International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 08 | Aug 2020 www.irjet.net p-ISSN: 2395-0072

Optical Character Recognition using Artificial Neural Networks

Anirudh Sanjay Patil[1], Jashneet Singh Monga[2]

[1,2]Student, Dept. of Computer Science Engineering, Prof. Ram Meghe Institute of Technology and Research, Maharashtra, India

---

Abstract – With no doubt keyboarding the most time consuming and labor-intensive operation is the most familiar method of data input into the computer. Optical Character Recognition is that the machine carbon of human reading and has been the subject of intensive research for three decades. OCR can be described as Mechanical or Electronic conversion of scanned images where images are often handwritten, typewritten or printed text. It is a way of digitizing printed texts in order that they will be electronically searched and utilized in machine processes. It converts the images into machine-encoded text that can be utilized in machine translation, text-to-speech and text mining. This paper presents a simple, efficient, and low-cost approach to construct OCR for reading any document that has fix font size and style or handwritten style. To achieve effectiveness and less computational cost, OCR in this paper uses database to recognize English characters which makes this OCR very simple to manage.

Key Words: Optical Character Recognition (OCR), Artificial Neural Network, Hough Transform, ASCII characters, automation processing, bilevel images.

1. INTRODUCTION

The growth in pattern recognition has accelerated recently because of the varied emerging applications which aren't only challenging but also computationally more formidable, such perceptible in Optical Character Recognition (OCR), Document Classification, Shape Recognition, Computer Vision, Data Mining, and Biometric Authentication. The area of OCR is becoming an integral part of document scanners, and is adopted in many applications like postal processing, script recognition, banking, security (i.e. passport authentication) and language identification. The research is in progress for over half a century and therefore there are astounding outcomes with successful recognition rates for printed characters exceeding 99%. The performance advancements for handwritten running-hand character recognition has exceeded the 90% mark. Now, multiple organizations depend on OCR systems to eliminate the human interactions for better performance and efficiency. OCR system gives full alphanumeric recognition of printed or handwritten characters by simply scanning the document. Documents are scanned using a scanner and are given to the OCR systems which recognizes the characters within the scanned documents and converts them into ASCII data. In OCR a database is employed at the backend for recognition purpose. In the proposed systems the process consists of following processing steps:

(1) Scanning of Image
(2) Pre-Processing of Image
(3) Character Extraction
(4) Feature Extraction and Recognition
(5) Post-Processing.

The first step involves a scanner which is used to scan the handwritten or printed documents. The quality of the scanned document is based on the scanner. Therefore, a scanner with high speed and color quality is desirable. The recognition process involves multiple complex algorithms, pre-loaded templates and dictionary which are reviewed with the characters and the machine editable ASCII characters. The verifying is completed either randomly or chronologically by human intervention. Optical Character Recognition is assessed into two types, Offline recognition and Online recognition. In offline recognition the original document is either an image or a scanned form of the document, although in Online recognition the consecutive points are given as a function of time and order of strokes which are available. The offline recognition is explained through the following paper.
The proposed OCR system provides the following features:

(1) No more retyping

(2) Quick Digital Searches

(3) Edited Text

(4) Save Space

Optical Character Recognition deals with the matter of recognizing optically processed characters. Optical recognition is performed offline once the writing or printing has completed. Both hand-written and printed characters could also be recognized, but the performance is directly dependent upon the quality of the input documents.

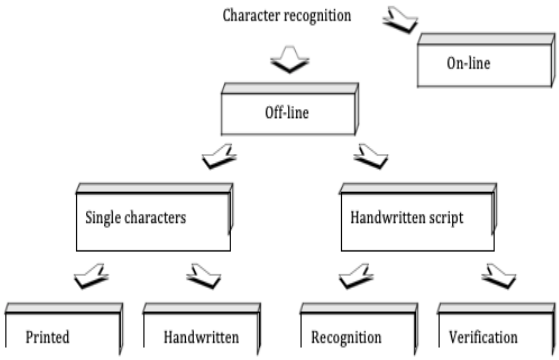

Fig -1: The different areas of character recognition

The more constrained the input is the better the performance of the OCR system. However, when it involves totally unconstrained handwriting, the OCR machines still lack as compared to the humans. However, the pc reads fast and the technical advances are constantly bringing technology closer to ideal.

2. The History of OCR

Analytically, character recognition is a sub-member of the pattern recognition area. However, it was character recognition that gave the motivation for making pattern recognition and image analysis developed fields of science.

2.1 The very first attempt

The motivation was to duplicate the human functions by machines, i.e. by making the machine able to perform tasks like reading. The origins of character recognition can be found effective back in 1870s. C.R.Carey of Boston Massachusetts innovated the retina scanner which was an image transmission system making use of a montage of photocells. Two decades later the Polish P. Nipkow innovated the sequential scanner which was a remarkable advancement eventually for modern television and reading machines.

In the initial decades of the 19th century multiple strives were made to develop devices to help the blind through trials with OCR. However, the modern version of OCR did not pop up until the middle of the 1940s with the evolution of the digital computer.

2.2 The start of OCR

By 1950 the technological rising was advancing at a high speed, and electronic data processing had become an important field. Data entry was performed through punched cards and a cost-efficient way of handling the escalating amount of data was needed. The technology for machine reading was sufficiently developed for application, and by mid 1950's OCR machines became commercially available.

The first accurate OCR reading machine was inaugurated at Reader's Digest in 1954. This machine was used to permute typewritten sales reports into punched cards for input to the pc.

2.3 First generation OCR

The merchant OCR systems materializing in the period from 1960 to 1965 are called the first generation of OCR. This generation of OCR machines were mostly distinguished by the constrained letter shapes read. These symbols were peculiarly designed for machine reading, and the initial ones didn't look very natural. Along time came multi font machines, which could study up to ten various fonts. The number of fonts it could recognize were limited by the pattern recognition method applied, template matching, which compared the character image with a library of prototype images for each character with font.

2.4 Second generation OCR

In the 1960s and early 1970s the reading machines of the second generation materialized. These systems were able to acknowledge regular machine printed characters and also had advancements in hand-printed character recognition. When the hand-printed characters were observed, the character set was contented to numerals and only a few letters and symbols.

The first and prominent system of this kind was the IBM 1287, which was revealed at the World Fair in New York in 1965. Also, in the same period Toshiba revolutionized its first automatic letter sorting machine for postal-codes and Hitachi developed the first OCR machine which provided high performance at low cost.

In this period, noteworthy work was concluded in the area of standardization. In 1966, an in-depth study of OCR requirements was accomplished and an American standard OCR character set was interpreted; OCR-A. This font was highly formalized and designed to assist optical recognition, while still readable to humans. A European font was also interpreted, OCR-B, which had more natural fonts than the American standard. And even some attempts were made to merge the two fonts into one standard, but instead of that machines being able to read both standards were materialized.

Fig -2: OCR-A(top), OCR-B(bottom)

2.5 Third generation OCR

In the 1970s, the third generation of OCR machines were fabricated, the vital challenge was documents of poor quality, large printed and hand-written character sets. The most significant objective of the machine was to maintain low cost and high performance, which were assisted by the exceptional advancements in hardware technology.

Whilst more sophisticated OCR-machines began to appear at the market, simple OCR devices were still very convenient. Within this period i.e. before the personal computers and laser printers began to overwhelm the area of text production, typing was a significant opening for OCR. The steady print spacing and a small quantity of fonts made classically designed OCR devices very convenient. Rough drafts could be created on typewriters and fed into the pc through an OCR device for final polishing. In this way word processors, which were an upscale resource at this point, could support several people and therefore the costs for equipment could be cut.

2.6 OCR today

Although, OCR machines became commercially available already in the 1950s, only a few thousand systems had been sold worldwide up to 1986. The major reason for this was the over costly systems. However, as hardware got cheaper, and the OCR systems started to become obtainable as software packages, the sales increased greatly. Nowadays, a few thousands of systems are sold each week, and the cost of an omni font OCR has dropped with a factor of ten every other year for the last 6 years.

1870	The very first attempts
1940	The modern version of OCR.
1950	The first OCR machines appear
1960 - 1965	First generation OCR
1965 - 1975	Second generation OCR
1975 - 1985	Third generation OCR
1986 ->	OCR to the people

Fig-3: A short OCR chronology

3. Methods of OCR

The main principle in automatic recognition of patterns, is first to show the machine which classes of patterns which will occur and what they appear like. In OCR the patterns are letters, numbers and a few special symbols like commas, question marks etc., while the various classes correspond to the different characters. The teaching of the machine is performed by showing the machine samples of characters of all the various classes. Based on these samples the machine builds a prototype or a description of every class of characters. Then, during recognition, the unknown characters are compared to the previously obtained descriptions, and assigned the class that provides the best match.

In most commercial systems for character recognition, the training process has been performed beforehand. However, some systems include facilities for training within the case of inclusion of latest classes of characters.

4. Components of an OCR system

A typical OCR system embodies of several components. In figure 4: a standard setup is illustrated. The first step within the process is to digitize the analog document using an optical scanner. Through the segmentation process symbols are extracted once texts are located. These extracted symbols are then preprocessed, for eliminating noise, to facilitate the extraction of features in next step.

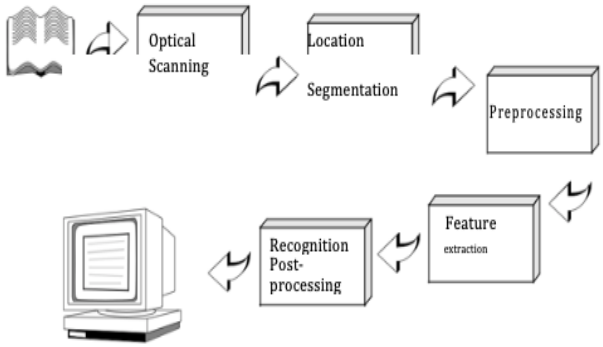

Fig-4: Components of an OCR system

4.1 Optical scanning

Through the scanning process a digital image of the original document is apprehended. In OCR optical scanners are used, which contains a transport mechanism and a sensing device that tabulates light intensity into gray-levels. Printed documents usually consist of black print on white background. Hence, when performing OCR, it is a regular practice to convert the multilevel image into a bilevel image of black and white. Often the process is referred as thresholding, which is performed on the scanner to save lots of memory space and computational effort.

The thresholding process is vital because the results of the subsequent recognition is completely dependent on the quality of the bilevel image. Still, the thresholding performed on the scanner is usually basic. A fixed threshold is employed, wherever, the gray-levels are below this threshold it is said to be black and for the levels above are said to be white. For a high-contrast document with uniform background, a fixed threshold can be sufficient to determine. However, the documents encountered in practice have a fairly substantial range in contrast. In such cases, more advanced methods for determination of thresholding are required to obtain an accurate result.

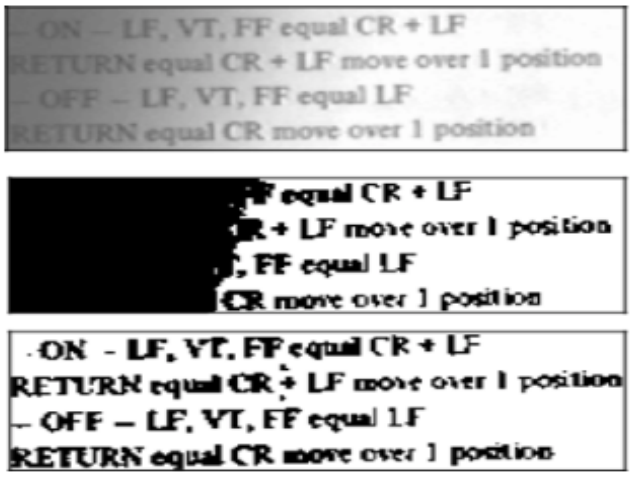

Fig-5: Problems in thresholding

The methods having the best output for thresholding are usually those which are ready to vary the threshold over the document adapting to the basic properties such as contrast and brightness. However, such methods usually depend on a multilevel scanning of the document which needs more memory and computational capacity. Therefore, these techniques are less frequently used in reference to OCR systems, although they result in better images.

4.2 Location and segmentation

Segmentation is a procedure that discovers the constituents of an image. It is important to locate the regions of the document where data is printed and then distinguish them into figures and graphics for processing. For example, when performing automatic mail-sorting, the address must be located and then segregated from the other embossing on the envelope such as stamps and company logos, prior to the recognition process.

The segmentation process is isolation of characters or words. Such that the majority of optical character recognition algorithms separate the words into isolated characters which are then recognized individually. Generally, the segmentation is performed by isolating each connected component, that is each connected black area. This technique is easy to implement, but problems when the characters touch or if the fragmented characters consist of several parts. After segmentation, it must look for mistakes such as texts considered as graphics or geometry and vice versa.

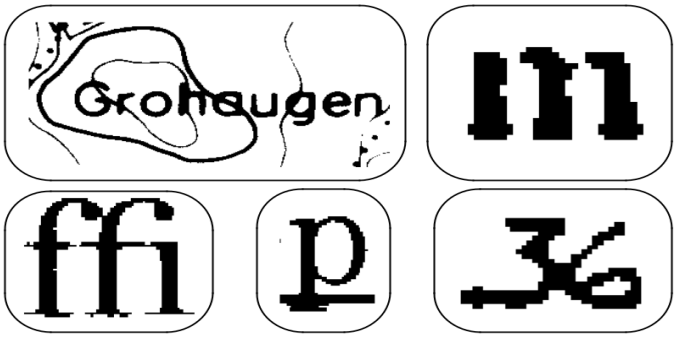

Fig-6: Degraded symbols

4.3 Preprocessing

The image obtained after the scanning process contain an explicit quantity of noise. Depending on the resolution on the scanner and therefore the success of the applied technique for deriving the thresholding value, the characters could also be soiled or broken. Due to these defects, there may be poor recognition rates, which can be eliminated by employing a preprocessor to smooth the digitized characters.

The smoothing implies each filling and thinning. Filling eliminates small breaks, gaps and holes within the digitized characters, whereas thinning reduces the span of the line. The most employed techniques for smoothing, is moving a window across the binary image of the character while

applying certain rules to the contents of the window for a better and accurate output.

In addition to smoothing, preprocessing typically includes normalization. The normalization process is applied to obtain characters of uniform size, slant and rotation. To be ready to correct for rotation, the angle of rotation should be obtained. Variants of Hough Transform are commonly used for detecting skew of rotated pages or lines of text. However, to obtain the rotation angle of a single symbol is not possible till the symbol has been recognized.

Fig-7: Normalization and smoothing of a symbol

4.4 Feature Extraction

The objective of feature extraction is to capture the essential characteristics of the symbols, and it's usually accepted that this can be one of the most faced troublesome issue of pattern recognition. The straightforward way of describing a character is by the actual raster image. Another approach can be to extract certain features that can be used to characterize the symbols, but they leave out the unimportant attributes.

The techniques for extraction of such features are often divided with features are found from into three main groups:

- The distribution of points.
- Transformations and series expansions.
- Structural analysis.

4.5 Post Processing

After the recognition stage, if there still remain unrecognized characters, then these characters are given their meaning in the post-processing stage. Additional templates can be added to the system for improving the compatibility checking within the systems database.

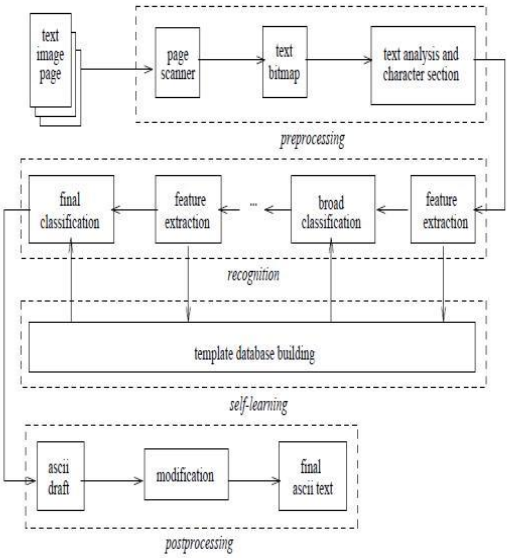

Fig -8: Method of processing

5. Applications of OCR

In the recent years there is a widespread appearance of commercial optical character recognition products which meet the requirements of different users. Three main application areas of OCR are commonly distinguished as data entry, text entry and process automation.

5.1 Data entry

The data entry area encloses technologies which enter large amounts of restricted data. Initially these document reading machines would be used for banking applications. These systems are characterized by reading solely an extremely limited set of printed characters, usually numerals and certain special symbols. They're designed to scan data like account numbers, customers identification, article numbers, amounts of money etc. The paper formats are constrained with a restricted range of fixed lines to be scanned per document.

Because of such restrictions, readers of this type might have a very high output of up to 150.000 documents per hour. Single character error and reject rates are approximately 0.0001% and 0.01% respectively. Also, because of the limited character set, these readers are remarkably tolerant to bad printing quality. These systems are specifically designed for their applications and therefore have high cost factors.

5.2 Text entry

The next branch of data reading machines is of page readers for text entry, used for office automation. Here the restrictions on paper format and character set are exchanged for constraints regarding the font and printing quality. The reading machines are employed to enter large amounts of text, typically in an exceedingly word processing environment. These page readers are in robust competition

with direct key-input and electronic exchange of data. This area of application is thus of diminishing importance.

As the character set read by these machines is quite large, the performance is very keen about the quality of the printing. However, under controlled conditions the single character error and reject rates are approximately about 0.01% and 0.1% respectively which is very considerable. The reading speed is usually in the order of a few hundred characters per second. Also, for various languages the probabilities of two or more characters appearing consecutively in a sequence can be computed, which may be utilized to detect errors more effectively.

5.3 Process automation

Within this area of application, the main concern isn't to read what is printed, but to control some processes. This is the technology required for automatic address reading for the mail sorting process. Hence, the objective is to direct each letter into the appropriate bin regardless whether the characters were correctly recognized or not. The general approach to this method is to read all the information provided and use the postcode as a redundancy check for accurateness.

The properties of the mail are very important for the acceptance rate of these systems. Therefore, this rate may vary with the percentage of handwritten mails. Even though, the reject rate for mail sorting is large, the missort rate is still close to zero. The sorting speed is around 30.000 letters per hour with a very minute error rate.

5.4 Other applications

The areas mentioned above are the ones in which OCR is proved successful and most widely used. However, the other areas of applications exist, and some of these are mentioned below.

Aid for blind.

Speech synthesis system was introduced such that a reader would enable the blind to understand printed documents. However, the major issue to this approach was the high costs of such machines. But this is certainly changing as the costs of microelectronics fall

Automatic number-plate readers.

The automatic number plate reader is a technology that uses the OCR tech on images to read and detect vehicle registration plates to create a vehicle location database. This can be employed by using road enforcement cameras or by cameras used specifically for the purpose.

6. The Future of OCR

Through the years, the methods of character recognition have prominently improved from quite primitive schemes, which were suitable only for reading stylized printed numerals, to those which are more complex and sophisticated for the recognition of a great variety of typeset fonts and also handprinted character sets. Below the future of OCR when it comes to both research and areas of applications, is briefly discussed.

6.1 Future Improvements

New methods for character recognition are still expected to appear, with the decreasing computational restrictions and the developing computer technology new approaches open up. With the future developments there might be a potential in performing character recognition techniques directly on grey level images. However, the greatest potential still seems to lie in exploiting of existing methods and by mixing the current methodologies and making more use of context.

The integration of segmentation and contextual analysis can not only improve recognition of split characters but also joined characters. Actually, there is a potential in using context to a much larger extent than at what level it is done today. In addition, the combinations of multiple independent feature sets and classifiers, may improve the recognition of individual characters as the weakness of one method may get compensated by the strength of another method.

The research in character recognition has advanced towards the recognition of cursive script which is handwritten connected or calligraphic characters. Promisingly the techniques in this area, deal with recognition of entire words rather than individual characters.

6.2 Future Needs

Today optical character recognition is most successful for constrained material having fixed character sets to deal with or that is documents produced under some control. However, in the future it seems that the need for constrained OCR will be decreasing. The major concern for this will be that control of the production process which means that the document is created from material prestored on the computer. Hence, if a computer readable version is already available, then this means that data can be exchanged electronically or printed in a more computer readable form, for instance barcodes so the constrained OCR systems won't be required.

The applications for future OCR systems exist in recognition of documents wherever the control over production process is not possible. This could be the material where the recipient is cut off or interrupted from an electronic version and also has no control on the production process or any older material which could not be generated electronically at production time. This also makes certain that the future OCR

systems that are intended for reading printed text must be omni font.

Further consideration of important area for OCR development can be said as the recognition of manually produced documents. Considering the example of postal applications, OCR must focus on reading addresses produced by people without any access to computer technology. Already, due to advancements it isn't unusual for companies, which have access to computer technology to do mail marking with the help of barcodes. Therefore, the relative importance of handwritten text recognition is expected to increase.

7. Artificial Neural Network

There are multiple approaches to solve an optical character recognition problem. One among the most common and popular approaches is relies on neural networks, which might be applied to various tasks, like pattern recognition, statistics and time-period prediction, function approximation, clustering, etc. An Artificial Neural Network (ANN) is an information processing (IP) paradigm that is influenced by the working of human brain. An ANN is designed for a particular application, like pattern recognition or data classification, through learning process which develops over time. Learning mostly involves adjustments directly to the synaptic connections that exist between the neurons.

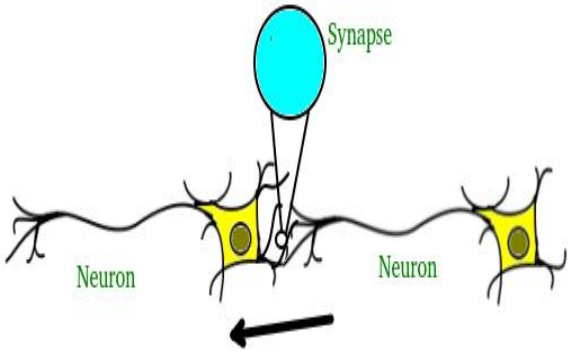

Fig -9: A Neuron

The brain consists of hundreds of billions of cells known as neurons. These neurons are connected along with the help of synapses which are nothing but the connections through which a neuron can send an impulse to the neighbor chain neuron. Once a neuron sends an excitatory signal to another neuron, then the same signal is going to be added to all or any inputs of that neuron. If it exceeds a given threshold then it'll cause the target neuron to fire an action signal forward — this is often how the thinking process works internally.

In the Computer Science field, we first model this process by creating "networks" on a computer with the help of matrices. These constructed networks can be understood as abstraction of neurons without considering the biological complexities. But to keep things simple, we will just construct a model of simple NN, which has two layers capable of solving linear classification problem.

Basically, there are 3 different layers in a neural network: -

Input Layer (All the input data is fed into the model from this layer)

Hidden Layers (There can be more than one hidden layer which are used for processing the inputs received from the input layers)

Output Layer (After processing, the output data is made available at the output layer)

Following is the manner in which these layers are laid

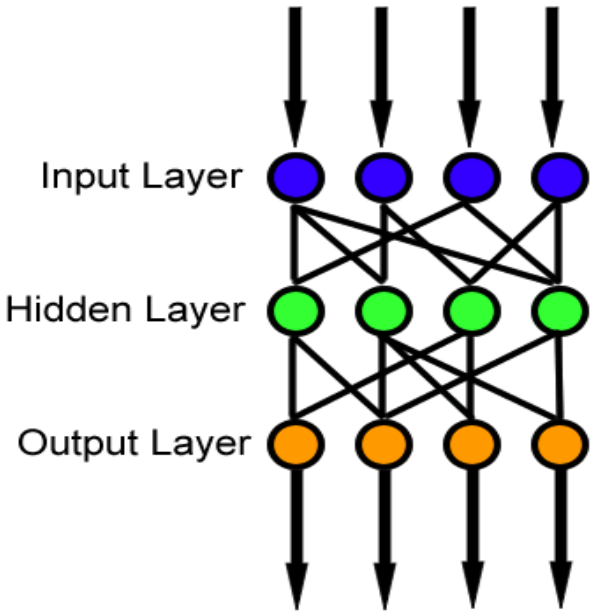

Fig -10: Depicting the different layers of a neural network

Input Layer

The most important task of the Input layer is to communicate with the external environment that generates a pattern for the neural network to work on. Its job is to deal with all the types of inputs. And to make sure that the input gets transferred to the hidden layers. The input layer should also represent the condition for which we are training the neural network to get optimum result.

Every input neuron should represent an independent variable that will have an influence on the output produced by the neural network.

Hidden Layer

The intermediate layer found between the input and output layer is known as hidden layer which, is responsible for the collection of neurons that have an activation function applied to solve the input data. Its job is to process the inputs generated as input by its previous layer. So, it is the layer which is in charge of extracting all the required features from the generated input data. Many researches have been made in evaluating the number of neurons in the hidden layer but still none of them was successful in finding the accurate result. In a neural network there can be multiple hidden layers. Therefore, the number of hidden layers required to solve a certain kind of problem will always differ. To find the accurate number of hidden layers required to solve a certain problem one must suppose, that if we have data which can be separated linearly, then there will be no need to use hidden layer as the required activation function can also be implemented to input layer which will solve the problem. But if we get a certain problem which deals with complex decisions then, we can use 3 to 5 hidden layers which will be proportional to the degree of complexity of the problem or the degree of accuracy required to solve the in-hand problem. That certainly does not mean that the accuracy of neural network keeps increasing if we keep on increasing the number of layers. We may also come across a stage when the accuracy becomes constant or falls if we add any extra layer. Also, there is a need to calculate the number of neurons in each network. So that if the number of neurons is less compared to the complexity of the problem data then there will be fewer neurons required in the hidden layers to sufficiently detect the signals in any data set. If there are unnecessarily more neurons present in a network then the phenomenon of Overfitting may occur. Multiple methods are considered to obtain the exact formula required for calculating the number of hidden layer as well as number of neurons in each hidden layer.

Output Layer

The function of the output layer of the neural network is to collect and transmit the data in a way that it has been designed to display. The output layers obtained pattern can be directly traced back to the input from the input layer. The number of neurons in output layer is directly proportional to the kind of work that is performed by the neural network. Therefore, to determine the number of neurons required by the output layer, we must first consider the kind of use of the neural network.

9. CONCLUSIONS

The artificial neural network systems based has shown some promising results due to the fact that despite trained on a single set of templates (independent of predefined fonts) in not only get trained in seconds but can also recognize the fonts for which it was not trained with high efficiency. The feature extraction step of optical character recognition can be used with existing OCR methods, especially for English text.

OCR being a very remarkable technology holds a lot of potential. OCR can be considered a good example for explaining how AI solutions are driving database advancements due to their affordability and accessibility.

REFERENCES

[1] "α-Soft: An English Language OCR", 2010 Second International Conference on Computer Engineering and Applications. Junaid Tariq, Umar Nauman Muhammad Umair Naru.
[2] "A Review on the Various Techniques used for Optical Character Recognition", Pranob K Charles, V.Harish, M.Swathi, CH. Deepthi
[3] International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622, Vol. 2, Issue 1,Jan-Feb 2012.
[4] Miletzki, Siemens Electrocom GmbH D-78767 Konstanz, Germany.
[5] "Character Recognition in practice Today and Tomorrow", 1996, Udo
[6] "Prototype Extraction and Adaptive OCR" IEEE Transaction on pattern analysis and Machine Intelligence, VOL. 21, NO. 12, DECEMBER 1999, Yihong XU, Member, IEEE, George Nagy, Senior Member, IEEE.
[7] "Contextual Focus for Improved Recognition of Hand-Filled Forms",
[8] 1999. Wing Seong Wong, Nasser Sherkat, Tony Allen IRIS, Department of Computing.
[9] "Image processing Algorithms for Improved Character Recognition and Components Inspection", 2009 World Congress on Nature & Biologically Inspired Computing (NaBIC 2009), Anima Majumder.
[10] "A System for Automated Data Entry from Forms", 1996 IEEE Proceedings of ICPR '96, Raymond A. Lorie, V. P. Riyaz , Thomas K. Truong.
[11] Combination of Document Image Binarization Techniques", 2011
[12] International Conference on Document Analysis and Recognition.
[13] ICAR: Identity Card Automatic Reader, 200 1 IEEE, Josep Lladbs, Felipe Lumbreras, Vicente Chapaprieta, Joan Queralt.
[14] "Implementing Optical Character Recognition on the Android Operating System for Business Cards", IEEE 2010, Sonia Bhaskar, Nicholas Lavassar, Scott Green EE 368 Digital Image Processing.
[15] "Document Analysis and Recognition", 2005. Eighth International Conference on 29 Aug.-1 Sept. 2005, Alon, Jonathan.
[16] Pre-processing Techniques in Character Recognition, Yaseer Alginahi.

BIOGRAPHIES

Anirudh Sanjay Patil
Pursuing Bachelor of Engineering.
(Computer Science & Engineering).

Anirudh Sanjay Patil
Pursuing Bachelor of Engineering.
(Computer Science & Engineering).