International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 07 Issue: 04 | April - 2023 Impact Factor: 8.176 ISSN: 2582-3930

Regional Language Optical Character Recognition Using Machine Learning

Devyani Parmar[1], Prof. Ankur J Goswami[2]
[1]Department of computer science, sankalchand Patel College of Engineering, visnagar
[2]Faculty of Sankalchand Patel College of Engineering, Visnagar,India

---

Abstract - The character recognition is fast-growing and composite area of machine learning and pattern recognition. The world is moving away from traditional pen and paper kind of communication tool and moving towards a digital world. As the world is becoming digital, more and more people are adopting digital technologies which result in increasing use of handheld devices. Digital world opens many opportunity and areas for researchers. There is a need of online handwritten character recognition system to provide an easy and efficient tool to communicate with digital gadgets in a traditional language and traditional way.

Optical Character Recognition (OCR) is a process or technology in which text within a digital image is recognized. It is mainly used for converting the transcribed, handwritten or any printed text to the text data that can be edited and reused.

Keywords: Optical character recognition, feature extraction, image processing, handwritten documents , character extraction

Introduction:

Today is the era of paperless office and governance. It comes with numerous advantages like increased productivity and efficiency, storage optimization, robustness and eco-friendliness. Hence there is a need of converting paper documents into machine editable form. This leads to development of OCR (Optical Character Recognition). OCR is a technique to convert, mechanically or electronically an image, photo or scanned document of a handwritten text or printed text into digital text.

Machine Learning-based Optical character recognition (OCR) Scanner will convert images of a typed, handwritten or printed text into machine-encoded text. It has been man's ancient dream to develop machines which replicate human functions. One such replication of human functions is reading of documents encompassing different forms of text. Over the last few decades machine reading has grown from dream to reality through the development of sophisticated and robust Optical character recognition (OCR) systems.

The focus of this application is to help various educators, lecturers, and students to make a text document of their
handwritten notes. The process of character recognition can be divided into two parts, namely, printed and handwritten character recognition.

The basic process of OCR involves capturing an image of text, such as a scanned document or a photograph, and using software to identify the individual characters within the image. The software then maps each character to a corresponding letter or number in a computer-readable format.

OCR technology has many practical applications, including automated data entry, document management, and digital archiving. For example, OCR can be used to convert paper-based documents into electronic format, making it easier to search, store, and retrieve information.

While OCR technology has come a long way, it is still not perfect, and there are many factors that can affect its accuracy, including image quality, font type, and language complexity. However, recent advancements in machine learning algorithms have enabled OCR systems to improve their accuracy by learning from large datasets of images and text.

Overall, OCR technology has revolutionized the way we handle and process information, making it faster, more efficient, and more accessible to a wider audience.

English language

The English Language is the most widely used language in the world. It is the official language of 53 countries and articulated as a first language by around 400 million people. Bilinguals use English as an international language character recognition for the English language has been extensively studied throughout many years. the English language has the highest number of publications. The OCR systems for the English language occupy a significant place as a large number of studies have been done in the era of 2000-2018 on the English language. The English language OCR systems have been used successfully in a wide array of commercial applications.[2]

OCR technology can recognize and convert English text from images into machine-readable format. The accuracy of OCR systems for English language text recognition can be quite high, especially when using advanced techniques such as deep learning algorithms. However, there are some challenges that can affect the accuracy of OCR systems for English text recognition, such as:

Image quality: OCR systems may struggle to recognize English text from images that are blurry, poorly lit, or have low resolution.

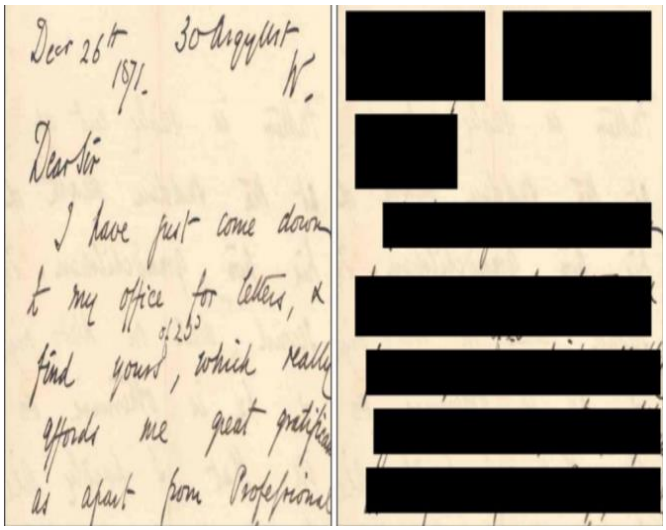
Font type: OCR systems may struggle to recognize certain fonts or handwriting styles, especially if they are unusual or difficult to read.

Language complexity: OCR systems may struggle to recognize English text that contains complex phrases or sentences, such as legal or technical documents.

Despite these challenges, OCR technology is widely used for English language text recognition in various industries, including banking, healthcare, and legal services. The accuracy of OCR systems for English language text recognition is constantly improving due to advancements in machine learning algorithms and the availability of larger datasets for training and testing OCR models.

Segmentation Phase

The critical and major component of an Optical Character Recognition (OCR) system is the segmentation of text line from images. In general, Text segmentation from a document image merges line segmentation, word segmentation and then character segmentation. Segmentation is the process of isolating text component within an image from the image's background. For appropriate reorganization of the editable text lines from the recognized characters, firstly, segmenting the line of text, then the words are segmented from the segmented line and then from that the characters are segmented. Document segmentation is a major pre-processing phase in implementing an OCR system. It is the process of classifying a document image into homogeneous zones, i.e., that each zone contains only one kind of information, such as text, a figure, a table, or a halftone image. In many cases, the accuracy rate of systems related to the OCR heavily depends on the accuracy of the page segmentation algorithm used.[3]



[a] **[b]**

Semantic segmentation also requires the existence of a parallel mask for every image in the dataset. A mask is a binary image, not unlike a 'label' in typical machine learning datasets, wherein the pixels that are of the class to be segmented are highlighted in black and the rest of the image is kept white. Generating these masks for every image required writing an algorithm that would append a box of black pixels precisely the same dimensions of a sentence to another blank image at the exact location that the sentence is being appended to in its image.

Blurring and Degradation :-

For the best accuracy of character recognition and character segmentation, character sharpness is required. At large apertures and short distances, uneven focus can be observed when a small point of view changes. For the most part connected with photography, there are two kinds of obscure which is: out of focus obscure and movement obscure [4]. At the point for catching a moving item, when the shade rate of the camera is not sufficiently high, the sensor gets presented to a continually changing scene. Accordingly, blurring will observed in parts in motion.

Pre-processing

As we seen above, some noise may occurred during scanning process. This results in poor recognition of characters. This usually occurred problem is overcome by preprocessing. It consists of smoothing and normalization. In

smoothing, certain rules are applied to the contents of image with the help of filling and thinning techniques. Normalization is responsible to handle uniform size, slant and rotation of characters. [9]

Deals with Improving quality of the image for better recognition by the system. Ocr software often "pre-processes " image to improve the chances of successful recognition.

Techniques Include :

- De-skew
- Despeckle
- Binarization
- Line removal
- Zoning
- Script recognition
- Segmentation

Improve OCR Accuracy with Advanced Image Preprocessing

Optical Character Recognition (OCR) technology got better and better over the past decades thanks to more elaborated algorithms, more CPU power and advanced machine learning methods. Getting to OCR accuracy levels of 99% or higher is however still rather the exception and definitely not trivial to achieve. [13]

First, Let's Define OCR Accuracy
When it comes to OCR accuracy, there are two ways of measuring how reliable OCR is:

- **Accuracy on a character level**
- **Accuracy on a word level**

In most cases, the accuracy in OCR technology is judged upon character level. How accurate an OCR software is on a character level depends on how often a character is recognized correctly versus how often a character is recognized in correctly. An accuracy of 99% means that 1 out of 100 characters is uncertain. While an accuracy of 99.9% means that 1 out of 1000 characters is uncertain.

Measuring OCR accuracy is done by taking the output of an OCR run for an image and comparing it to the original version of the same text. You can the neither count how many characters were detected correctly (character level accuracy), or count how many words were recognized correctly (word level accuracy).[13]

To improve word level accuracy, most OCR engines make use of additional knowledge regarding the language used in a text. If the language of the text is known (e.g. English), the recognized words can be compared to a dictionary of all existing words (e.g. all words of in the English language corpus). Words containing uncertain characters can then be "fixed" by finding the word inside the dictionary with the highest similarity.

When it comes to improving OCR accuracy, you basically have two moving parts in the equation.

The Quality of Your Source Image

If the quality of the original source image is good, i.e. if the human eyes can see the original source clearly, it will be possible to achieve good OCR results. But if the original source itself is not clear, then OCR results will most likely include errors. The better the quality of original source image, the easier it is to distinguish characters from the rest, the higher the accuracy of OCR will be.

The OCR Engine

An OCR engine is the software which actually tries to recognize text in whatever image is provided. There are various OCR engines available, ranging from free open source OCR engines to proprietary solutions with a hefty price tag.

While many OCR engines are using the same type of algorithms, each of them comes with its own strengths and weaknesses.OCR accuracy comparison is difficult and choosing the right OCR engine mostly depends on your specific use-case, the allocated budget and how it integrates into an existing system.

APPLICATIONS

Banking
Another imperative use of OCR is in banking [2], where it is utilized to process cheques without human intervention. A cheque can be embedded with a machine where the framework filters the sum to be issued and the right measure of cash is exchanged.

Healthcare
To process printed material, medicinal services [5] have likewise seen an expansion in the utilization of OCR innovation. Medicinal service experts continuously need to manage extensive volumes of documents for each patient, including protection frames and in addition general health forms. To stay aware of every one of this data,it is valuable to input relevant information Into an electronic database. With OCR processing tools, we can
extract data from structures and put it into databases, so that each patient's information is quickly recorded and retrieved when needed in future.

Legal Industry
Legal industry is likewise one of the recipients of the OCR innovation. OCR is utilized to digitize documents, and to specifically enter into PC database. Legitimate experts can further search documents required from tremendous databases by basically writing a few keywords.

Feature Extraction

It extracts the features of symbols. Features are the characteristics. In this, symbols are characterized and unimportant attributes are left out. The feature extraction technique does not match concrete character patterns, but rather makes note of abstract features present in a character such as inter sections, open spaces, lines, etc. [7] Tesseract algorithm is used to implement feature extraction. Feature extraction is concerned with the representation of the symbols. The character image is mapped to a higher level by extracting special characteristics of the image in the feature extraction phase.

There are methods for extracting features in OCR:

In these method , pattern recognition works by identifying the entire character.

FIG.1 : PATTERN RECOGNITION ON A SINGLE CHARACTER. [1]

ADVANTAGES

Following are some of the advantages of Optical Character Recognition to Increased Productivity: Optical Character Recognition helps the organizations to increase the productivity as it increases the efficiency by allowing the faster data fetching, whenever it is needed. Time taken by people for the data entry is now saved as OCR does the same work within no time. Staffs also do not need to go through the files of the company for accessing any data as he/she can access it digitally from his/her desk. [6].

High Accuracy: The OCR software uses large amount of data for training and testing, this solves the problem of low accuracy and provides consistency. It also avoids the mistakes by employees that may occur at the time of data entry.

In addition, OCR also helps to address the problems such as data loss. Problems such as accidentally entering incorrect information can be resolved.[6]

Gujarati Script

The Gujarati language has large and complex character set and many characters have similar strokes, which makes OCR more challenging.

This paper deals with an optical character recognition (OCR) system for handwritten Gujarati numbers. One may find so much of work for Indian languages like Hindi, Kannada, Tamil, Bangala, Malayalam, Gurumukhi etc, but Gujarati is a language for which hardly any work is traceable especially for handwritten characters. Here in this work a neural network is proposed for Gujarati handwritten digits identification. A multi layered feed forward neural network is suggested for classification of digits. The features of Gujarati digits are abstracted by four different profiles of digits. Thinning and skew-correction are also done for preprocessing of handwritten numerals before their classification. This work has achieved approximately 82% of success rate for Gujarati handwritten digit identification.[9]

Gujarati is the official language of Gujarat and is spoken by 60.3 million people. It is derived from Devanagari script. In Gujarati script there are 11 vowels and 36 consonants[3], few additional characters are also there. Consonant-vowel combination is denoted by attaching symbol of vowels to the consonants. Every vowel can be represented by a unique symbol, called vowel modifiers. Vowel modifier can appear to the right, left, top, bottom and middle of the character. A combination of two or more characters can make new characters (Conjuncts - Jodakshar). Unlike many North Indian languages, Gujarati has no shirorekha (Upper Horizontal line on the top of a word, as in Devanagari script) and hence all the characters in a word are isolated. The character set of Gujarati is almost double than that of English language.

These are the few characteristics of Gujarati script which can be considered as a reason for the slow progress in development of Gujarati character recognition. [10]

Problem statement :

The problem here is that when information is scanned through a paper documents the computer system for character recognition as we know that we have a number of newspapers and books which are in printed formats related to different subjects. Whenever we scan through the scanner , the documents are stored in the computer system as images like jpeg, gif, etc…

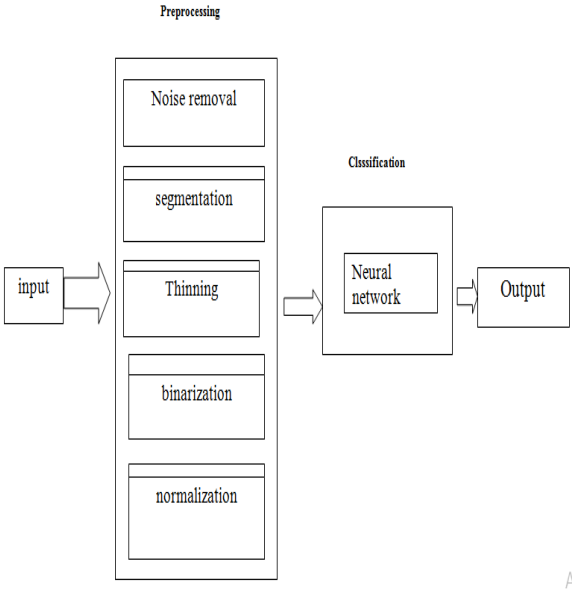
This image cannot be read or edited by the user. And it is very difficult to read individual materials and find line-by-line and word-by-word contents of these documents to reuse the information. This Days there is a huge demand in "storing the information in these paper documents into a computer storage disk and then later editing or reusing this information by searching process".

CNN :

Neural networks are designed by taking inspiration from brain. Convolution neural network is mainly used for image classification. CNN consist of many layers depending on requirements. The OCR can be implemented using convolution neural network (CNN). Which is a popular deep neural network architecture . The traditional CNN classifiers are capable of learning the important 2D features present in the image and classify them , the classification is performed by using soft-max layer.

CNNs are made of a large number of interconnected neurons that have learnable weights and biases.

CNNs can be trained to recognize individual characters within an image and classify them into the correct category, allowing for accurate OCR.



Artificial Neural Networks

Artificial neural networks (ANN) are used for determining what character is shown in an image. A simple definition of an ANN is '... a computing system made up of a number of simple, highly interconnected processing elements, which process information by their dynamic state response to external inputs.A neural net works by performing a computation

on a set of inputs, resulting in a set of outputs. The computation is specifically chosen such that the output says something meaningful about the input. An ANN consists of many nodes, called 'neurons'. Each neuron is a computational unit with one or more inputs and one output.[11]

Biological neuron inspired architecture , Artificial Neural Networks (ANN) consists of numerous processing units called neurons. These processing elements (neurons) work together to model given input data and map it to predefined class or label.

The main unit in neural networks is nodes (neuron). Weights associated with each node are adjusted to reduce the squared error on training samples in a supervised learning environment (training on labeled samples / data). Figure 8 presents pictorial representation of Multi Layer Perceptron (MLP) that consists of three layers i.e. (input, hidden and output).Feed forward networks / Multi Layer Perceptron (MLP) achieved renewed interest of research community in mid 1980s as by that time "Hopfield network" provided the way to understand human memory and calculate state of a neuron .Initially, computational complexity of finding weights associated with neurons hindered application of neural networks. With the advent of deep (many layers) neural architectures i.e. Recurrent Neural Network (RNN) and Convolutional Neural Networks (CNN), neural networks has established itself as one of the best classification technique for recognition tasks including OCR..[12]

Convolutional Neural Networks (CNNs): This is a deep learning approach that has become popular in recent years for OCR. CNNs can automatically learn the features of the input image and use them for classification.

Recurrent Neural Networks (RNNs): RNNs are neural networks that are designed to handle sequential data. They can be used for OCR by treating each character in the input image as a sequence of pixels.

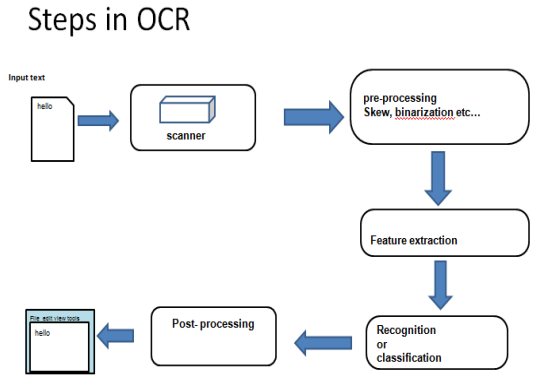
Attention Mechanisms: Attention mechanisms are a recent addition to deep learning models for OCR. They allow the network to focus on specific parts of the image that are most relevant for recognition.

Transfer Learning: This is a technique where a pre-trained model is fine-tuned for OCR. This can significantly reduce the amount of training data required and improve the accuracy of the OCR system.

Overall, deep learning approaches such as CNNs and RNNs have shown to be very effective for OCR, outperforming traditional computer vision methods in many cases. However, the choice of solution depends on the specific requirements of the application and the available resources for training and deployment.

Methodology

There are few steps through classification and recognition of character could be done.



Preprocess the images to improve the accuracy of the OCR model. This may involve converting the images to grayscale , applying noise reduction techniques , and performing image enhancement.

Segment the characters in the images using techniques such as morphological operations, edge detection, and thresholding. This step is important to isolate each character and improve the accuracy of recognition.

Extract features from each segmented character, such as the shape, size, and orientation. These features are used to train the OCR model.

Train the OCR model using a machine learning algorithm, such as a convolutional neural network (CNN) or a support vector machine (SVM). The model should be trained on a subset of the dataset, with the remaining data reserved for testing.

Proposed solutions :

Using OCR technology for your work may be quite tricky and needs time to learn from mistakes. Here are ways to better perform your OCR accuracy:

1. Good Quality of input text

Before using OCR, make sure you can read the images with your own eyes. If you, with your own eyes, can't see the image clearly, make sure the original source images are not damaged AND wrinkle-free. So, use the cleanest and most original files for better results.

2. Right Size of Images

OCR engine needs to read source images not only the ones with the best quality but also the right resolution. Make sure the image or PDF file is resized to the correct size, which is usually about 1 / 10 of the original size (1.5 mm x 1 mm) or less. This way, the result will be more accurate.

3. Remove Noise / Denoise

Human eyes can't even read documents that have many noises, so does the OCR engine. Noises make the engine difficult to read original sources and it can decrease the OCR accuracy. If the image has background or foreground noise, remove it to get a higher quality data extraction.

4. Increase Image Contrast

How do you see white papers with light grey ink? You -and the OCR engine must be uncomfortable reading such papers. Thus, try to increase the contrast between text and background brings more clarity to the output. The best contrast will help the OCR engine to read images accurately.

5. De-skew Original Source

No one wants to read papers upside down. Thus, make sure you get the image in the right format and shape (text should appear horizontal and not inclined). The image can be rotated by tilting it to one side, turning it clockwise or counter-clockwise, and turning it back to the other side.

Human and OCR accuracy is actually the same since both of them are working in the same ways. The only difference is that OCR uses engines to get the jobs done.

Result and conclusion :

The study shows that the system needs more training for hand-written characters than computerized fonts.

In most cases, 98-99% accuracy is the acceptable accuracy rate, measured at the page level (not algorithm level). This means that in a page of around 1,000 characters, 980-990 characters should be accurately identified by the OCR software

Future work :

Our next works with OCR Mobile Application will include the improvement of the results. OCR application will also display the signatures and the other symbols as it is in the document. It will also update its features including the translation of one language to another. So that it will helpful for people from other countries who can't understand the local language.

REFERENCES :

[1]HTTPS://MOOV.AI/EN/BLOG/OPTICAL-CHARACTER-RECOGNITION-OCR/

[2] Ganis MD , Wilson CL, Blue JL." Neural network-based systems for handprint OCR applications ." IEEE Transactions on Image Processing. 1998 Aug ; 7(8):1097-112.

[3] Patel C, Patel A, Patel D. "Optical character recognition by open source OCR tool tesseract: A case study. " International Journal of Computer Applications. 2012 Jan 1;55 (10).

[4] Jain A, Dubey A, Gupta R, Jain N, Tripathi P. "Fundamental challenges to mobile based ocr." vol. 2013 May;2:86-101.

5] Ganis MD, Wilson CL, Blue JL. "Neural network- based systems for handprint OCR applications. " IEEE Transactions on Image Processing. 1998 Aug;7(8):1097-112.

[6] Abin M Sabu and Anto Sahaya Das, "A Survey on various Optical Character Recognition Techniques," 2018 Conference on Emerging Devices and Smart Systems (ICEDSS)

[7] C. C. Tappert, C. Y. Suen, and T. Wakahara, ''The state of the art in online handwriting recognition,'' IEEE Trans. Pattern Anal. Mach. Intell., vol. 12, no. 8, pp. 787–808, Aug. 1990, doi: 10.1109/34.57669

[8] Honggang Wang, Ming C. Leu and Cemil Oz, "American Sign Language Recognition Using Multidimensional Hidden Markov Models", Journal of Information Science and Engineering, January 2006.

[9] Ravina Mithe, Supriya Indalkar, Nilam Divekar " optical character recognition " International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-2, Issue-1, March 2013

[10] Jyoti pareek, dimple singhaniya, rashmi rekha kumari "Gujarati Handwritten Character Recognition From Text Images " Third International Conference on Computing and Network Communications (CoCoNet'19)

[11] T. (Tijmen) van Dien "Information retrieval through optical character Recognition" Department of Mathematics and Computer ScienceSystem Architecture and Networking (SAN),Eindhoven, April 2018

[12] Jamshed Memon , Maira Sami, and Rizwan Ahmed Khan " Handwritten Optical Character Recognition (OCR): A Comprehensive Systematic Literature Review (SLR)" IEEE 2020

[13] https://docparser.com/blog/improve-ocr accuracy/

[14] https://www.codersarts.com/post/optical-character-recognition-using-convolutional-neural-network

[15]https://gleematic.com/what-is-ocr-accuracy-how-to-improve-it/

[16] Ponvizhi. U, Ramya. P, Ramya. R "Optical Character Recognition Using Python" International Journal of Trend in Scientific Research and Development (IJTSRD) Volume 5 Issue 3, March-April 2021 Available Online: www.ijtsrd.com e-ISSN: 2456 – 6470