference on Mining Intelligence and Knowledge Exploration (pp. 78-88). Springer, Cham.

[15] Sharma, S., Gupta, S., Gupta, D., Juneja, S., Singal, G., Dhiman, G., & Kautish, S. (2022). Recognition of Gurmukhi handwritten city names using deep learning and cloud computing. Scientific Programming, 2022.

[16] Joseph, S., & George, J. (2023). A Review of Various Line Segmentation Techniques Used in Handwritten Character Recognition. Information and Communication Technology for Competitive Strategies (ICTCS 2021), 353-365.

[17] Sanjay Krishna, R., Jaya Suriya, E., & Shana, J. (2022). Study of Classification Algorithms for Handwritten Character Recognition. In Computational Vision and Bio-Inspired Computing (pp. 461-470). Springer, Singapore.

[18] Patel, M. B., Parikh, S. M., & Patel, A. R. (2022). Comparative Study of Handwritten Character Recognition System for Indian Languages. In ICT with Intelligent Applications (pp. 797-806). Springer, Singapore.

[19] Jadhav, S., & Inamdar, V. (2022). Convolutional Neural Network and Histogram of Oriented Gradient Based Invariant Handwritten MODI Character Recognition. Pattern Recognition and Image Analysis, 32(2), 402-418.

[20] Obaidullah, S. M., Mondal, A., Das, N., & Roy, K. (2014). Script identification from printed Indian document images and performance evaluation using different classifiers. Applied Computational Intelligence and Soft Computing, 2014.