Advances and Applications in Mathematical Sciences
Volume 19, Issue 5, March 2020, Pages 321-331
© 2020 Mili Publications

A SURVEY ON CHARACTER RECOGNITION FROM HANDWRITTEN DOCUMENTS

GAGANDEEP KAUR, VARINDER SINGH, SUNIL KUMAR CHAWLA and MAHIMA BHASIN

Computer Science and Engineering Department
College of Engineering
Chandigarh Group of Colleges
Landran, India
E-mail: gagandeep.4421@cgc.edu.in
varinder.4403@cgc.edu.in
sunil.3550@cgc.edu.in
mahimabhasin02@gmail.com

Abstract

In the past few years, OCR technique has occupied very fascinating stretch of research. In this paper we describe an inclusive analysis of Handwritten Character Recognition (HCR) in cursive writing script. HCR is utilized in numerous fields like Finance sectors, Health care, commerce and several such officialdoms wherever handwritten forms are used. The Handwritten Character recognition technique is rephrasing of handwritten manuscript into machine understandable form. In order to decrease the complication of recognizing handwritten manuscript in modern researches special types of techniques, classifiers and features has used.

1. Introduction

The approach of reformation of pictures of printed typescript, handwritten script and typewritten documents into a text recognized by a machine is what we call as Optical Character recognition. This can be used for editing, reduction in storage space etc. So, in a nutshell, it relates to the possession of teaching and recognition of data. With the purpose of increasing

2010 Mathematics Subject Classification: 68T10.

Keywords: Handwritten database, HCR (Handwritten Character recognition), TIE (Text Information Extraction), MARY (Modular Architecture for Research on Speech Synthesis), TTS (Text-to-speech), Intelligent word recognition, OCR (Optical character recognition) etc.

This work was supported in part by the Department of Computer Science and Engineering faculties.

Received August 31, 2019; Accepted December 16, 2019

efficiency, accelerating turn out time, processing period and enabling swift access across the board for application users, congregating advantages of OCR and cloud computing in unique place is the key idea. In this manuscript, we are using the handwritten Characters for the recognition into the machine editable format. The Machine Learning method which depends on man-made neural network is used for identification and recognition. The task of identity recognition is done using the approach of Machine learning which will depend on the man-made neural network. That method focuses on valid features of the image in order to recognize them without seizing much help. The Neural Network advanced so as to accord the concept of biological neurons. The use of deep learning for character recognition draws to exactitude and efficacy. In differentiation to the innovation of profound learning, once mechanical progressions are cooking perfect and basic grounds to turn up with an unbeatable and most savvy answer for this troubling hitch. Visual debilitations add to weighty monetary figures both straightforwardly and in a roundabout way due to the expense of treatment and diminished capacity of work individually. With the assistance of OCR cloud, Optical Character Recognition (OCR) Cloud Base Reading Aid is only building up a product application for outwardly disabled individuals. The application for an assistive framework is kept forward going for the act of spontaneity of the possibility of the outwardly impeded individuals by perusing the content from examined record and renovating the printed data as discourse. The two innovations in particular OCR (Optical Character Recognition) for Text Information Extraction (TIE) and MARY TTS (Modular Architecture for Research on Speech Synthesis-Text-to-discourse) are the two advancements which are generally utilized for the improvement of such frameworks. OCR based MARY TTS framework have great highlights like perceptibility, multi-language support yet the later have more points of interest like the independent/disconnected mode combination, voice recording and less idleness. Other than being the most basic capacity of any assistive understanding framework, the capacity of Text Information Extraction (TIE) to decide the understandability of the yield discourse saves its indispensable space in OCR. For the procedure of transformation of discourse, there are two fundamental sorts of OCR, which may create a positioned rundown of competitor characters. The technique which includes examination of a picture with the put away glyph on a pixel-by-pixel premise is Matrix Matching and is otherwise called "Example Recognition" or "Picture Correlation". This strategy works best with typewritten message and doesn't function admirably when new text styles are experienced. Be that as it may, the universe of print media, for example, papers, books, sign sheets and menus regularly stay past the limits of outwardly debilitated individuals in spite of the dug in research endeavors in this circle. Wherefore, an assistive innovation based arrangement called BH (Binary Hearing) application is created and tried so as to look for an answer for this persevering issue. There is a requirement for the updation of the highlights of OCR to incorporate language interpretation include so as to help those (from various nations) to whom a specific language is incogitable.

2. Handwritten Character Recognition

Handwriting is a fundamental skill crucial for literacy success. Handwriting teaches the letter formation, supports reading and language acquisition. Writing by hands engages the brain in learning, handwriting based on the evidence neuroscience to plays an important role in learning of letters. Today's digital libraries not only included the printed text but also include the handwritten text. There are various tools for manipulate the handwritten text. Sundry libraries around enclose peculiar handwritten documents involving different handwriting styles. Written by hand Character Recognition (HCR) can be characterized as the marvel of change of transcribed content into machine meaningful structure. The vacillation of the penmanship styles, which can be totally extraordinary for various journalists is the real hitch in Handwritten Character Recognition (HCR). The HCR went for the execution of easy to understand PC helped character portrayal permitting fruitful extraction of characters from transcribed archives and digitalization, interpretation of the written by hand message into machine meaningful content.

Manually HCR framework is isolated into two categories:

- On-line character recognition: It is framework in which acknowledgment is performed when characters are under creation.
- Off-line character recognition: It is framework in which originally transcribed records are produced, filtered, put away in PC and then they are perceived.

There are four techniques for cursive manually written word recognition.

(1) Holistic Approach: The technique for perceiving the whole word, by removing highlights of the whole word without parting them.

(2) Segmentation based Approach: Characters are divided from word.

(3) Recognition based separation Approach: Character all the while by utilizing reasonable learning method.

(4) Mixed Approach: The blend of above techniques subsist in this framework. In this paper we present brief overview of accessible HCR for English language. HCR methods are examined with their quality and shortcomings. With the reason to characterize the information characters, various sorts of highlights are removed and various kinds of classifiers are utilized. The present examination focusses to test potential procedures with the mean to build up a disconnected HCR framework for both separate characters and cursive words for English language.

2.1. Intelligent Word Recognition

For the conversion of any cursive handwriting document into the digital form, the intelligent word recognition technique is used. IWR recognizes only meaningful words.

For instance, when an offline character recognition system extracts the word "hello" from a document, it will recognize "*h*", "*e*", "*l*", "*l*" and "*o*". IWR will match the letters to dictionary and extract the whole word, "hello" based on neural networks.

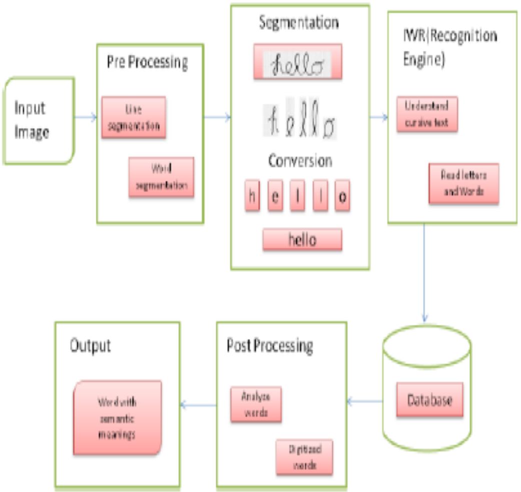

Figure 1. Architectural project of handwriting transformation.

2.2. Motivation

To gather information from clients, a majority of organizations make use of credentials. Usually, these papers are handwritten. Forms, cheques etc. are the examples of such documents. Mutating and stockpiling the documents in digital formats eases the retrieval of information. Manually filling same data into computer is the common practice to handle that information. It would exasperating and tedious to deal with such archives physically. Subsequently need of an exceptional Handwritten Character Recognition Software exudes which could consequently perceive writings from pictures and archives. With the advancement of Handwritten Character Recognition (HCR) Software, it gives no sweat to extract data from handwritten documents and garner it in the form of electronic data. Handwritten documents are used regularly in various sectors such as Banking sectors and Health care industries etc.

HCR systems are finding its usage in newly emerging breadths of handwritten data entry required sectors, such as development of electronic libraries, multimedia database etc.

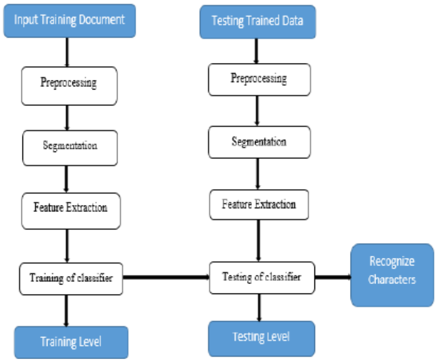

Figure 2. Block diagram of HCR System.

Preparing information and testing information are the two sections wherein the gathered databases are separated. With the emphasis on preparing the framework, preparing information is utilized, further using this prepared framework to perceive test information.

(1). Pre-processing: The arrangement of activities performed on the examined info picture is the thing that we call as pre-handling. The picture is made appropriate for further preparing as it is considerably upgrades the picture. Clamor evacuation, binarization, slant rectification and so forth are some different assignments performed on the picture in pre-handling stage.

(a) Noise removal: It incorporates the utilization of suitable channels in order to expel commotion from the filtered picture. These channels are Smoothing Linear Filter, Order

Statistic Filter and so on. Expulsion of little subtleties from the picture separating enormous items alongside lessening commotion and obscuring of picture is finished utilizing Smoothing.

(b) Binarization: The worldwide thresholding procedure like Ostu's strategy for thresholding is used for the transformation of a dark scale picture into a twofold picture. An ideal estimation of limit is given by Ostu's.

(c) Skew correction: For the best possible further division of the filtered archive, slant is evacuated. Slant redress strategies are executed mandatorily due to the way that transcribed notes are not necessarily adjusted flat splendidly. Cross-connection, projection profile examination, closest neighbor bunching, hough changes, piece-wise covering by parallelogram and so on are a few models.

(2) Segmentation: The atomization of the picture into sub-pictures of sole characters, is done in division arrange. It incorporates:

- Line division for example division of line from section.

- Word division for example division of word from line.

- Character recognition for example separation of character from words. For cursive word recognition, character division is performed if division based technique is embraced.

(3) Feature Extraction: This stage means to separate those the highlights of the characters which are cardinal for their order at acknowledgment organize. This stage is viewed as the most vital stage as on its effective activity, the acknowledgment rate ameliorates other than the reduction in misclassification. With the extraction of angles, for example, paired highlights, directional highlights and so forth. Include vector is made. Beneath referenced are a few classifications amongst which highlight extraction falls:

- Statistical highlights: This component lays on the hypothesis and theory of likelihood. Variety recorded as a hard copy styles are recognized utilizing factual dissemination of pixels of a picture. The idea of factual highlights is separated from measurable appropriation of focuses. For instance Projections histogram, intersections, separations, zoning and so on.

- Structural highlights: Besides depicting the structure of picture, basic highlights additionally portrays the geometrical and topological properties of character, for example, up, down, left and right projection profiles, branches, circles, stroke width, stroke length, crossing focuses and so forth.

- Global change highlights: Ample delineation of the state of the picture is portrayed by Global Transformation based highlights. . It is spatial space to recurrence area interpretation of picture. It can go along vitality smallness because of its capacity to cherish data contained a total picture into couple of coefficients. Different kinds of worldwide change based highlights are: Discrete Cosine Transform, Discrete Fourier Transform, Discrete Wavelet Transform.

3. *D*-Classification

Other than being the basic leadership part of recognition framework, the characterization organize similarly uses the highlights refined from the past stage. The element vector is meant as X where $X = (f_1, f_2, \ldots, f_d)$, f indicates highlights, d implies the quantity of highlights separated from character. Characters are effectively ordered into proper class and perceived based on examination highlights. Classifiers depend on two sorts of learning techniques.

- Supervised learning: In request to prepare a model, preparing information with right detail of class is applied in administered learning. For the testing of information for legitimate order, this model is utilized. Both information and wanted outcomes are incorporated into preparing information. On continuing learning process and dependent on this learning, this model groups test information. For instance: SVM, HMM and so forth.

- Unsupervised learning: In solo learning model isn't given preparing information. Learning isn't required in this. Test information is being ordered on measurable properties and by their spatial gathering and considering their closest neighbor. For instance: *k* implies, Clustering, and so forth.

4. *E*-Post Processing

In order to perform high level concepts such as semantic analysis, syntax analysis etc. adapted to check the recognized character, dictionary is coupled to the system so as to further dilate the accuracy of recognition. This stage isn't obligatory in HCR framework.

5. Related Work

Improvement and progress of different ways to deal with extraction of content data from the picture and video have been proposed for explicit application. Voluminous research, it isn't easy to model broadly useful frameworks. It is because of the explanation that there are various practical wellsprings of variety when blackmailing content. Concealed from the finished foundation or from the low-differentiate, compound pictures or pictures with dissimilarities in text dimension, style, shading, direction, and alignment. As a result of this variety it is exceptionally hard for the issue to draw proverbially. All in all, content identification techniques are partitioned into three distinct sorts. The first has associated part based system, which depends on the supposition that the content areas have uniform hues and convince certain size, shape, and basic arrangement limitations. For the most part, these strategies are not impertinent when the content has coordinating hues with the foundation. The subsequent one contains the surface based strategies which presumes that the content regions have excellent surface. Be that as it may, these methods are comparably less thoughtful to foundation hues, they may not group the writings from the content like foundations. The last one has the edge-based methods. The content districts are recognized under the assumption that the edge of the foundation and the article areas are inadequate than those of the content locales. However, these sorts of methodologies are not fruitful in recognizing writings with enormous text dimension. thought about the Support Vector Machines (SVM) based technique with the multilayer perceptrons (MLP) based one for content check more than four autonomous highlights, in particular, the separation guide include, the grayscale spatial subordinate component, the steady slope fluctuation attribute and the DCT coefficients qualities. They presumed that prevalent discovery results are gotten by SVM as opposed to by MLP. Multi-goals based content location strategies are regularly embraced to distinguish messages on various scales. Writings with various scales would have unique: -

VIDEO - > EDGE EXTRACTION - > SEGMENTATION - > TEXT CLASSIFICATION - > WRITING EXTRACTION - .> OCR

6. Conclusion

In this paper we have examined a few methodologies that were utilized for character recognition to be specific, cursive stroke succession system, neural system, division calculation and canny word acknowledgment. Gigantic work and research have been done in the manually written separate character acknowledgment. Till date all out exactness has not been accomplished which furnishes us with domain of extra exertion toward this path. Various characters furnish us with predominant precision yet word acknowledgment is bombastic with non-indistinguishable composition style. Comprehensive strategy cancels the entangled division yet it uses limited jargon. Division based system because of its multifaceted nature has slighter exactness. Satisfactory precision has been found in the classifier where extent of words is limited to fix numbers as it needs to manage confined number of variety.

References

[1] S. Joseph James, C. Lakshmi, Uday Kiran and P. Parthiban, An Efficient Offline Hand Written Character Recognition utilizing CNN and Xgboost, International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-8 Issue-6, April 2019.

[2] Zheheng Rao, Chunyan Zeng, Minghu Wu, Zhifeng Wang, Nan Zhao, Min Liu and Xiangkui Wan, Research on a written by hand character acknowledgment calculation dependent on an all-encompassing nonlinear bit remaining system, Ksii Transactions On Internet And Information Systems VOL. 12, NO. 1, January 2018.

[3] D. T. Mane and U. V. Kulkarni, Visualizing and Understanding Customized Convolutional Neural Network for Recognition of Handwritten Marathi Numerals, International Conference on Computational Intelligence and Data Science, ISSN: 132 (2018), 1123-1137, (ICCIDS 2018).

[4] Chowdhury Md Mizan, Tridib Chakra borty and Suparna Karmakar Text Recognition using Image Processing, International Journal of Advanced Research in Computer Science, Volume 8, No. 5, ISSN No. 0976-5697, May – June 2017.

[5] Usha Yadav and Satya Verma A deep learning based character recognition system from multimedia document, International Conference on Innovations in Power and Advanced Computing Technologies [i-PACT2017].

[6] Ravneet Kaur, Handwriting Recognition of Gurmukhi Script: A Survey of Online and Offline Techniques, International Journal of Computer Trends and Technology (IJCTT) – Volume 49 Number 1 July 2017.

[7] Jaswinder Kaur and Mrs. Rupinder Kaur, Review of the Character Recognition System Process and Optical Character Recognition Approach" International Journal of Computer Science and Mobile Computing, Vol.6 Issue.5, pg. 45-49, May- 2017.

[8] Michael Reynaldo Phangtriastu, Jeklin Harefa and Dian Felita Tanoto Comparison between neural network and support vector machine in optical character recognition, 2nd International Conference on Computer Science and Computational Intelligence 2017, ICCSCI 2017, 13-14 October 2017.

[9] Manoj Sonkusare and Narendra Sahu A survey on handwritten character recognition (HCR) techniques for english alphabets, Advances in Vision Computing: An International Journal (AVC) Vol.3, No.1, and March 2016.

[10] Amit Verma and Gagandeep Kaur, A review of character recognition from handwritten document, International Journal of Applied Engineering Research ISSN 0973-4562 Volume 10, pp 37580-37584, Number 16 (2015).

[11] Amit Verma and Gagandeep Kaur, Character recognition from handwritten documents using neural networks, International Journal of Applied Engineering Research ISSN 0973-4562 Volume 10, pp 37574-37579, Number 16 (2015).

[12] Amit Verma and Gagandeep Kaur, A Comparative Analysis of Back Propagation and Random Forest Algorithm for Character Recognition from Handwritten Document, Journal of Computer Science and Applications, ISSN 2231-1270 Volume 7, pp. 59-66, Number 1 (2015).

[13] Amit Verma and Gagandeep Kaur, Character recognition from handwritten documents, 2nd National Conference on Advances in Computer Science, Communication Engineering and Applications (ACSCEA-), pp 187-191, (2015).

[14] Monica Patel and Shital P. Thakkar, Handwritten Character Recognition in English: A Survey International Journal of Advanced Research in Computer and Communication Engineering, ISSN (Online) 2278-1021, ISSN (Print) 2319-5940 , Vol. 4, Issue 2, February 2015.

[15] Ramandeep Kaur and Shruti Gujral, Recognition of Similar Shaped Isolated Handwritten Gurumukhi Characters Using Machine Learning, 5th International Conference-Confluence The Next Generation Information Technology Summit (Confluence), 2014 IEEE.

[16] D. Kavitha and P. Shamini, Department of Computer Applications, Easwari Engineering College, Chennai and Tamilnadu, Handwritten Document into Digitized Text Using Segmentation Algorithm, Special Issue, 4th National Conference on Advanced Computing, Applications & Technologies, May 2014.

[17] Nisha Vasudeva, Hem Jyotsana Parashar and Singh Vijendra, Offline Character Recognition System Using Artificial Neural Network, International Journal of Machine Learning and Computing 2(4) (2012).

[18] C. Pornpanomchai, V. Wongsawangtham, S. Jeungudomporn and N. Chatsumpun, Thai Handwritten Character Recognition by Genetic Algorithm (THCRGA), IJET 2011 Vol. 3 (2): 148-153 ISSN: 1793-8244.

[19] Subhash Panwar and Neeta Nain, Cursive Stroke Sequencing for Handwritten Text Documents Recognition, Department of Computer Engineering Malaviya National Institute of Technology, Jaipur.

[20] Chhaya Sambhaji Gochade and R. C. Thool, Handwriting Recognition Using Neural Networking, International Journal of Computer, Information Technology & Bioinformatics (IJCITB) ISSN: 2278-7593, Volume-1, Issue-4 IEEE & IEEE 1st International Conference on Robust Technology for Human Identification.