Journal of Population Therapeutics & Clinical Pharmacology

RESEARCH ARTICLE
DOI: 10.47750/jptcp.2023.30.07.031

A Framework to Detect Digital Text Using OCR Machine Learning

Arun Kumar R[1*], Mathanagopal.V[2], Kaviyarasan.R[3], Srivaratharaj.K[4]

[1]Assistant Professor, Department of Information Technology, K.S.Rangasamy College of Technology Namakkal, TamilNadu

[2,3,4]Department of Information Technology, K.S.Rangasamy College of Technology, Namakkal, TamilNadu.

***Corresponding author:** Arun Kumar R, Assistant Professor, Department of Information Technology, K.S.Rangasamy College of Technology Namakkal, TamilNadu, Email: rarunkumar@ksrct.ac.in

Submitted: 16 February 2023; Accepted: 17 March 2023; Published: 10 April 2023

ABSTRACT

The deep learning algorithm used in this paper to explain optical character recognition Deep learning and character recognition have recently caught the attention of numerous scholars. In many classification and recognition problems, deep neural networks operate at the cutting edge. Ocular character recognition is referred to as OCR. utilizes a character's optical picture as input and outputs that character. Numerous uses for it exist, such as robotics, traffic monitoring, and the digitization of printed materials. OCR can be implemented using Convolutional neural networks are examples of deep neural network designs (CNN),which is a well-known example. Traditional CNN classifiers can classify pictures using the soft-max layer by learning the most important 2D characteristics that are present in the medical images.

Keywords: *OCR, machine learning, recognition, character recognition, CNN*

INTRODUCTION

The OCR (optical character recognition) technology converts writing into code that can be read by computers [1]. Today, OCR is used to transform typewritten documents into digital form in addition to digitizing handwritten mediaeval manuscripts. The OCR (optical character recognition) technology converts writing into code that can be read by computers [1]. OCR is now used to convert typewritten documents in addition to digitizing handwritten mediaeval manuscripts. An OCR system's primary components are the extraction of characteristics, categorization, and discrimination of these qualities (based on patterns).As a subset of OCR, handwritten OCR is becoming more and more common. Additionally, it is categorized as an offline device. The assistant editor in charge of organizing the evaluation of this piece and approving it for publication used a web-based system based on Jenny Mahoney. The data entry Online systems' input is more dynamic and is based on the movement of a pen tip with a defined velocity, projection angle, position, and locus point, as opposed to offline systems' input, which is static and takes the form of scanned medical images. Since an online system avoids the issue of input data duplication that an offline system has, it is judged to be more complex and sophisticated than an offline system. digitalizing paper papers [3]. Because one no longer needs to sift through mountains of documents and files to find the information they need, retrieving crucial

information has become simpler as a result. Organizations are addressing the demands for the digital preservation of historical data, legal papers, and other documents.

The assistant editor in charge of organizing the evaluation of this piece and approving it for publication used a web-based system based on Jenny Mahoney. input of data The input for the online systems is more dynamic and is based on the movement of a pen tip with a certain velocity, projection angle, position, and locus point, as opposed to the static input of scanned pictures used by offline systems. An online system is stated to be more advanced and complex than an offline system since it has solved the problem of input data overlap that arises in the latter. In the 1940sone of the early OCR systems was developed; as technology advanced, the system improved its capacity to handle both printed and handwritten characters, which made OCR equipment more widely available. In 1965, the "IBM 1287" early reading gadget was revealed at the New York World's Fair. For the first time ever, an optical reader was able to read handwritten digits.In the 1970s, research concentrated on improving the OCR system's efficiency and response time. An OCR software system was created between 1980 and 2000 and is now used in educational institutions, survey OCR, and the identification of imprinted characters on metal bar. In order to preserve historical records in digital form and give researchers access to these materials, binarization techniques were created in the early 2000s.These programme supported the growth of these people's reading and writing skills. In the last 10 years, academics have examined a wide range of machine learning methods, including Support Vector Machine (SVM), Random Forests (RF), k Nearest Neighbor (KNN), Decision Tree (DT), and Neural Networks.Researchers combined machine learning and medicalimage processing methods to increase the optical character recognition system's accuracy. Deep learning has been the main technology used recently by researchers to create methods for digitizing handwritten documents. The use of GPUs and cluster computing, as well as the increased effectiveness of deep learning architectures like recurrent neural networks (RNN), convolutional neural networks (CNN), and long short-term memory (LSTM) networks, have all contributed to this paradigm shift. Many people prefer to use pen and paper to draught essential physical documents, even though a variety of technologies can be used to produce text-based documents. Data from traditional papers must be stored and retrieved with care. The use of handwriting detection is growing rapidly in this era of globalization. Handwriting recognition is the process by which a machine converts human writing into digital shape. On PDAs and mobile Computers, handwriting detection is a method for identifying handwritten symbols. One benefit of handwriting recognition is the ability to use electronic storage, which requires less staff to organize and order documents.

Deep learning

Deep Learning is a subdivision of Machine Learning, which is a subset of Artificial Intelligence. Artificial intelligence is a method that allows technology to replicate human behaviour. Deep Learning is a type of machine learning that draws inspiration from the structure of the human brain. Machine learning is a technique for creating artificial intelligence that makes use of programmes that have been trained on datasets.. A machine learning method called "deep learning" is used to extract data traits and duties. Facts can be conveyed visually, verbally, or acoustically. Since Deep Learning characteristics are recognised by Neural Networks without the need for human input, deep learning is also referred to as end-to-end learning. The training process for deep neural networks takes hours, if not months.

Convolutional Neural Network

An picture classification network called the Convolutional Neural Network (CNN) uses deep learning. The fundamental principle of CNN is to recognize entire objects, such as animals, people, and vehicles, by first using preset convertible filters to identify patterns in picture edges and parts of objects. Since 1959, Hubel and Wiesel have covered news for CNN. Despite the algorithms' success, learning could not be automated. CNNs are frequently used to identify

pictures, group and categories medical images, find items, and other tasks. Comparatively speaking, CNNs require much less preparation than conventional medical image analysis methods. The CNN's transmission system and the structure of an animal's visual cortex are comparable. Comparatively speaking to other picture classification methods, CNNs require incredibly little planning.

Over all Description

Handmade text recognition is known as HTR. We can convert paper documents into digital files that can be viewed by users, such as scanned photographs or medical images, by using a deep learning programme.

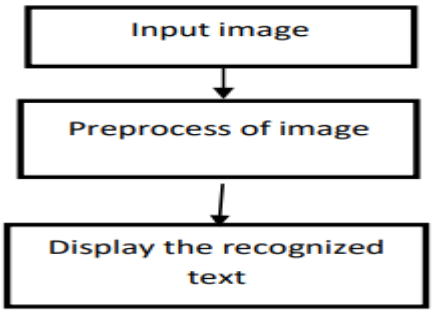

FIGURE 1: preprocessed image

LITERATURE SURVEY

It can identify text in medical images and convert it. Humans have been recording their thoughts in the form of letters, transcripts, and so forth for a long time in order to share them with others. However, when computer technology advanced, the handwritten text format was simply transferred to a digitally produced machine. Many believe that such a way of translating handwritten writing to digital language is necessary since processing such data is simple and convenient. More academics are working on algorithms for handwritten text recognition (HTR).

Hand Written Recognition

By fusing the Long Short Term Memory (LSTM) and Convolutional Neural Network (CNN) recurrent neural networks, they used line level approaches. These techniques were used to identify characteristics. You teach the model using Connectionist Temporal Classification. As opposed to more traditional techniques that use hand-engineered features, many data-driven deep-learning-based systems will take and select pertinent characteristics from the training sample set to be used. The HTR, however, is essential in a future where handwritten words must be incorporated into technology, even though the aforementioned methods have resulted in substantial advancements in recognition. By utilizing movement aids, a large number of people can utilize HTR services. According to the literature analysis, this is the primary article that tackles the interaction between online HTR systems and offline handwriting recognition with the aim of building a completely functional HTR system using an existing training dataset. The HTR reported in this study [14] was taught using medical images generated from online HTR systems' motion data. Training samples for HTR to OCR systems have been created using adaptive picture degradation techniques in order to generate adequate precision for real-time uses for handwritten text [14]. following handwriting picture enhancement.

The LSTM method is one of the many approaches that scholars have tested for line identification in HTR system architectures; it is similar to a wide range of other tried-and-true methods. Recurrent models have a number of drawbacks, including the inability to function on specialized machinery like Feed Forward Networks and the lack of training operations. In order to accomplish designs based on LSTM with a similar level of accuracy, many experts created a fully Feed Forward Network model. Even though handwritten text line detection is frequently the first stage in HTR systems, it is only one of several essential parts of a complete system. We'll go over the updated processes that were applied during the merger.

PROPOSED METHDOLOGY

Problem statement

There is currently no single method that effectively solves the problem of handwriting text identification due to the great difficulty in recognizing handwritten writing. The most recent

works finished to date use character identification as a method, but the accuracy and precision they achieve fall short of ideal.

METHODOLOGY

The goal of the project is to develop Handwritten Text Recognition (HTR) so that handwriting content that is only permitted for text can be accurately recognized. In this assignment, we will first gather data for training handwritten texts, then extract features from those text samples, and finally train the model using a Deep Learning method. In order to improve precision, we will use the technique in this research to recognize in terms of words rather than characters. The LSTM deep model-based algorithm is remarkably precise. Image 1 shows the basic HTR methodology. This diagram shows how HTR devices operate. The steps involved in pre- processing, training, and categorization are shown in Figure 2. The descriptions of each pre-processing step are provided below.

Preprocessing

When a paper is scanned or the source data is provided, some preparatory processing might be required. The ultimate form of the paper, which will be processed by a handwritten text recognition technology, is created with the aid of pre-processing. The main pre-processing goals are as follows: Segmentation, narrowing, binarization, noise reduction, and normalization.

Noise removal

Several kinds of noise may be present in the data during scanning, which is undesirable during the process. Noise, which is a black pixel in a white pixel or vice versa, is an example of unwanted pixels in an medical image. Background disturbance can be seen in Figure 2's initial image. As a consequence, before continuing with the editing, some of the noise must be removed. Among the noise reduction methods, a median filter with a filter size of roughly 3 3 was chosen. Figure 4 shows how the median filtering noise reduction method was used to remove the background from the original picture.

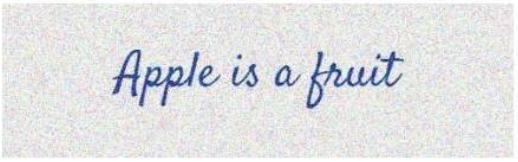

FIGURE 2: noise removal result

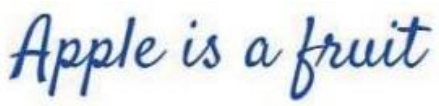

FIGURE 3: normalization

Segmentation

The act of segmenting sentences is word separation. Since the structure can only recognize one word at a time in HTR, it is crucial. The basic connection concept was used. According to this hypothesis, a word's pixels were once linked to one another. Separating the medical image from the background is another aspect of segmentation.

FIGURE 4 :segmented results

Binarization

Another crucial step in the editing of medical images is called binarization, which separates the pixels of the medical image into the foreground (black) and backdrop (white) groups (black). Only white and black are available in binary medical images. Therefore, an universal grey scale intensity threshold is used in the proposed binarization method. Figure 5 depicts the binaries picture.

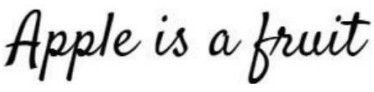

FIGURE 5: Binarization results

Normalization

Different typeface widths should be recognized by the framework. As a consequence, before

being sent to the classifier, every new character needs to be changed to a standard height. The process of adjusting the picture scale to one that the classifier approves is known as normalisation. The neural network is composed of input layers, the number of which is set and which receive medical image pixels as input.

Classification

After preprocessing is complete, a classifier receives the result from the preprocessing procedures, which is a picture of a standard dimension. The word's pixel location is fed into the algorithm as data. Although a neural network has a special technique for classification that could be used, the only strategy for back propagation has been used in this case.

Deep Learning Algorithm

An programme built on deep learning The characteristics are collected and the CNN model is used to build the model. Modern HTR systems frequently use LSTM and other Recurrent Neural Networks (RNN) for text line identification stages. In this research, a model influenced by CLDNN was outlined (Convolutions, LSTMs, and Deep Networks). The genesis designs were used to build each and every CNN layer [20]. Four layered bidirectional LSTMs (BLSTMs) were used for the LSTM layers. In this setup, the model only has a feed-forward network. This research provided a description of the strategy. Use of a licence that is authorized is limited to: The mix of 391 and LSTM- compliant devices offers tolerable accuracy.

sources to improve the Handwritten Text Recognition systems' accuracy: Researchers can create an HTR model in a variety of languages using a huge collection of ink samples at their disposal. We must educate the algorithm on a large number of datasets in order to attain its high accuracy.

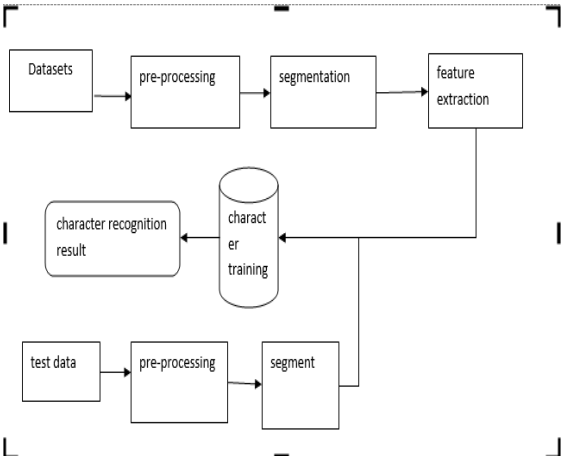

FIGURE 6: Architectures design

FIGURE 7: trainingacuuracy

Training Dataset

Both offline and online handwritten samples are used to show the effectiveness of the HTR method. The Lancaster Oslo/Bergen (LOB) texts corpus served as triggers for the handwritten text images in the IAM offline handwriting library [30] (offline), which consists of scanned papers of handwritten text pictures created by roughly 500 distinct people. Photos with text lines make up the dataset, which will be classified using various training, approval, and test datasets. These studies use a variety of information

We gathered the datasets and saved them in the block. We also gathered the picture and resized it to [0, 255] pixels. Using CNNs (convolution neural networks), form and color characteristics are extracted by using epochs. Using h5 files, we train the algorithm and distribute it. The levels that are typically seen in CNNs are listed below. Information from the Input Layer is sent to the Convolution Layer, which is in charge of the main feature extraction process. The supplied data is subjected to convolution. Moving the kernel across the input and running the sum of the

product at each position allows for convolution. The kernel's sliding steps are quantified by the stride. The depth of a convolutional layer is another name for the number of feature maps it generates.. Several convolution processes are performed on the data using different kernels. A rectified linear unit is a linear measure (ReLU).The job at hand is the focal point of CNN's architecture. Like all neural networks, CNN is conscious and capable of learning. Training facilitates learning (supervised). CNNs are feed forward networks that learn through the process of back-propagation. There are two movements in the exercise: one forward and one rearward. During the forward iteration, small random integers are used to initialize the network weight and bias. Calculate the network output using the training data. By comparing the network output with the anticipated training output, the variance is calculated. The mistake propagates backward in the reverse pass, and all weights and bias are changed to decrease it. The process is carried out repeatedly until the desired outcome is attained. After the network is finished.

IMPLEMENTATION RESULTS

The validity of the research published in Procedia Computer Science 167 (2020) 2403-2409 2407 by Mayur Bhargab Bora and associates Mayur Bhargab Bora et al./Procedia Computer Science 00 (2019) 000-000 5 obtained 88% and 93% after training and testing on the NIST dataset. The objective of this work is to use the ECOC classifier to increase the performance of the CNN character recognition system. The collection is divided into 26 files, each containing 2473 unique upper case English alphabet pictures (1483 training photos and 990 testing photos). The network is trained using the NVIDIA GeForce GT 730 graphics driver. All tests were run on a single processor with 8 GB of RAM. The extracted features are used to train the ECOC classifier once the CNNs have been trained to learn the features. Pictures are resized and transformed from RGB to grayscale in order to prepare the data for neural networks. The use of numerous popular CNN models as feature extractors in tandem with the ECOC classification has been explored. Following are the modelling parameters and training data.

CONCLUSION

In order to provide more engaging and reliable training, we may use additional preprocessing methods like jittering in our web- based character recognition apps. We can split each medical image by the standard deviation to standardize the data. Due to time and resource constraints, we could only use 20 training instances for each word to rapidly analyze and refine our model. Another method to enhance our character segmentation model is to go beyond a greedy search for the choice that seems most probable. To solve this, we would employ a more involved but still effective deciphering method like beam search. To each of the potential final beam search candidate pathways and their merged individual softmax, we may employ a character/word-based language-based model. The suggested method offers greater precision.

REFERENCE

1. Chaudhuri, Arindam and Mandaviya, Krupa and Badelia, Pratixa and Ghosh, Soumya K and others. (2017) “Optical Character Recognition System. In Optical Character Recognition Systems for Different Languages with Soft Computing Springer: 941.
2. Li, Haixiang and Yang, Ran and Chen, Xiaohui. (2017) “License plate detection using convolutional neural network. 3rd IEEE International Conference on Computer and Communications (ICCC),IEEE:17361740
3. Rajavelu, A and Musavi, Mohamad T and Shirvaikar, Mukul Vassant. (1989) “ A neural network approach to character recognition.Neural Network 5,Elsevier (2): 387393.
4. Bai,Jinfeng and Chen, Zhineng and Feng, Bailan and Xu, Bo.(2014) “Image character recognition using deep convolutional neural network learned from different languages. IEEE International Conference on medical Image Processing (ICIP):25602564.
5. Maitra, Durjoy Sen and Bhattacharya, Ujjwal and Parui, Swapan K. (2015) “CNN based common approach to handwritten character recognition of multiple scripts.13th International Conference on Document Analysis and Recognition (ICDAR),IEEE:10211025.

J Popul Ther Clin Pharmacol Vol 30(7):e259–e265; 10 April 2023.

6. Jakkula,Vikramaditya. (2006)“Tutorial on support vector machine (svm). School of EECS, Washington State University 37.
7. Ciresan, Dan Claudiu and Meier,Ueli and Gambardella,Luca Maria and Schmidhuber,Jurgen. (2011)“Convolutional neural network committees for handwritten character classification. International Conference on Document Analysis and RecognitionIEEE:11351139.

J Popul Ther Clin Pharmacol Vol 30(7):e259–e265; 10 April 2023.