OCR APPLICATION FOR MULTIPLE LANGUAGE DETECTION

[1]Tanvi Modak, [2]Shweta Bhoje, [3]Gargee Tikekar, [4]Prof.Mangesh Balpande

[1]Student, [2]Student, [3]Student, [4]Assistant Professor

[1]Information Technology,
[1]Finolex Academy of Management & Technology, Ratnagiri, India.

Abstract: **In today's modern era of digitization there is a huge demand for recognizing the characters from printed documents. Newspaper articles and books are available on paper. So there is a need to store them on computer storage so that they can be searched and made available whenever necessary. For this, a software system is required to scan the documents and store them in digital format. Optical character recognition is such a tool which will convert the scanned images into a machine-encoded text. OCR helps in editing and searching such printed documents or handwritten text. The proposed system is reliable and accurate and can recognize multiple languages.**

Keywords: Optical character recognition, artificial neural network, back propagation

I. INTRODUCTION

Optical character recognition abbreviated as OCR is a part of document image analysis which is used to convert the scanned printed documents in the electronic form readable by a machine. It is a widespread technology to recognize text inside images which can be printed or handwritten. OCR technology came into existence in the 1990s in attempting to digitize historical documents. Since then, it has undergone many changes and new research is made to improve accuracy. OCR can be useful in converting the scanned documents so that it can be edited with word processors like Microsoft Word or Google Docs.

OCR falls into two categories printed document recognition and handwriting recognition. Printed document recognition is simple because characters are in uniform dimension. Handwriting recognition is difficult because each users writing style is different. Handwriting recognition is further divided into two subparts online and offline. Online OCR is performed on real-time data whereas Offline OCR operates on static data which is scanned in the form of an image.

OCR can be used for a variety of applications. Banking industry uses this technique to handle handwritten checks for verification of the signature. The legal industry makes use of OCR for storing legal documents like affidavits, judgments, statements, wills and other. OCR can be used in the healthcare industry for searching medical history of patients and hospital records. There are various benefits of OCR like search-ability, edit-ability, accessibility, storability, backups, and translatability.

OCR is helpful to reduce the storage space. Images require very large space as compared to text documents. For example, if a document in text format requires 2KB of space whereas image requires 5KB. Also, it is difficult to read the images because they are a blur and searching for a word becomes impossible.

There are six traditional phases of implementing OCR which is as follows:-

1. Pre-processing

Pre-processing is required to remove the noise and skew from the image to improve the quality of an image. Binarization is the method of converting a gray-scale image to a binary image. Thresholding is required to separate information from the background. Binarization of image is done by thresholding which includes approaches like global thresholding and local thresholding. Global thresholding is used to find a single threshold value and each pixel is either assigned to foreground or background based on this value. Local thresholding can be done using clustering approaches, entropy, and neighbourhood information.

Image quality is improved by noise reduction. The skewness of the document is corrected by adjusting the rotated image. Various morphological operations like adding or removing pixels of the characters that have excess pixels are done.

2. Segmentation

Segmentation is the process of separating the text from the image. It includes approaches like the top-down approach, bottom-up approach, and hybrid approach. In the top-down approach segments, large regions into sub regions and the process will stop when a criterion is met. In a bottom-up approach, first interest pixels are searched and then combined into words and lines or text blocks. The combination of both top-down and bottom-up approach is called a hybrid approach.

Line segmentation can be done by horizontal projection profile (HPP) which is a Histogram of ON pixels on every row of the image. Word segmentation is done to find the spacing between words. It is done by vertical projection profile (VPP) which is Histogram of ON pixels on every column of the image.

3. Normalization

This system is used to reduce the segmented text in a particular size for the next phase. Normalization will eliminate unnecessary information from the image without losing any influential information. In normalization, the intensity value of the pixel is changed to the range of [0, 1]. The various dimension images are converted into fixed dimensions in normalization. Normalization can reduce the complexity of the image so that feature extraction can be done effectively.

4. Feature extraction

Various features of the text are extracted in this phase. They can be geometric features like loops or strokes and statistical feature like moments. There are two types of features in OCR. The first type is based on zones. The image is divided into horizontal zones and vertical zones. For each zone, we calculate the density of the characters. In the second type, upper and lower as well as left or right area of character is calculated.

5. Classification

It is used to classify the characters inappropriate category. Classification is the procedure of assigning classes in which the pattern falls. Classification can be done by K-nearest neighbour and support vector machines (SVM).

K-nearest neighbour is one of the simplest classification algorithms. An object is classified by a majority of vote of its neighbour and object is assigned to the class which is common among its nearest neighbours. The k value is always positive. To find k nearest objects Euclidean distance is used.

Support vector machines (SVM) is a supervised learning method. In SVM the data is divided into training and testing set. The SVM takes the input data and classifies them into one of two distinct classes. A model is created by training the data and this model is used to classify test data.

6. Post-processing

This phase includes improving the efficiency of OCR. Post-processing is used to reduce the number of errors. The post-processing is done by creating more than one classifier and choosing the classifier which has the highest accuracy.

II. LITERATURE REVIEW

There is various research work done on OCR. The first research introduces the basic information about OCR and various techniques used in OCR. There are various stages in implementing OCR [1]. Pre-processing is the first stage to remove noise from the image. Then comes character segmentation which will separate characters from the image. Feature extraction step will extract features of the text. And then the characters are classified into some category. The last phase post-processing includes improving the efficiency of OCR. Also, various applications of OCR are mentioned like number-plate recognition, smart libraries, and other real-time application.

Grid infrastructure is a technique which supports character recognition of multiple languages [2]. It supports editing and searching for documents. Accuracy is improved. It speeds up the process of character recognition. There are different modules in the proposed system. Document processing module for storing images and creating a grid infrastructure data structure. System training module for training fonts. Document recognition module for recognizing characters from an image. Document editing module for editing the documents and Document searching module for searching the documents.

There are numerous errors that emerge during the process because of environmental or camera related factors [3]. They include Skewness, Blurring and degradation of images, Conditions of Uneven lighting, Tilting, fonts italic, Multilingual environments, etc. Various phases of OCR are also mentioned like Pre-processing, Segmentation, Normalization, feature extraction, Classification, and post-processing. For languages like Hindi or Marathi, there are various fonts and features of these fonts, recognizing of these fonts is complex and requires more processing [6]. The first step is to take the printed binary image as input. The second step is to extract pixel information and store it into memory. After successful completion of the second step find the skeleton of character and check if it matches with pixel information. After the skeleton is available to find out various features like Horizontal lines, Vertical lines, Crosslines, loops, and curves or geometrical shapes in that skeleton. All the features are stored in the database. Then the features in the image are compared with the database and the character is printed in editable format.

Character recognition is not a new problem but its roots can be traced back to systems before the inventions of computers. The earliest OCR systems were not computers but mechanical devices that were able to recognize characters, but very slow speed and low accuracy. In 1951, M. Sheppard invented reading and robot GISMO that can be considered as the earliest work on modern OCR [6]. GISMO can read musical notations as well as words on a printed page one by one. However, it can only recognize 23 characters. The machine also has the capability to could copy a typewritten page. J. Rainbow, in 1954, devised a machine that can read uppercase typewritten English characters, one per minute. The early OCR systems were criticized due to errors and slow recognition speed. Hence, not much research efforts were put on the topic during the ’60s and ’70s. The only developments were done on government agencies and large corporations like banks, newspapers, and airlines, etc. Because of the complexities associated with the recognition, it was felt that three should be standardized OCR fonts for easing the task of recognition for OCR. Hence, OCRA and OCRB were developed by ANSI and EMCA in 1970 that provided comparatively acceptable recognition rates [7]. During the past thirty years, substantial research has been done on OCR. This has led to the emergence of document image analysis (DIA), multi-lingual, handwritten and Omni-font OCRs [7]. Despite these extensive research efforts, the machine’s ability to reliably read text is still far below the human. Hence, current OCR research is being done on improving the accuracy and speed of OCR for diverse style documents printed/ written in unconstrained environments. There has not been the availability of any open source or commercial software available for complex languages like Urdu or Sindhi etc.

III. PROPOSED SYSTEM

The proposed system uses Artificial Neural Network (ANN) for training and recognition of data. ANN is a wonderful tool that will help to resolve the problems in OCR. The advantage of using ANN in OCR is to simplify the development and achieves the highest quality of recognition and good performance.

3.1 Artificial Neural Network (ANN)

An artificial neural network is an engineering approach which is vaguely inspired by the structure and function of neurons in the human brain. ANN is composed of multiple nodes which are connected by links and they interact with each other. The basic neural network consists of an input layer, a hidden layer, and output layer and each layer is comprised of multiple neurons. The number of input neurons is equal to the input data values, the number of output neurons is equal to the output data values and the number of hidden layers is an arbitrary value.

The input layer receives information from the outside world that the network will attempt to learn and recognize. The layer which response to the information learned is called an output layer. Between the input layer and output layer are hidden layers which together form the majority of the network. The connections between one layer and the other are assigned a number called weight, which can be either positive or negative.

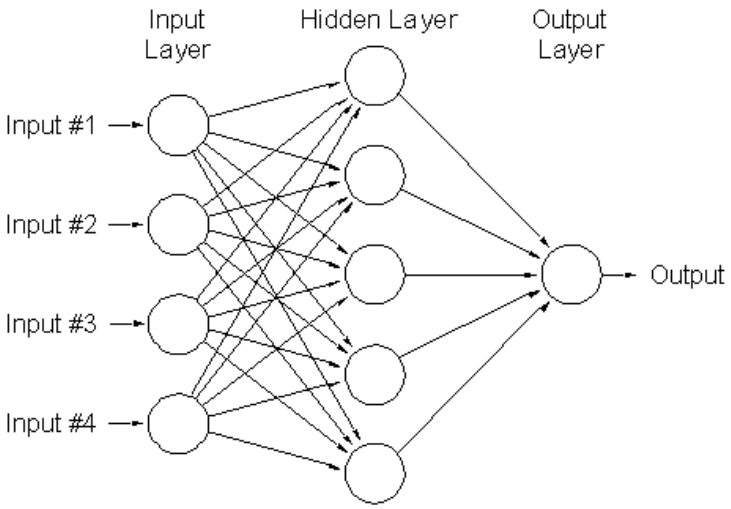

Fig1. Structure of Artificial Neural Network

3.2 Backpropagation algorithm

The backpropagation algorithm is the multilayer feedforward algorithm which .is the oldest and most powerful supervised learning method proposed by Rumelhart, Hinton and Williams. While designing a Neural Network we initialize weights with some random values. But if the values we have selected are not correct then the output changes that is the error value is huge. Backpropagation algorithm is used to train the neural network and reduce the errors.

There are four steps of implementing back propagation:

1. Compute how fast error changes when an output unit is changed. This error derivative (EA) is the difference between the actual and desired activity.

$$EA_j = \frac{\partial E}{\partial y_j} = y_j - d_j \qquad (3.2.1)$$

2. Compute how fast error changes as total input received by an output unit is changed. This quantity (EI) is the multiplication of answer from step 1 and the rate at which the output of unit changes as total input is changed.

$$EI_j = \frac{\partial E}{\partial x_j} = \frac{\partial E}{\partial y_j} X \frac{\partial y_j}{\partial x_j} = EA_j y_j - \left(1 - y_j\right) \qquad (3.2.2)$$

3. Compute how fast error changes as total input received by an output unit is changed. This quantity (EW) is the answer from step 2 multiplied by activity level of unit from which the connection originates.

$$EW_{ij} = \frac{\partial E}{\partial W_{ij}} = \frac{\partial E}{\partial k_j} X \frac{\partial k_j}{\partial W_{ij}} = EI_j y_j \qquad (3.3.3)$$

4. Compute how fast the error changes as activity of a unit in previous layer is changed. To calculate overall effect on error we add all separate effects on output unit. Each effect is calculated by multiplying answer in step 2 and weight on the connection of that output unit.

$$EA_i \frac{\partial E}{\partial y_i} = \sum_j \frac{\partial E}{\partial k_j} X \frac{\partial k_j}{\partial y_i} = \sum_j E_{ij} W_{ij} \qquad (3.3.4)$$

By using steps 2 and 4, we can convert the EAs of one layer into EAs for previous layer. This procedure can be repeated to get the EAs for as many as previous layers as desired. Once we know the EAs of the unit we can use step 3 and 3 to calculate EWs.

IV. CONCLUSION

Optical character recognition is a useful technique which recognizes the characters from an image. It reduces the manual work of retyping the printed or handwritten text to store it in electronic format. It is used in many sectors and helps in storing, editing and searching the text from images.

V. REFERENCES

[1] Noman Islam, Zeeshan Islam,Nazia Noor, "A Survey on Optical Character Recognition System", Journal of Information & Communication Technology-JICT Vol. 10 Issue. 2, December 2016

[2] Najib Ali Mohamed Isheawy, Habibul Hasan, "Optical Character Recognition (OCR) System", IOSR Journal of Computer Engineering (IOSR-JCE) e-ISSN: 2278-0661, p-ISSN: 2278-8727, Volume 17, Issue 2, Ver. II, Apr. 2015

[3] Karez Abdulwahhab Hamad, Mehmet Kaya, "A Detailed Analysis of Optical Character Recognition Technology", International Journal of Applied Mathematics, Electronics and Computers, ISSN: 2147-82282147, 3rd September 2016

[4] Chowdhury Md Mizan, Tridib Chakraborty, Suparna Karmakar, "Text Recognition using Image Processing", ISSN No. 0976-5697,Volume 8, No. 5, May – June 2017

[5] Prasanta Pratim Bairagi,"Optical Character Recognition for Hindi", International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056 p-ISSN: 2395-0072, Volume: 05, Issue: 05, May-2018

[6] Satti, D.A., 2013, Offline Urdu Nastaliq OCR for Printed Text using Analytical Approach. MS thesis report Quaid-i-Azam University: Islamabad, Pakistan. p. 141.

[7] Mahmoud, S.A., & Al-Badr, B., 1995, Survey and bibliography of Arabic optical text recognition. Signal processing, 41(1), 49-77.

[8] Sameeksha Barve, "Artificial Neural Network Based On Optical Character Recognition", International Journal of Engineering Research & Technology (IJERT), Vol. 1 Issue 4, June – 2012, ISSN: 2278-0181