IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 3, November 2022
Impact Factor: 6.252

A Review on Multi Model Recognition and Mining Alphabets Identification using NLP

Mayur Sakhare[1], Vishal Sonawane[2], Rushikesh Bhalerao[3] , Nikita Karpe[4] , Prof.Niranjan Bhale[5]
Students , Department of Information Technology[1,2,3,4]
HOD, Department of Information Technology[5]
Matoshri College of Engineering and Research Centre, Nashik, Maharashtra, India

Abstract: *Optical character recognition, usually abbreviated to OCR, is the electronic conversion of scanned or photographed images of typewritten or printed text into machine encoded / computer - readable text. It is widely used as a form of data entry from some sort of original paper data source, whether passport documents, invoices, bank statement, receipts, business card, mail, or any number of printed records. It is a common method of digitizing printed texts so that they can be electronically edited, searched, stored more compactly, displayed on-line, and used in machine processes such as machine translation, text-to-speech, key data extraction and text mining. OCR is a field of research in pattern recognition, Artificial Intelligence and Computer Vision. Optical Character Recognition (OCR) is the electronic translation of handwritten, typewritten or printed text into machine translated images. It is widely used to recognize and search text from electronic documents or to publish the text on a website. In our proposed methodology we developed our system on a desktop system.*

Keywords: OCR, python, AI, Image processing, NLP, Photo, image, character.

I. INTRODUCTION

The optical character recognition (OCR) technology converts many types of documents (PDF, BMP, TIFF, JPEG, PNG) into machine-readable text. It has become one of the most notable technological applications in the fields of Artificial Intelligence and Pattern Recognition. In contrast to the human brain, which can easily distinguish characters and language from an image, machines are still a long way from being able to understand the information contained in a picture. As a result, numerous research projects have been launched in an attempt to efficiently transform a document image to a machine-readable format .OCR is a method for converting checked images into editable content. OCR is the electrical conversion of handwritten images or typewritten material into machine-editable content. Many antithetic types of Optical Character Recognition OCR implementations are commercially available today, and it is a field of study in Pattern Recognition, Artificial Reasoning, and Machine Visual Perception. OCR systems have become one of technology's most successful uses. OCR technology allows us to convert many sorts of documents, such as scanned paper documents, pdf files, or photos acquired by a digital camera, into editable and searchable data [1].

OCR is also known as Text Recognition. The particular mechanisms that allow individuals to notice things are still unknown, but three very basic characteristics that scientists are currently aware of that aid in recognition are integrity, purpose, and adaptability. OCR programmers have grown with the development of Hardware and Software that are used to convert data from visual to computer-editable text. Advanced processing is handled by software, while hardware, such as optical scanners, is utilized to copy or read text from visual text. Software utilizes cutting-edge technology such as Artificial Intelligence (AI) to produce accurate and advanced character recognition (ICR) techniques. The programme first determines the format of the provided image, and then the hardware reads the data in the image. The data is then divided into pages, which are further divided into blocks of text, tables, graphics, and so on, which are further divided into words and characters [2].

Optical character recognition (OCR) is a system that converts input text into machine encoded format. Today, OCR is helping not only in digitizing the handwritten medieval manuscripts, but also helps in converting the typewritten documents into digital form. Organizations are satisfying the needs Of digital preservation of historic data, law documents educational persistence etc.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 3, November 2022
Impact Factor: 6.252

An OCR system depends mainly, on the extraction of features and discrimination of these features (based on patterns). Handwritten OCR have received increasing attention as a subbed of OCR. It is further categorized into offline system, and online system based on the associate editor coordinating the review of this manuscript and approving it for publication. The offline system is a static system in which input data is in the form of scanned images while in online systems nature of input ismore dynamic and is based on the movement of pen tip having certain velocity, projection angle, position and locus point. Therefore, an online system is considered more complex and advance, as it resolves the overlapping problem of input data that is present in the offline system.

The main concept in automatic recognition of patterns is first to teach the machine which class of patterns that may occur and what they look like [3]. In OCR patterns are letters, numbers and some special symbols like commas, question marks as well as different characters .The teaching of machine is performed by showing machine examples of characters of all different classes. Based on these examples the machine builds prototype or description of each class of characters. During recognition the unknown characters are compared to previously obtained descriptions and assigned to class that gives the best match. In most commercial systems for character recognition training process is performed in advance. Some systems however include facilities for training in the case of inclusion of new classes of characters. A typical OCR system consists of several components [4]. The first step is to digitize analog document using an optical scanner. When regions containing text are located each symbol is extracted through segmentation process. The extracted symbols are pre-processed, eliminating noise to facilitate feature extraction. The identity of each symbol is found by comparing extracted features with descriptions of symbol classes obtained through a previous learning phase. Finally contextual information is used to reconstruct words and numbers of the original text. These steps are briefly presented here. Interested readers can refer for more elaborate discussion of OCR system components [1].

Character Recognition is a subset of Pattern Recognition area. Several concepts and techniques in to replicate human functions by machines and making the machine perform common tasks like reading is an ancient dream. The origins of character recognition dates back to 1870 when C.R. Carey of Boston Massachusetts [5] invented retina scanner which was an image transmission system using a mosaic of photocells. Early versions needed to be trained with images of each character and worked on one font at a time

II. LITERATURE SURVEY

In “[6]”, proposed algorithm could be damaged as a kernel utilized for solitary alphabet finding within an entire OCR description method with no requirement for any compound mathematical operations. The highlights of this approach are that, it doesn’t employ any databases or libraries of picture matrices to differentiate alphabets, but it has a special algorithm to differentiate alphabets in its position. The enthusiasm for the improvement of that algorithm was the straightforward actuality that English alphabets are permanent glyphs and they shall not be altered forever. Caused by this reality, procedure of non-expected neural networks and vector-based information education supply approximately perfect outcomes, but these are performing many outmoded work.

In “[7]”, a license number plate is considered as the exclusive acknowledgment of a medium, which builds the license number plate detection a crucial operation in intellectual transportation systems (ITS). A multifilter dependent LPR architecture is suggested in that paper to effectively identify license plates, distinguish and understand alphabets in a database. The key advantage of the system is the potential to merge several filters to guarantee that the missed figure is near to zero with high versatility in localization. In particular, filters with different unique approaches may be inserted or altered depending on the physical world.

In “[8]”, Writer provides a well-organized algorithm for habitual recognition of any license plate, through an emphasis on the Lebanese license plates wherever a few of their features have been exploited well to decrease the identification mistakes. The planned algorithm has been implemented through the illustration Processing Toolbox during MATLAB R2013b (8.2.0.701). Their experiments exhibit to the detection inaccuracies have been focused well upon developing the actuality that the Lebanese license shields are printed in two designs. Also, their method has an alternative to profit from the corporation of the license shields in the front together with the back end of the car to develop the appearance

In “[9]”, author presented alphabetical extraction technique in usual scene images. Along with expand alphabet recognition scheme using alphabetical extraction technique in natural scene images. A process of graph matching using letter structured alphabetically. It is technique of identification to promote analysis of position relative and systemic

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 3, November 2022
Impact Factor: 6.252

connection. Shift in text or rotation is stable technique. In planning to show it; they encountered two occasions where teaching font and check font are similar cases or separating cases.

In "[10]", they evaluate a fresh alphabet recognition technique of certify plate numeral based on similar BP neural networks. This enhanced the correctness of the identification scheme that aims to understand writing repeatedly the Chinese license plate. In the planned method, the quality is binarized with the sound is eliminate in the pre-processing stage, then the alphabet alphabetical is extract with frame by means of the alphabet is normalize to size 8*16 pixels. As a final point, the alphabet attribute is place addicted to the similar neural network and the alphabet is documented. The anticipated technique in alphabet recognition is effectual, and hopeful grades have been obtained in experiment on Chinese certify plates. In "[11]", they tried to know printed and handwritten alphabets with projecting them on dissimilar sized grids (5x7, 7x11, and 9x13). The consequences showed that the accuracy of the quality recognition depends on the resolution of the alphabet projection. As well, they realized that not each inscription approach could be expected with the similar network through the identical accuracy. This shows that the diversity of individual handwriting habits couldn't be completely enclosed through single neural network.

In "[12]", authors analyzed programmed learning systems utilized to extract functional information from musical scripts, participate a major role in optical music recognition. Visual music acknowledgment or OMR has been extensively utilized to take out the melodious notations and understanding as of old scripts and thus encloses lot of significance in retrieving chronological data. The paper designed fresh automated music recognition systems that are capable of being used efficiently to identify sheet or written scores in a playable environment under unlike blur situations, differences in description, noise etc. The mentioned scheme is capable of extracting ordinary images as well as images from older scripts through noise, images as well.

In "[13]" They researched the topic of improving the productivity of common; an ALPR program consists of four step of dispensation. When selecting the ALPR scheme camera, a range of points need to be remembered in the picture gaining process, such as camera movement and shutter speed.

In "[14]" a license plate dataset consisting of 141 photographs was used in the work. Consider that the dataset is not homogeneous, as it includes photographs in varying dimensions, aspect ratios, background material, plate sizes, lighting situations, camera angles, tilts, pans, etc. Much of the photographs used were collected in Canada-Ontario and thus the dataset mainly includes Ontario plates. An aspect ratio of 1.33 is retained to normalize the input images, and all images are sized to 1024x768 pixels. In pictures of various aspect ratios, either the height or the width is calibrated to preserve a defined aspect ratio.

In "[15]", that paper presents, a variety of milestones are often accomplished through handwriting recognition, primarily for unavailable hand written typescripts and phrases. The newly implemented PDAs haw sensible meaning, for example, in the online scenario. Correspondingly, over the last few years, some electronic signature confirmation program has been on the market and educational tools to help children learn writing are beginning to emerge. Generally offline victories have arrived in inhibited fields, for instance postal addresses, bank cheques, and study structures. The assessment of authorize with complex layouts, credit of degraded printed text, and the detection of running handwriting maintain to stay mainly in the investigate arena

In "[16]" visual quality detection refers toward a procedure whereby printed credentials be distorted into ASCII records intended for the reason of compact storage, editing, quick recovery, along with other file manipulations during the exploit of a processor. The appreciation phase of an OCR procedure is prepared hard by additional noise, image deformation, in addition to the variety of alphabet typefaces, size, with fonts to a text could contain. During to learn a neural network come near is introduce to execute elevated correctness detection on multi-dimension along with multi-basis typescript; a fresh centered hesitant preparation procedure through a short noise-sensitivity normalization method is old to attain elevated correctness grades. The next part trades accuracy for supplementary font and size potential, and a better two-encrusted neural system is taught to be familiar with the complete set of 94 ASCII quality descriptions for every point size beginning S to 32 along with for 12 frequently utilized fonts. The arrangement of these two systems is estimated founded on an evidence of supplementary than discrete million alphabet depictions from the harsh dataset.

In "[17]", that papers present a straightforward, well-organized, and reasonable approaches to create OCR for interpretation some manuscript that has fix font size and method or handwritten style. To full capability and less calculation rates, OCR in along the purpose of manuscript employs database to identify English alphabets which

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 3, November 2022
Impact Factor: 6.252

constructs this OCR system extremely straightforward to handle. This paper tells about OCR system intended for offline handwritten alphabet detection. The system has the ability to yield excellent results. Predispensation technique utilized in document images as a first step in alphabet acknowledgment systems was obtainable. This scheme suggests a higher boundary through having an improvement specifically its extensibility, specifically while it is organized to interpret a pre described set of article configures, at present English articles; it could be organized to identify original categories.

III. CONCLUSION

OCR is known to be the recognition of certain letters to the branch of computer sciences and to use this certain technology to differentiate the printed and handwritten letters in digital images of the displayed text. The fundamental process of an OCR is to examine the type of text and text present in a document and translate the letters in the document into code that can be used to process data. This proposed methodology describes the OCR system functioning by breaking it down into steps, which helps in identifying the loops in the current method of character recognition. The system is capable of giving some really commendable results.

Pattern recognition done using neural network can tolerate noise and trained properly, can be used to recognize unknown patterns. Neural networks if constructed using proper architecture and trained correctly can be used for different scientific and commercial applications. They can be used for data entry, text entry, data automation etc.

Various image processing algorithms were studied and tested and the algorithm that supplied the best result is included in the architecture and discussed. Segmentation is one of the important processes in image processing. Isolating a single character from a bunch of characters can be achieved using classical approach, recognition based or holistic approach.

A neural network using back-propagation algorithm is one of the most popular algorithms for training. It is a time-consuming algorithm for training network with a large number of nodes. Adjusting the size of the input, error margin and addition of hidden node will give a better result.

REFERENCES

[1]. Chaudhuri, K. Mandaviya, P. Badelia, and S. K. Ghosh, Optical character recognition systems, vol. 352. 2017. doi: 10.1007/978-3-319-50252-6_2.

[2]. P. Divya et al., "Web based optical character recognition application using flask and tesseract," Mater. Today Proc., no. xxxx, 2021, doi: 10.1016/j.matpr.2020.10.850

[3]. A. T. Sahlol, C. Y. Suen, H. M. Zawbaa, A. E. Hassanien, and M. A. Elfattah, "Bioinspired BAT optimization algorithm for handwritten Arabic characters recognition," 2016 IEEE Congr. Evol. Comput. CEC 2016, pp. 1749–1756, 2016, doi: 10.1109/CEC.2016.7744000.

[4]. A. Dobroczoni, R. Takacs, B. M. Cermak, and Shchokin, "Design Of Machines And Structures," vol. 18, no. 1, pp. 2–4, 2014.

[5]. H. Singh and A. Sachan, "A Proposed Approach for Character Recognition Using Document Analysis with OCR," Proc. 2nd Int. Conf. Intell. Comput. Control Syst. ICICCS 2018, no. Iciccs, pp. 190–195, 2019, doi: 10.1109/ICCONS.2018.8663011.

[6]. Sushruth Shastry, Gunasheela G, Thejus Dutt, Vinay D S and Sudhir Rao Rupanagudi, "i" - A novel algorithm for Optical Alphabet Recognition (OCR). 978-1-4673-5090- 7/13/$31.00 ©2013 IEEE.

[7]. Lulu Zhang, Xingmin Shi, Yingjie Xia, Kuang Mao, "A Multi-filter Based License Plate Localization and Recognition Framework". 978-1-4673-4714-3/13/$31.00 ©2013 IEEE.

[8]. Ibrahim El Khatib, Yousef Samir-Mohamad Omar, and Ali Al Ghouwayel, "AN EFFICIENT ALGORITHM FOR AUTOMATIC RECOGNITION OF THE LEBANESE CAR LICENSE PLATE. ISBN: 978-1-4799-5680-7/15/$31.00 ©2015 IEEE.

[9]. Jieun Kim, and Ho-sub Yoon "Graph Matching Method for Alphabet Recognition in Natural Scene Images. 978-1-4244-8956- 5/11/$26.00 ©2011 IEEE

[10]. Feng Yanga，and Fan Yangb, "Alphabet Recognition Using Parallel BP Neural Network". 978-14244-1724-7/08/$25.00©2008IEEE.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 3, November 2022
Impact Factor: 6.252

[11]. Rókus Arnold, and Póth Miklós "Alphabet Recognition Using Neural Networks. 11th IEEE International Symposium on Computational Intelligence and Informatics • 18– 20 November, 2010 • Budapest, Hungary.

[12]. Amarjot Singh , Ketan Bacchuwar , Akash Choubey, and Devinder Kumar , "An OMR Based Automatic Music Player". 978- 1-61284-840-2/11/$26.00 ©2011 IEEE.

[13]. Shan Du, Member, IEEE, Mahmoud Ibrahim, Mohamed Shehata, Senior Member, IEEE, and Wael Badawy, Senior Member, IEEE "Automatic License Plate Recognition (ALPR):A State-ofthe-Art Review, IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, VOL. 23, NO. 2, FEBRUARY 2013.

[14]. Imran Shafiq Ahmad, Boubakeur Boufama, Pejman Habashi, William Anderson and Tarik Elamsy, "Automatic License Plate Recognition.

[15]. A Comparative Study". 2015 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT).

[16]. Rejean Plamondon, Fellow, IEEE, and Sargur N. Srihari, Fellow, IEEE, "On-Line and Off-Line Handwriting Recognition:A Comprehensive Survey 1EEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE. VOL. 22, NO.

[17]. Hadar I. Avi-Itzhak, Thanh A. Diep, and Harry Garland, "High Accuracy Optical Alphabet Recognition Using Neural Networks with Centroid Dithering IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, VOL. 17, NO. 2, FEBRUARY 1995.