Detection and Recognition of Text from an Image using Computer Vision toolbox in MATLAB

Sejal Wasule
Department of Electronics Engineering, Shri Ramdeobaba College of Engineering and Management, Nagpur 440013, India

Abstract:
Optical character recognition or OCR, is transformation of captured images of printed text containing images to computer accessible text. It has a variety of applications like data acquisition from paper printed data sources like bank passbook, bank statement, business cards, invoices, certificates or any government issued documents like Aadhar card, Driver's license, PAN Card. A lot of them exist digitally on our laptops and phones, but rarely the text on them is accessible. These kinds of documents are physical hard clones having been published or handwritten for illustration. This highlights an important problem and brings with it a unique challenge — while the information in digital documents can fluently be read manually, the text in them cannot be accessed digitally (Copied or pasted). If we are to reproduce the text on these documents in digital form, we would have to type the information manually. Taking a print of a document with a camera or a scanner does not really help because this simply produces an image that a computer can display, reproduce or edit, but not understand. In order to allow computer to make sense of information the image, the device needs the capability to look at it and fetch the abstract shapes we call letters and figures in short; it would need to be suitable to read and reuse the information intelligently.
Keywords: OCR, text, documents, data, information, computer, letters.

I. Introduction:

Throughout the history attempts to develop a computing system with the power to recognize, extract and display the text from pictures containing text have been made. In this age where everything is driven by data, there's a huge demand for storing information in accessible ways like digitally written documents to later re-utilize this information. One simple way to store information to the computing system from the paper documents may well be to 1st scan the documents then store them as image files. But, text in those cannot be accessed digitally. Optical Character Recognition is the method by which conversion of any style of text-containing documents having handwritten, printed or scanned text, into editable digital format can be possible. This technology allows a machine to acknowledge text in such documents and enables to turn the simple image files, into absolutely searchable and editable documents with text content recognized by computers. Extraction of the relevant information rather than the standard approach of manually retyping the text from documents is less laborious process. This technique saves a huge amount of time, money and labor with its variety of applications in sectors like banking, business, Health, education, Transport etc. OCR is handy when it comes to real life applications like vehicle number plate recognition and extraction, text to speech conversion, text translation, transcription of medical and pharmaceutical labels etc. For good quality and high accuracy character recognition, OCR techniques expect top quality or high-resolution pictures with some basic structural properties like high differentiating text and background. The manner pictures square measure generated is a crucial consideration of the accuracy and success of

OCR, since this usually affects the standard of pictures dramatically. Usually, OCR with pictures created by scanners provides high accuracy and good performance. In distinction, pictures created by cameras usually aren't nearly as good as input scanned pictures to be used for OCR because of the environmental or camera connected factors. Numerous errors may occur. Our objective was to create such an algorithm that can minimize noise from environmental factors and accurately recognize and display the text.

II. Literature survey:

Till now, various approaches have been made to make OCR as simple and accurate as possible. In this section we shall see the various OCR techniques that are discovered and their variety of applications available in the market.

In [1] the authors used deep learning and reasoning to detect the address and whereabouts of candidates on the captured or scanned pages of documents to assign the correct location details for each individual. Primarily, the standard algorithms of Binarization were used for preprocessing for removing noise, after that they used machine learning for segmentation, text recognition, and Identification of Address Candidates. At last, they used deep learning and reasoning with Rule-based Address-API to Extract and validate the address data. In paper [2] authors discussed about a Sanskrit verse meter (Chanda) identification and utilization system based on web. Majority of Sanskrit literature is in the form of poetry which is studied in meters called as 'Chandas'. The analysis of text files of these Chandas is done and a useful aid for digitization of Sanskrit texts using the methodology of post-OCR manual correction is discussed. Paper [3] explored the development in OCR technology from 1876 till now. The different methodologies used in different times and the accuracies were compared. Paper [4] discusses various applications of OCR and different methodologies through which optimum accuracy and precision can be achieved for OCR. In Paper [5] authors used MATLAB. They used dataset of Arabic characters with 261 features. They performed the preprocessing, feature extraction first then selection was done by the algorithm Hybrid whale optimization, and the last step classification was done with python. 96% accuracy was achieved. In paper [6] the method proposed was implementing only CNN for OCR. CNN or convolutional neural network has variety of applications including image classification. They first obtained sample written text from different people having different handwritings. The training dataset includes 1500 samples of each Devanagari character and 250 samples of each English character. This was the dataset used to train CNN. Classification of Devanagari characters might have been tricky because of so many features, intricacies of characters and uniqueness of penmanship by each individual. The accuracy from this method was 91%. In Paper [7] datasets used were hand written. Deep neural networks (DNNs) were used for classification and segmentation of characters. It was applied as tutoring applications for children. In paper [8] the system for classifying the CV and resumes of applicants for the suitable job is discussed. Here first the documents are processed and text is extracted from them using various OCR techniques with accuracy up to 97%. Then this extracted text is processed by removal of stop words, punctuations and lemmatization and analyzed using NLP. Next this extracted text is fed as an input to the unsupervised machine learning algorithm, Latent Dirichlet Allocation (LDA), for classification and matching the resumes with various job descriptions in various sectors. In [9] this paper the application of OCR for the benefit of the visually impaired people is given. Here they proposed the hardware system i.e. wearable smart reader spectacle with a micro

camera. In its assistance a raspberry pi system containing OCR and Text to speech (TTS) algorithms with a separate portable rechargeable power bank arrangement is provided. Firstly, the system will perform preprocessing on input images through camera like deblurring, motion sensing and pixel correction etc. Then, the good quality image is fed as an input to Raspberry pi through which text is extracted from image and with the aid of text-to-speech (TTS) engines the audio output can be finally heard through speakers of headphones or earphones. In paper [10] 3 different datasets were used to give a new method which changed many problems and perspectives of viewing them. This model assures better results than traditional OCR algorithms. In survey paper [10] authors focused on the work done to for recognition of various Indian scripts. Most of the Indian scripts like Marathi, Hindi, Gujarati, Malayalam, Bangla, Tamil etc. having more than 500 Characters and Symbols each sounding unique. They proposed to study consonant-vowel combinations in sentences to overcome the errors while recognition of characters. They discussed about the development in OCR approaches and overcoming the difficulties in recognizing and extracting Indian scripts in detail. [11] In this paper the application of OCR in recognition of hateful and derogatory content and cyberbullying on social media is discussed. Multimodal deep framework which used deep learning models (like BERT and Electra), computer vision, OCR and Natural language processing are used for comprehending the complex memes in various Indian languages and classifying them as good or hateful content. They created their own dataset which contained memes on Facebook. In paper [12] application of OCR for the vehicle number plate recognition is discussed. For the classification of characters on the number plate they trained a Convolutional neural network and tested the network with different datasets. In paper [13] the techniques like machine learning and genetic algorithm by which OCR is done are Compared. In genetic Algorithm patterns or chromosomes are selected and initialized and their fitness is calculated. Now randomly selection of genes and their crossover is done. Again, fitness is calculated after mutation and whole process goes in loop until we get best fitness score. Paper [14] discussed about application of OCR in bus attendance and tracking system. The author proposed to use faster RCNN for detection and classification. The database is SQLite. Steps include segmentation and localization of number plate, then extraction of data, and entering the data in system. In paper [15] The authors proposed a system which will read the medical reports using Image processing and OCR and analyze them using NLP. The system will also help processing processing the reports and predicting the stage of diabetes, type of diabetes probability of contracting diabetes to future generations using machine learning models like linear and logistic regression etc.

In this section we saw various approaches for OCR and how inventors and researchers applied that with various degrees of successes in the fields like Medicine, Transport and trade, Business, Social welfare, Education etc. with numerous advanced approaches.

III. Methodology:

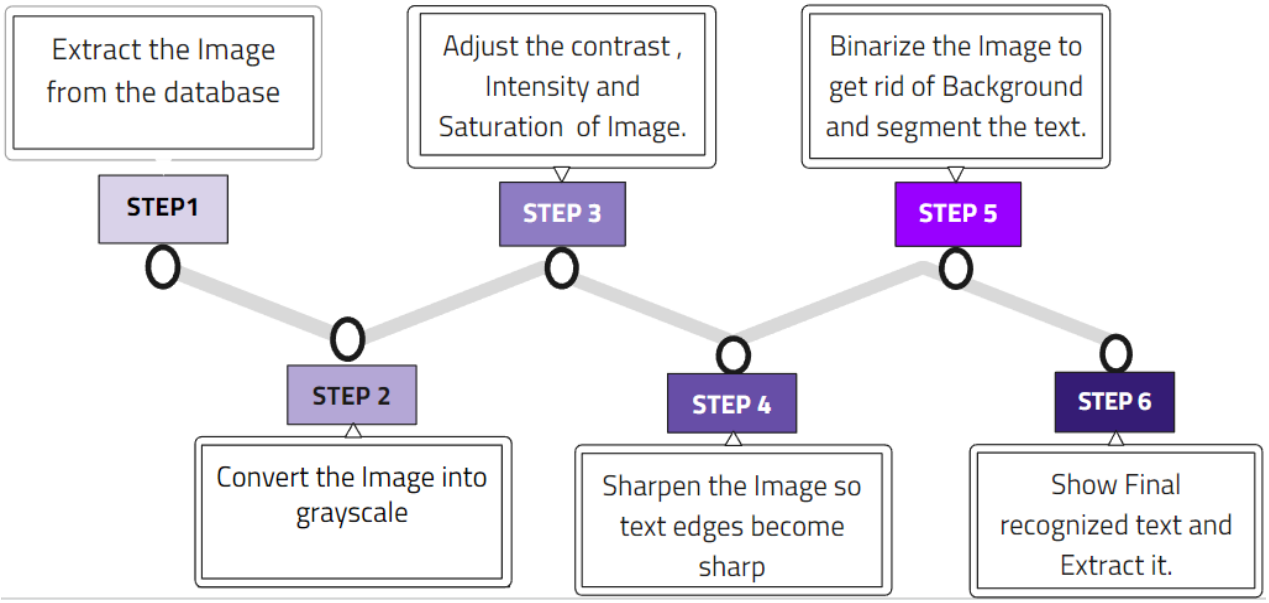


Fig 1. Block diagram for text recognition and extraction

About our Approach:
In recent years the research in OCR has advanced to such a high level of understanding no matter machines and computers. This part of the research paper will explain what we've done to simplify the method and make OCR more generic for all types of text images. The input of OCR is often of two types either it may be handwritten or machine printed recognition system. During this research paper we've solely focused on the printed text because within the majority of sectors like finance, Medicine, Education etc. documents are printed and not handwritten and nowadays availability of Things in digital format has become a necessity.

1)Image Extraction:
Through the scanning or image capturing an image of the document is captured in JPEG, PNG, BMP or JPG formats. Then we extract the Image and everyone its information is stored within the style of pixels. We process the Image so we are able to extract further information. In this step Quality of the captured image matters significantly because clearer the image lesser the possibilities of errors in output. So, it's better if image is captured through an HD Camera or on high focus while using smartphone cameras. we've tried to enhance the standard of Image by some preprocessing methods but every image is exclusive so requirements for preprocessing and enhancement are different for each image but we've tried to make it work for many of the unscrewed and unblurred text images. For the Skewed and blurred images, the preprocessing technique will include more steps for deblurring and rotating the image so text vertically fits the screen.

2) Image Preprocessing:

a) Gray scaling and Binarizing:

Majority of input images are colorful in nature and as a result of that, we tend to convert a multicolor image into grayscale to avoid wastage of storage space and complex computation. The thresholding process is vital as it makes various factors easier for computation. The color of text and color of background is always distinct in nature for text to be visible. Even if only edges of text are of different color thresholding image can create features which can be distinguished and separated as character easily. When we grayscale the image we set a certain value for every pixel which determines its tone or how much black or white it is. The shade or closeness to a black (1) or white (0) determines value of pixel between 0 and 1. When set a certain value as threshold all the pixels below that value become 0 or white and all the pixels above that value become black. This process is called as thresholding or binarization.

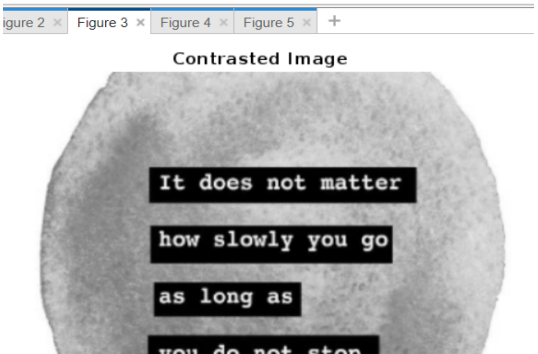
As we are using the computer vision tool kit here the threshold value is calculated automatically with threshold function. However, it can also be given manually as a lot of documents we come across may have a rather large or little contrast or lot of noise.

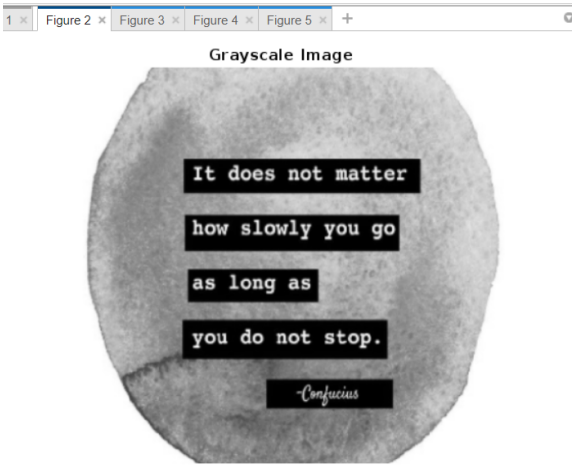
b) Contrasting and Sharpening:

The image generated from the scanning or capturing process may or may not contain a specific amount of noise. When we sharpen a picture, we are improving the definition between tones. In other words, you're adding contrast between edges, which makes those edges appear sharper. Which helps to separate text from background easily. Changing the contrast means you're adjusting the range of tones—when you give a picture more contrast, you're giving it a broad range of tones between the blacks and also the whites. This helps in removing the fuzziness of image and adjust the intensity in order that there's no chance of missing a letter while text extraction.

3) Text Recognition and Extraction:

The Function boxes = locate Text (ocrText, pattern) returns the placement and size of bounding boxes stored within the object. The locate Text function returns only the locations of bounding boxes which correspond to text within a picture that exactly match the input pattern. We can also use pretrained CRAFT deep learning framework from computer vision toolbox in MATLAB. This tool is very useful for locating text in image. The pixels have certain values as we saw earlier, they can also be viewed in form of matrix which will contain their spatial coordinates with respect to entire image. These coordinates give locations of individual pixels and when we apply detectTextCRAFT function the the coordinates of these pixels having text are computed. As we have done all the preprocessing earlier this model can localize the characters in sentences or words and then, we can extract the text from detected text locations OCR function. OCR function has inbuilt and pretrained character matching algorithm which returns the text output matched with the recognized character. Similarly with numbers and special characters it tries to return the best fit match.





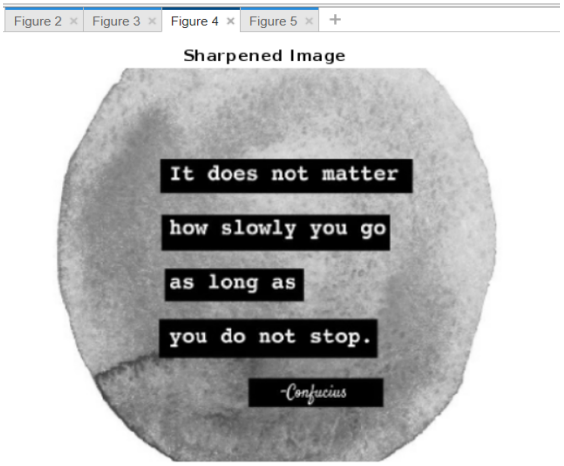


Fig 2. Output images after each step

IV. Results:

Fig 3. Final output image of sample Image 1

Fig 4. Output of Sample Image 2

To test the accuracy of our method we fed various sample images containing English text as input, and measured the accuracy based on the ratio of total letters and words given as input with number of letters and spaces or special characters it successfully recognized. Most of the text recognized was correct. In few of the images some recognized part contained mistakes, but they have been unreadable or damaged during the capturing process. Majority of times algorithm was able to successfully recognize and extract even special symbols and numbers. In few cases it got confused between some symbols (like? / and ! or, and.; and:) and some characters (0 and O, B and 8) which is just minor error and can be curable manually. Handwritten characters cannot be recognized using this algorithm due to uneven lines and spaces between characters. It can generally predict any quite normal font used in documents. When an image contained text with two completely different fonts (Calibri and Italic) it recognizes the matter written in one of them (Calibri) clearly, while other one contained error. In majority of the recognition where error occurred was due to over preprocessing during which the pixels from the text which define

the character and make it unique were lost. We used MATLAB (R2022.a/64-bit) for the implementation of OCR algorithm. The recognition accuracy was 80% to 90% counting on the quality of the image captured.

V. Conclusion and future scope:

In this paper, we've gone through different methods of feature extraction and classification techniques for OCR. We have also seen its various applications in different sectors. Finally, we've presented our own method for extracting the text from image using MATLAB. Our approach is simple to implement thanks to its algorithmic simplicity and ease of implementation. Recognition is fastest and strongest in uniform single column script and doesn't require training of any models or any huge datasets. Today OCR accuracy is almost perfect and every sector is flourishing thanks to deep research and innovation in this field.

REFERENCES:

[1] Matthias Engelbach, Dennis Klau, Jens Drawehn, Maximilien Kintz1, " Combining Deep Learning and Reasoning for Address Detection in Unstructured Text Documents" , arXiv:2202.03103v1 [cs.AI] , [Online accessed 1 September 2022]

[2] Hrishikesh Terdalkar Arnab Bhattacharya,"Chandojnanam: A Sanskrit Meter IdentificationandUtilizationSystem", https://sanskrit.iitk.ac.in/jnanasangraha/chanda/,https://github.com/hrishikeshrt/chanda/, [Accessed 5 September 2022]

[3] Proddutur Shruthi and Dr. Devaraj Verma,"A Detailed study and recent research on OCR", International Journal of Computer Science and Information Security (IJCSIS),Vol. 19, No. 2, February 2021.

[4] Jay Dilipbhai Thanki, Priyank Dineshbhai Davda, Dr. Priya Swaminarayan,"A Review on OCR Technology", International Journal of Emerging Technologies and Innovative Research, ISSN:2349-5162, Vol.8, Issue 4, page no.716-720, April-2021

[5] A. T. Sahlol, M. Abd Elaziz, M. A. A.Al-Qaness and S. Kim, "Handwritten Arabic Optical Character Recognition Approach Based on Hybrid Whale Optimization Algorithm With Neighborhood Rough Set ," in IEEE Access, vol. 8, pp. 23011-23021,2020, - 30 January 2020.

[6] B. Dessai and A. Patil, "A Deep Learning Approach for Optical Character Recognition of Handwritten Devangiri Script," 2019 2nd International Conference on Intelligent Computing, Instrumentation and Control Technologies (ICICICT), Kannur, Kerala, India, 2019, pp. 1160-1165,doi: 10.1109/ICICICT46008.2019.8993342.-2019.

[7] T. T. Zin, S. Z. Maw and P. Tin, "OCRPerspectives in Mobile Teaching andLearning for Early School Years in BasicEducation," IEEE 1st GlobalConference on Life Sciences andTechnologies (LifeTech), Osaka, Japan,2019, pp. 173-174,doi:10.1109/LifeTech.2019.8883978.-2019.

[8] Alghazal, Mohammed. "Talent Acquisition Process Optimization Using Machine Learning in Resumes' Ranking and Matching to Job Descriptions." Paper presented at the SPE Middle East Oil & Gas Show and Conference, event canceled, November 2021. doi: https://doi.org/10.2118/204534-MS, December 15 2021

[9] S. Anbarasi, S. Krishnaveni,R. Aruna K. Karpagasaravanakumar , Smart Reader Glass for Blind and Visually Impaired People, Recent Trends in Intensive Computing M. Rajesh et al. (Eds.) © 2021 The authors and IOS Press, doi:10.3233/APC210303.

[10] Prof. Bamb Kalpesh K, Prof. Tated K.S, Prof. Mutha H.H., Prof. Chopda P.P.,"A Literature Survey on Character Recognition of Indian Scripts for New Researchers" 6th International Conference on Recent Trends in Engineering & Technology (ICRTET - 2018), International Journal for Research in Engineering Application & Management (IJREAM) Special Issue – ICRTET-2018 ISSN : 2454-9150.

[11] Rajat Subhra Bhowmick, Isha Ganguli, Jayanta Paul, Jaya Sil, A Multimodal Deep Framework for Derogatory Social Media Post Identification of a Recognized Person, ACM Transactions on Asian and Low-Resource Language Information Processing Volume 21 Issue 1 January 2022 Article No.: 2pp 1–19, https://doi.org/10.1145/3447651, 02 November 2021

[12] Luna, A.C., Trajano, C., So, J.P., Pascua, N.J., Magpantay, A., Ambat, S. (2022). License Plate Recognition for Stolen Vehicles Using Optical Character Recognition. In: Fong, S., Dey, N., Joshi, A. (eds) ICT Analysis and Applications. Lecture Notes in Networks and Systems, vol 314. Springer, Singapore. https://doi.org/10.1007/978-981-16-5655-2_56

[13] Arafat A. Muharram, Khaled M. G. Noaman, K.S.A Ibrahim Abdulrab Ahmed, K.S.A Jamil A. M. Saif, Optical Character Recognition based on Genetic Algorithms and Machine Learning, International Journal of Computer Applications (0975 – 8887) Volume 172 – No.2, August 2017.

[14] P. Bhavani, S. Vaishnavi, P. Vennila, V. Vijayalakshmi, Bus Attendance System using Optical Character Recognition, Retrieval Number: D7732049420/2020©BEIESP DOI: 10.35940/ijeat.D7732.049420 Journal Website: www.ijeat.org, International Journal of Engineering and Advanced Technology (IJEAT) ISSN: 2249 – 8958 (Online), Volume-9 Issue-4, April, 2020.

[15] W. A. J. R. Silva, H. M. K. Shirantha, L. J. M. V. N. Balalla, R. A. D. V. K. Ranasinghe, N. Kuruwitaarachchi and D. Kasthurirathna, "Predicting Diabetes Mellitus Using Machine Learning and Optical Character Recognition," 2021 6th International Conference for Convergence in Technology (I2CT), 2021, pp. 1-6, doi: 10.1109/I2CT51068.2021.9417941.

About Authors

She is a Second-year student, pursuing Electronics Engineering from Shri Ramdeobaba College of Engineering and Management, Nagpur. She is also studying Artificial intelligence, Machine learning, Deep learning for Visual recognition as Honors.