IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 3, March 2024

DOI: 10.17148/IJARCCE.2024.133146

HANDWRITTEN TEXT CONTENT CLASSIFICATION SYSTEM USING ANDROID

Dr.V. Suganthi[1], Vasanth.S[2]

Associate Professor, PG & Research Department of Computer Science, Sri Ramakrishna College of Arts & Science, Coimbatore 641006 Tamil Nadu India[1]

UG Student, PG & Research Department of Computer Science, Sri Ramakrishna College of Arts & Science, Coimbatore 641006 Tamil Nadu India[2]

Abstract: Character Recognition is a technology that enables to convert different types of documents, such as scanned paper documents into editable data. The ability to understand these inputs varies in each person according to many factors. OCR is a technology that functions like human ability of reading. Although OCR is not able to compete with human reading capabilities, it can convert the content from the image files. Automatic text recognition aims at limiting these errors by using image preprocessing techniques that bring increased speed and precision to the entire recognition process. In the proposed system, the written content is converted into text content using the pattern recognition system and the same is stored in file. The proposed method uses the Learning based Spatio-Temporal Algorithm to extract the written contents. The primary objective of this system is to written content recognition system and to create the android based mobile application to save the handwriting in to text file. Robust data capture solutions handle multiple document formats and can be used with both electronic and paper documents, eliminating paper and reducing manual identification and data entry of document content into other systems.

I. INTRODUCTION

Handwriting recognition is a computer’s ability to recognize and interpret handwritten input. It’s sometimes known as ‘handwritten text recognition’. The input is usually in the form of an image such as a picture of handwritten text that is fed to pattern-recognition software, or as real-time recognition using a camera for optical scanning. It explores the content including alphabets, numbers typed or printed in the paper. Optical character recognition (OCR) technology is a business solution for automating data extraction from printed or written text from a scanned document or image file and then converting the text into a machine-readable form to be used for data processing like editing or searching. A common application of OCR technology is the automated conversion of an image based PDF, TIFF or JPG into a text based machine-readable file. OCR capabilities, the ability to extract machine-printed text from a digital image, are only one aspect of a data capture solution. Data can be extracted from documents in many different formats—hand printed text (ICR), check boxes (OMR), bar codes, etc. OCR can recognize both handwritten and printed text. But the performance of OCR is directly dependent on quality of input documents.

The OCR is designed to process images that consist almost entirely of text, with very little non-text clutter obtain from picture captured by mobile camera. This application is for the Android mobile operating system that combines Google’s open-source OCR engine. As OCR stands for optical character recognition, OCR technology deals with the problem of recognizing all kinds of different characters. Both handwritten and printed characters can be recognized and converted into a machine-readable, digital data format. Think of any kind of serial number or code consisting of numbers and 2 letters that you need digitized. By using OCR you can transform these codes into a digital output. The technology makes use of many different techniques. Put simply, the image taken is processed, the characters extracted, and are then recognized. Online handwriting recognition has recently been gaining importance for multiple reasons: (a) An increasing number of people in emerging markets are obtaining access to computing devices, many exclusively using mobile devices with touchscreens. Many of these users have native languages and scripts that are not as easily typed as English, e.g., due to the size of the alphabet or the use of grapheme clusters which make it difficult to design an intuitive keyboard layout. OCR system works with Tesseract algorithm which recognizes characters. Tesseract identifies characters in foreground pixels, called as blobs, and then it finds lines. Word by word recognition of characters is done throughout the lines. Recognition involves converting these images to character streams representing letters of recognized words. In short, recognition extracts text from images of documents

II. RELATED WORKS

Related work in the field of character recognition and OCR technologies has seen significant advancements in recent years. Researchers have explored various approaches to enhance the accuracy and efficiency of these systems. One notable avenue of research involves the utilization of machine learning algorithms for pattern recognition. Techniques such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs) have been employed to extract meaningful features from images and improve recognition accuracy.

Moreover, there has been a focus on developing robust data capture solutions that can handle diverse document formats, whether electronic or paper-based. These solutions aim to streamline the process of digitizing content, thereby reducing reliance on manual data entry and eliminating paper-based workflows. Additionally, advancements in image preprocessing techniques have contributed to enhancing the speed and precision of automatic text recognition systems.
In the realm of mobile applications, efforts have been directed towards creating user-friendly interfaces for capturing and converting handwritten content into editable text files. Integration with smartphones and tablets allows for on-the-go document digitization, empowering users to easily save and manage their handwritten notes.

Overall, the ongoing research in character recognition and OCR technologies, coupled with advancements in machine learning and mobile computing, holds promise for developing efficient and accurate systems for converting handwritten content into digital text files.

III. IMPLEMENTATION

Handwriting recognition requires much more advanced technology than OCR. Instead of using simple techniques to identify letter shapes, this type of OCR leverages a highly trained machine learning model and advanced computer vision engines to actually read what is written like a human would. The combination of highly trained machine learning models and computer vision engines is what makes it possible for handwriting OCR to replicate the way humans read handwriting.

The proposed method combines the machine learning and computer vision to improve the text recognition process. It also combines the deep learning methods for the handwritten content recognition. It replaces our previous segment-and-decode system, which first oversegments the ink, then groups the segments into character hypotheses, and computes features for each character hypothesis which are then classified as characters using a rather shallow neural network. The recognition result is then obtained using a best path search decoding algorithm on the lattice of hypotheses incorporating additional knowledge sources such as language models. This system relies on numerous preprocessing, segmentation, and feature extraction heuristics which are no longer present in our new system.

INPUT DESIGN: Handwritten text content written in the image is considered as input to the recognition system. The application system uses Android mobile camera as input capturing device. Camera captures the image of document. This is nothing but the process of scanning. In short we can say that scanning makes original document as digital image. Generally, original documents are made up of the black colored text print on the white colored background. Scanning comes with thresholding which makes the digital image as gray scale image. Human handwriting has a high degree of oscillation due to the non-uniform hand muscle forces being exerted while writing.

OUTPUT DESIGN: The character image is mapped to a higher level by extracting special characteristics of the image in the feature extraction phase. And it provides the extracted content in the form of text data representation. In this work, we use three types of augmentation schemes: (i) affine transformation, (ii) elastic distortion and (iii) multi-scale transformation both while training and testing. Under affine transformation we apply translation, scaling, rotation, and shearing. Here we restrict rotation to a random amount between 5 degrees, while shearing is restricted to 0:5 degrees along the horizontal direction which mimics the skew and cursiveness present in natural handwriting. We perform translation in terms of padding on all four sides, of upto 20 pixels in any direction, to simulate incorrect segmentation of 19 words. We randomly apply a combination of the above 3 transformations to an input image.

SYSTEM DEVELOPMENT: The idea behind CNNs is to combine a feature learning module with a trainable classifier, which often consists of a fully connected net- work. The feature learning module would replace a prior feature engineering stage, often performed by hand, to reduce data processing to a minimum. In fact, CNNs are intended to work with raw data (or data with very little preprocessing). After features have been learned from raw data, they are introduced to a train- able classifier. CNNs are interesting for solving many different problems since they provide invariance to translations or local distortions of the input.

DOI: 10.17148/IJARCCE.2024.133146

IV. CONCLUSION AND FUTURE WORK

Optical Character Recognition is a technology that functions like human ability of reading the data from image. Automatic text recognition aims at limiting these errors by using image preprocessing techniques that bring increased speed and precision to the entire recognition process. In the proposed system, the written content is converted into text content using the pattern recognition system and the same is stored in file. The proposed method uses the Learning based algorithm to extract the written contents. The system recognizes the written content and stores the extracted data in the form of regular text data. The CNN based learning model is developed to identify the written content in image file with higher accuracy.

In future, instead of being restricted to a fixed number of character sets, these new OCR programs will accumulate knowledge and learn to recognize any number of characters. The long-standing, intrinsic difficulty of character recognition itself has long blinded us to the reality that simple digitization was never the end goal for using OCR. And also the system can be improved with the assistance of the deep learning algorithm such as Boltzmann Learning Model. This can improve the accuracy and reduces the time to complete the conversion process.

REFERENCES

[1]. Keysers, D., Deselaers, T., Rowley, H.A., Wang, L.L. and Carbune, V., 2016. Multi-language online handwriting recognition. IEEE transactions on pattern analysis and machine intelligence, 39(6), pp.1180-1194.

[2]. Baldominos, A., Saez, Y. and Isasi, P., 2018. Evolutionary convolutional neural networks: An application to handwriting recognition. Neurocomputing, 283, pp.38-52.

[3]. Kulik, S.D., 2015. NEURAL NETWORK MODEL OF ARTIFICIAL INTELLIGENCE FOR HANDWRITING RECOGNITION. Journal of Theoretical & Applied Information Technology, 73(2).

[4]. Dutta, K., Krishnan, P., Mathew, M. and Jawahar, C.V., 2018, August. Improving cnn-rnn hybrid networks for handwriting recognition. In 2018 16th international conference on frontiers in handwriting recognition (ICFHR) (pp. 80-85). IEEE.

[5]. Wigington, C., Tensmeyer, C., Davis, B., Barrett, W., Price, B. and Cohen, S., 2018. Start, follow, read: End-to-end full-page handwriting recognition. In Proceedings of the European Conference on Computer Vision (ECCV) (pp. 367-383).

[6]. Voigtlaender, P., Doetsch, P. and Ney, H., 2016, October. Handwriting recognition with large multidimensional long short-term memory recurrent neural networks. In 2016 15th International Conference on Frontiers in Handwriting Recognition (ICFHR) (pp. 228-233). IEEE.

[7]. Carbune, V., Gonnet, P., Deselaers, T., Rowley, H.A., Daryin, A., Calvo, M., Wang, L.L., Keysers, D., Feuz, S. and Gervais, P., 2020. Fast multilanguage LSTM-based online handwriting recognition. International Journal on Document Analysis and Recognition (IJDAR), pp.1-14. 32

[8]. Bluche, T., Ney, H., Louradour, J. and Kermorvant, C., 2015, August. Framewise and CTC training of neural networks for handwriting recognition. In 2015 13th international conference on document analysis and recognition (ICDAR) (pp. 81-85). IEEE.

[9]. Suryani, D., Doetsch, P. and Ney, H., 2016, October. On the benefits of convolutional neural network combinations in offline handwriting recognition. In 2016 15th International Conference on Frontiers in Handwriting Recognition (ICFHR) (pp. 193-198). IEEE.

[10]. 10.Sueiras, J., Ruiz, V., Sanchez, A. and Velez, J.F., 2018. Offline continuous handwriting recognition using sequence to sequence neural networks. Neurocomputing, 289, pp.119-128.

[11]. 11.Poznanski, A. and Wolf, L., 2016. Cnn-n-gram for handwriting word recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2305-2314).

[12]. 12.Kumar, P., Saini, R., Roy, P.P. and Pal, U., 2018. A lexicon-free approach for 3D handwriting recognition using classifier combination. Pattern Recognition Letters, 103, pp.1-7.

[13]. 13.Sun, L., Su, T., Liu, C. and Wang, R., 2016, October. Deep LSTM networks for online chinese handwriting recognition. In 2016 15th international conference on frontiers in handwriting recognition (icfhr) (pp. 271-276). IEEE.

[14]. 14.Bluche, T. and Messina, R., 2017, November. Gated convolutional recurrent neural networks for multilingual handwriting recognition. In 2017 14th IAPR international conference on document analysis and recognition (ICDAR) (Vol. 1, pp. 646-651). IEEE.

[15]. 15.Du, H., Li, P., Zhou, H., Gong, W., Luo, G. and Yang, P., 2018, April. Wordrecorder: Accurate acoustic-based handwriting recognition using 33 deep learning. In IEEE INFOCOM 2018-IEEE Conference on Computer Communications (pp. 1448-1456). IEEE.