IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 9, May 2023

Multi Model Recognition and Mining Alphabet Identification using NLP

Mayur Sakhare[1], Rushikesh Bhalerao[2], Vishal Sonawane[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 Center, Nashik, Maharashtra, India

Abstract: *Optical character recognition (OCR), usually abbreviated to OCR, is the mechanical or electronic conversion of scanned or photographed images of typewritten or printed text into machine-encoded or computer-readable text. It is widely used as a data entry method for various original paper data sources, including passport documents, invoices, bank statements, receipts, business cards, mail, and other printed records. OCR serves as a common technique for digitizing printed texts, enabling electronic editing, efficient storage, online display, and utilization in machine processes such as machine translation, text-to-speech conversion, key data extraction, and text mining. OCR is a field of research encompassing pattern recognition, artificial intelligence, and computer vision. Optical Character Recognition or OCR is also used for the electronic translation of handwritten, typewritten, or printed text into machine-translated images. It finds widespread application in recognizing and searching text from electronic documents or publishing text on websites. In our proposed methodology, we developed our system on a Windows 11 PC, utilizing PYTHON as the frontend software*

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

I. INTRODUCTION

Optical character recognition (OCR) technology has emerged as a significant advancement in the fields of artificial intelligence and pattern recognition. It enables the conversion of various document types, including PDF, BMP, TIFF, JPEG, and PNG, into machine-readable text. While humans possess the innate ability to distinguish characters and language from images, machines still face challenges in comprehending information from pictures. To bridge this gap, extensive research projects have been undertaken to effectively transform document images into machine-readable formats. OCR serves as a method for converting scanned images into editable content, whether they are handwritten or typewritten. Commercially available OCR implementations offer diverse approaches to character recognition, and the field itself encompasses the study of pattern recognition, artificial intelligence, and machine vision. OCR systems have gained remarkable success as a technological application, enabling the conversion of scanned paper documents, PDF files, or digital camera-acquired photos into editable and searchable data [1].

OCR, often known as text recognition, is a technology that converts visual input into computer-editable text using hardware and software. To achieve accurate character recognition, software leverages advanced technologies such as artificial intelligence. Determine the image format, read the data with hardware, divide it into pages, blocks of text, and then into words and characters [2]. OCR is critical in digitizing handwritten manuscripts and typewritten documents, which helps with digital preservation initiatives. Handwritten OCR, which can be classified as offline or online depending on the input data, has gained popularity. Online systems handle dynamic input based on pen motions, whereas offline systems deal with scanned images. Automatic pattern recognition is the process of teaching machines to recognize patterns like letters, numbers, and symbols by supplying samples of each class [3]. Optical scanning, segmentation, feature extraction, character identification, and contextual reconstruction are common phases in OCR systems [4]. Character recognitionis a subset of pattern recognition, and the desire to replicate human capabilities through machines has a long history, reaching back to the 1870s [5].

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 9, May 2023

The demand for more advanced and reliable techniques of alphabet recognition and analysis motivates research in the field of multi-model recognition and mining alphabets identification using NLP. Traditionally, optical character recognition (OCR) systems concentrated on translating scanned or photographed text into machine-readable format. However, alphabets occur in a variety of forms, such as handwritten, typewritten, or printed text, posing difficulties for existing OCR systems.

We hope to improve alphabet identification and understanding across multiple input sources by incorporating natural language processing (NLP) tools into the OCR process. NLP can give language analysis and contextual knowledge, allowing for more accurate alphabet identification and overall recognition performance improvement.

Furthermore, the capacity to mine and analyse alphabets from many sources has numerous practical uses. In document processing, for example, efficient alphabet recognition can automate data entry, making it faster and more precise. The capacity to analyse and compare alphabets in different languages or handwriting styles can help linguistic researchers. Furthermore, data analysts can glean significant insights from massive amounts of alphabet-based data, benefiting sectors such as linguistics, education, and information retrieval.

We hope to overcome the constraints of existing OCR systems and explore new possibilities for letter recognition and analysis by designing a system that merges multi-model recognition and mining alphabets identification using NLP. This study has the potential to enhance the state-of-the-art in language processing, contribute to the creation of intelligent systems, and pave the way for practical applications in a variety of disciplines.

II. LITERATURE SURVEY

The technique provided in "[6]" is a revolutionary approach to single alphabet detection within an OCR framework that eliminates the need for sophisticated mathematical computations. This method, in particular, does not rely on image matrix databases or libraries to differentiate alphabets. It instead employs a specialized algorithm that takes into account the fixed nature of English alphabets. This approach produces almost optimal results while minimizing duplicate computations by utilizing non-traditional neural networks and vector-based data training.

The study "[7]"emphasizes the significance of license plate detection in intelligent transportation systems (ITS), as license plates are the fundamental means of identification. To address this, a multifactor dependent license plate recognition (LPR) architecture is developed, with the goal of efficiently identifying license plates while precisely recognizing the alphabets stored inside a database. The method has the distinct advantage of including numerous filters, resulting in fewer missed detections and greater localization flexibility. To adapt to varied real-world events and requirements, the filters can be customized and adjusted using distinct ways.

The author presents a well-structured algorithm for the automatic recognition of license plates in his article "[8]"with an emphasis on Lebanese license plates. To reduce identifying errors, the system makes use of special properties of these plates. The approach is implemented in MATLAB R2013b (8.2.0.701) using the Image Processing Toolbox. The experimental results show that detecting inaccuracies can be successfully targeted due to the fact that Lebanese license plates are produced in two separate designs. Furthermore, the system makes use of both the front and rear license plates of automobiles to improve recognition accuracy.

The author offers a strategy for extracting alphabets from natural scene photos in "[9]" withthe goal of advancing alphabet identification in such circumstances. The suggested methodleverages graph matching and takes advantage of letter structural information foridentification, allowing for the investigation of relative locations and systemic linkages. Theapproach is resistant to text changes or rotations. The authors met two scenarios during theirresearch: cases where the training and test fonts were comparable, and cases where theyneeded to separate various alphabets.

In "[10]"a unique alphabet recognition strategy for license plate numbers is assessed usingBP neural networks. With repeated writing, this method intends to enhance the accuracy of15recognizing Chinese license plates. Binarizing the quality, reducing noise in the pre-processing stage, extracting the alphabetical components, and normalizing them to a size of8*16 pixels are all part of the proposed approach. The alphabet properties are then suppliedinto the neural network for recognition. The anticipated technique has shown to be effective,with promising results obtained in trials on Chinese license plates.

In "[11]", the authors study the recognition of printed and handwritten alphabets byprojecting them onto different grid sizes (5x7, 7x11, and 9x13). The results show that theresolution of the projection affects the accuracy of alphabet

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 9, May 2023

recognition. Furthermore, it hasbeen discovered that not all handwriting styles can be successfully recognized using the sameneural network, emphasizing the need for numerous networks to accommodate variedindividual handwriting patterns.

The focus of "[12]" is on programmed learning systems used for extracting meaningfulinformation from musical scores, which play an important part in optical music recognition(OMR). OMR can be used to retrieve historical data by extracting musical notations andinterpretations from old writings. The research introduces novel automatic musicidentification systems capable of identifying sheet music or written scores under difficult conditions such as blurriness, description variations, and noise. Even in the presence of noiseand other image-related difficulties, the proposed technique is effective at extractinginformation from both regular and older scripts.

In "[13]" the emphasis is on increasing the productivity of an ALPR (Automatic LicensePlate Recognition) programme, which requires four processing steps. Choosing an ALPRcamera necessitates taking into account aspects such as camera movement and shutter spee throughout the picture capture process.

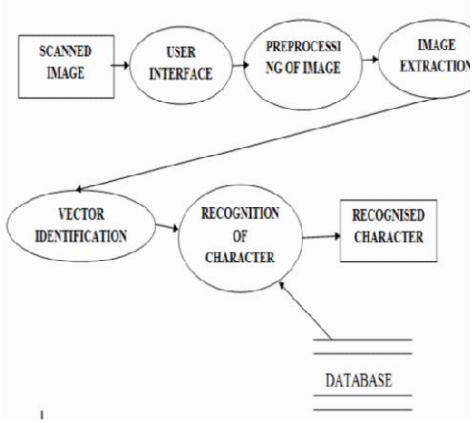
In "[14]" a license plate dataset of 141 photos is used. The dataset is diverse, includingimages of various sizes, aspect ratios, backdrop elements, plate sizes, lighting circumstances,camera angles, tilts, pans, and so on. The photos have been normalized to a 1.33 aspect ratioand shrunk to 1024x768 pixels. To retain the defined aspect ratio, the height or width ofimages with different aspect ratios is changed.

The paper demonstrates handwriting recognition milestones in "[15]" particularly forhandwritten manuscripts and words that are not easily available. With the introduction ofPDAs and electronic signature verification systems, the importance of handwriting16recognition has grown. Offline accomplishments have been obtained in categories such aspostal addresses, bank cheques, and forms, although complicated layouts, damaged printedtext, and cursive handwriting detection are still active study areas.Visual quality detection is covered in "[16]" where printed papers are turned into ASCIIrecords for computer storage, editing, and other manipulations. Noise, image deformation,and changes in alphabet types, sizes, and fonts all provide obstacles to the OCR process. Aneural network approach is used to detect multi-dimensional and multi-basis scripts withexcellent accuracy. A two-tiered neural network system is trained to recognize the entire setof 94 ASCII characters for various point sizes and fonts. A collection of over a millionalphabet images is used to investigate the trade-off between accuracy and font/sizevariability. In "[17]" a simple and cost-effective strategy for developing OCR systems that can analyse fixed font and method or handwritten style manuscripts is provided. The OCR systemrecognizes English alphabets using a database, making it simple to use. The study discusses ahandwritten alphabet detection algorithm that works offline and produces outstanding results.As a first step in alphabet recognition systems, a pre-processing technique is used to improvedocument images. The system's extensibility has enhanced, allowing it to be customized forparsing different categories of specified article formats other than English articles.

III. DATA FLOW DIAGRAM

Level 1 Data Flow Diagram

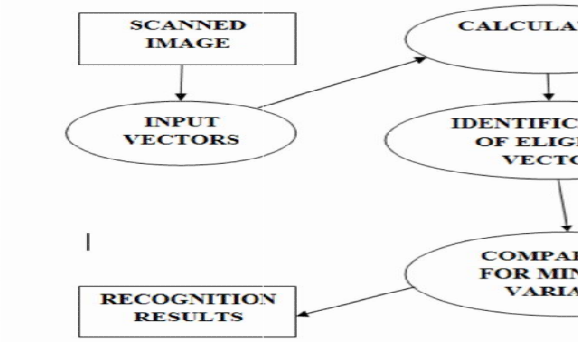
Level 1 DFD breaks down the processes identified in Level 0 into more detailed sub-processes, representing the flow of data between them and any external entities involved.



IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 9, May 2023

Level 2 Data Flow Diagram

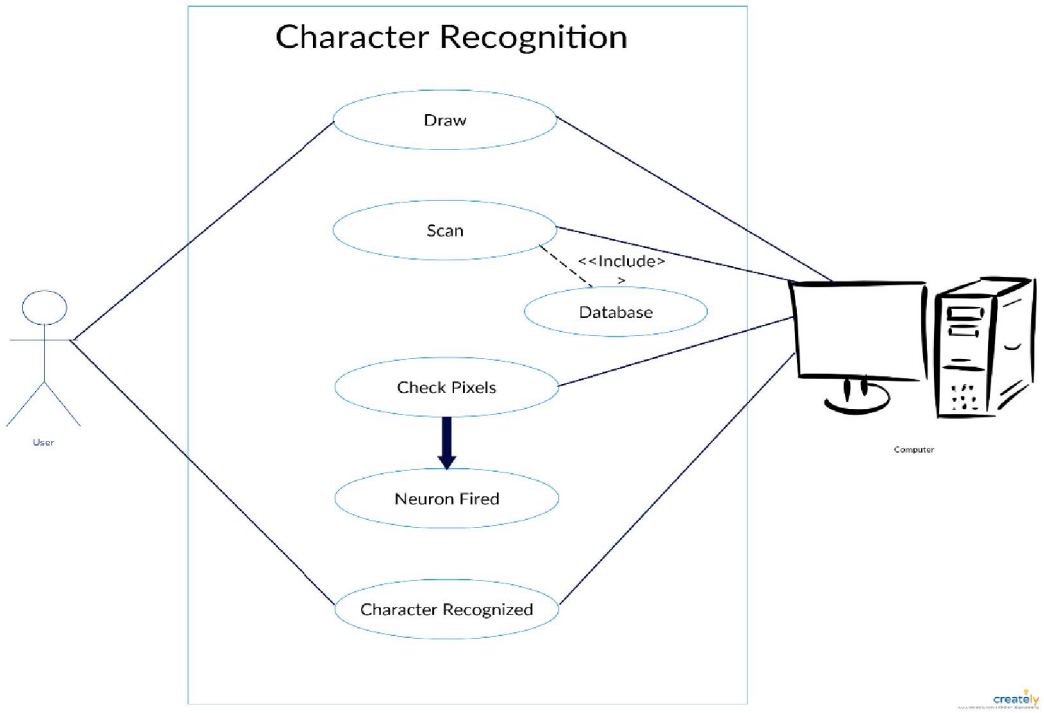
Level 2 DFD further expands the sub-processes of Level 1, capturing additional details and illustrating the data flow within each sub-process, including any data transformations or manipulations that occur.



Entity Relationship Diagrams

Data Description

Data description involves defining and describing the entities and their relationships within a system. Entities represent real-world objects or concepts, and relationships describe the associations between these entities. It also includes specifying the attributes of each entity, their data types, and any constraints or relationships they have with other entities



IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 9, May 2023

IV. CONCLUSION

OCR is recognized as a branch of computer science that distinguishes between printed and handwritten letters in digital images of displayed text. It examines and translates the text in a document into code, allowing for data processing. This proposed methodology breaks down the OCR system into steps, identifying loops in the current character recognition method and producing commendable results. Pattern recognition can effectively handle noise and, when trained properly, recognize unknown patterns. Proposed method has numerous scientific and commercial applications such as data entry, text entry, and automation. The study explored various image processing algorithms with the best results for inclusion in the architecture. Segmentation, an important process in image processing, can be achieved using classical, recognition-based, or holistic approaches to isolate individual characters from a group. The technique is widely used and can be time-consuming when dealing with a large number of nodes. Adjusting input size, error margin, and adding hidden nodes can improve the overall performance and results of the network

FUTURE SCOPE

The future scope of our project includes:

- Improving accuracy, enabling real-time recognition with the phone's camera
- Supporting multiple languages and text translation
- Providing document scanning functionality
- Ensuring continuous improvement through regular updates and user feedback.
- These enhancements aim to enhance accuracy, expand functionality, and provide a more seamless and user-friendly experience for users of the app.

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., 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, "Bio-inspired 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]. Bartz, C., Yang, H. and Meinel, C., "STN-OCR: A single neural network for text detection and text recognition". 2017. arXiv preprint arXiv:1707.08831.

[6]. 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, pp. 190–195, 2019, doi: 10.1109/ICCONS.2018.8663011

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

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

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

[10]. 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, 2015, IEEE.

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

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 9, May 2023

[12]. 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

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

[14]. 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.

[15]. Imran Shafiq Ahmad, Boubakeur Boufama, Pejman Habashi, William Anderson and Tarik Elamsy, "Automatic License Plate Recognition: A Comparative Study". IEEE International Symposium on Signal Processing and Information Technology (ISSPIT), 2015

[16]. Rejean Plamondon, Fellow, IEEE, and Sargur N. Srihari, Fellow, IEEE, "On-Line and Off-Line Handwriting Recognition:A Comprehensive Survey" IEEE Transactions on Pattern Analysis And Machine Learning Intelligence, Vol. 22, No. 1, Jan 2000

[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 Learning Intelligence, Vol. 17, No. 2, Feb 1995