Review Article | Impact Factor: 0.621 | ISSN: 2319-507X
Sachin Sawalkar, IJPRET, 2014; Volume 2 (9): 345-355 | IJPRET

INTERNATIONAL JOURNAL OF PURE AND APPLIED RESEARCH IN ENGINEERING AND TECHNOLOGY

A PATH FOR HORIZING YOUR INNOVATIVE WORK

A REVIEW OF OPTICAL CHARACTER RECOGNITION TECHNIQUES

SACHIN A. SAWALKAR[1], DR. R. M. DESHMUKH[2]

1. Department of Electronics & Telecommunication, IBSS college of Engineering, Amravati.
2. Professor & Head, Department of Electronics & Telecommunication, IBSS college of Engineering, Amravati.

Accepted Date: 27/02/2014 ; Published Date: 01/05/2014

Abstract: OCR is the acronym for Optical Character Recognition. This technology allows a machine to automatically recognize characters through an optical mechanism. Human beings recognize many objects in this manner our eyes are the "optical mechanism."But while the brain "sees" the input, the ability to comprehend these signals varies in each person according to many factors. By reviewing these variables, we can understand the challenges faced by the technologist developing an OCR system. Handwriting recognition has been one of the most interesting and challenging research areas in field of image processing and pattern recognition in the recent years. This paper describes the techniques for converting textual content from a paper document into machine readable form. The computer actually recognizes the characters in the document through a revolutionizing technique called Optical Character Recognition. Several techniques like OCR using correlation method and OCR using neural network or OCR using Matlab.

Keywords: Optical character recognition, feature extraction, template matching, structural analysis, learning and practical OCR systems.

Corresponding Author: MR. SACHIN A. SAWALKAR

Access Online On:

www.ijpret.com

How to Cite This Article:

Sachin Sawalkar, IJPRET, 2014; Volume 2 (9): 345-355

PAPER-QR CODE

INTRODUCTION

Recognition of Latin-script, typewritten text is still not 100% accurate even where clear imaging is available. One study based on recognition of 19th- and early 20th-century newspaper pages concluded that character-by-character OCR accuracy for commercial OCR software varied from 71% to 98%; total accuracy can be achieved only by human review. Other areas including recognition of hand printing, cursive handwriting, and printed text in other scripts (especially those East Asian language characters which have many strokes for a single character) are still the subject of active research. Accuracy rates can be measured in several ways, and how they are measured can greatly affect the reported accuracy rate. For example, if word context (basically a lexicon of words) is not used to correct software finding non-existent words, a character error rate of 1% (99% accuracy) may result in an error rate of 5% (95% accuracy) or worse if the measurement is based on whether each whole word was recognized with no incorrect letters.

On-line character recognition is sometimes confused with Optical Character Recognition (see Handwriting recognition). OCR is an instance of off-line character recognition, where the system recognizes the fixed static shape of the character, while on-line character recognition instead recognizes the dynamic motion during handwriting. For example, on-line recognition, such as that used for gestures in the Pinpoint OS or the Tablet PC can tell whether a horizontal mark was drawn right-to-left, or left-to-right. On-line character recognition is also referred to by other terms such as dynamic character recognition, real-time character recognition, and Intelligent Character Recognition or ICR.

On-line systems for recognizing hand-printed text on the fly have become well known as commercial products in recent years (see Tablet PC history). Among these are the input devices for personal digital assistants such as those running Palm OS. The Apple Newton pioneered this product. The algorithms used in these devices take advantage of the fact that the order, speed, and direction of individual lines segments at input are known. Also, the user can be retrained to use only specific letter shapes. These methods cannot be used in software that scans paper documents, so accurate recognition of hand-printed documents is still largely an open problem. Accuracy rates of 80% to 90% on neat, clean hand-printed characters can be achieved, but that accuracy rate still translates to dozens of errors per page, making the technology useful only in very limited applications.

Recognition of cursive text is an active area of research, with recognition rates even lower than that of hand-printed text. Higher rates of recognition of general cursive script will likely not be

possible without the use of contextual or grammatical information. For example, recognizing entire words from a dictionary is easier than trying to parse individual characters from script. Reading the Amount line of a cheque (which is always a written-out number) is an example where using a smaller dictionary can increase recognition rates greatly. Knowledge of the grammar of the language being scanned can also help determine if a word is likely to be a verb or a noun, for example, allowing greater accuracy. The shapes of individual cursive characters themselves simply do not contain enough information to accurately (greater than 98%) recognize all handwritten cursive script.

It is necessary to understand that OCR technology is a basic technology also used in advanced scanning applications. Due to this, an advanced scanning solution can be unique and patented and not easily copied despite being based on this basic OCR technology. For more complex recognition problems, intelligent character recognition systems are generally used, as artificial neural networks can be made indifferent to both affine and non-linear transformations. A technique which is having considerable success in recognizing difficult words and character groups within documents generally amenable to computer OCR is to submit them automatically to humans in the RECAPTCHA system.

1] CORRELATION METHOD FOR SINGLE CHARACTER RECOGNITION:

A. Preprocessing: The image is taken and is converted to gray scale image. The gray scale image is then converted to binary image. This process is called Digitization of image. Practically any scanner is not perfect, the scanned image may have some noise. This noise may be due to some unnecessary details present in the image. So, all the objects having pixel values less than 30 are removed. The denoised image thus obtained is saved for further processing. Now, all the templates of the alphabets that are preredesigned are loaded into the system.

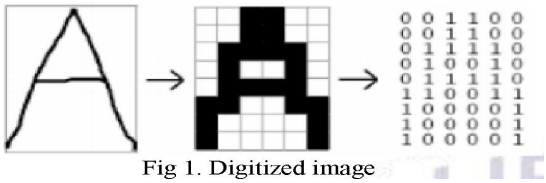


Fig 1. Digitized image

B. Segmentation: In segmentation, the position of the object i.e., the character in the image is found out and the size of the image is cropped to that of the template size.

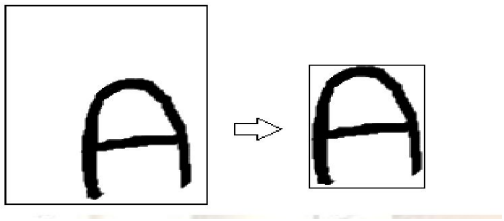


Fig 2. Segmented image

C. Recognition: The image from the segmented stage is correlated with all the templates which are preloaded into the system. Once the correlation is completed, the template with the maximum correlated value is declared as the character present in the image.

D. Conclusion: In correlation method there are many unnecessary comparisons and the efficiency of recognition is same for a particular pattern and the given set of templates. However extra templates can be added to the system for providing a wide range of compatibility but doing so will increase the computational intensity of the system. Another important drawback of this method is it requires lot of memory and execution time.

2] CORRELATION METHOD FOR CONTINUOUS CHARACTER RECOGNITION:

A. Preprocessing: A noisy image is read from the scanner and is converted to a binary image. The noise is removed from the image by removing all details of the image less than 30 pixels. Then the image is segmented by splitting the image into lines, each line representing a row of words in the image each with a separate label for identification. Now each line consists of different number of words each with many numbers of letters. Each letter should be separated and resized to the size of the preloaded templates. The recognition process is similar to 'correlation method for single character recognition'.

B. Creating Templates: Images from A to Z and numbers from 1 to 9 are taken into different variables and are preprocessed. All these variables are stored in the form of a cell in which each sub matrix represents a letter. The same process is done for printed upper case, printed lower case, printed numbers, hand written upper case, hand written lower case and hand written numbers. All the model inputs are saved under the same variable name like ‘templates.dat’ to the hard disk.

C. Dividing into lines: The image is first clipped and an array containing the coordinates of non-zero elements is found. Then the empty row i.e., the row with all elements as zero is found, this row is taken as the demarcation line to separate the top line from all the lines below it.

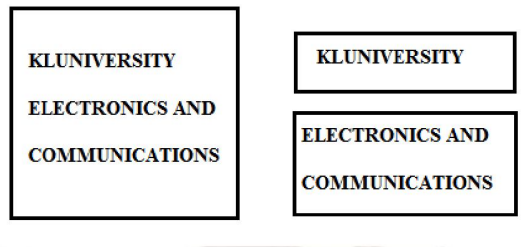


Fig 3. Dividing the image into lines

D. Conclusion: This method has the same disadvantages as that of 'correlation method for single character recognition'.

3. OCR USING ARTIFICIAL NEURAL NETWORKS:

A. Artificial Neural Networks: Artificial Neural Networks (ANN) can be likened to collections of identical mathematical models that emulate some of the observed properties of biological nervous systems and draw on the analogies of adaptive biological learning. The key element of an Artificial Neural Network is its structure. It is composed of a number of interconnected processing elements tied together with weighted connections, which take inspiration from biological neurons. The ability to make decisions about imprecise input data makes it useful as a medical analysis tool. There is no need to provide a specific algorithm on how to identify the disease when using a neural network. Neural networks learn by example so the details of how to recognize the disease are not needed. What is needed is a set of examples that is representative of all the variations of the disease. The quality of examples is not as important as the quantity.

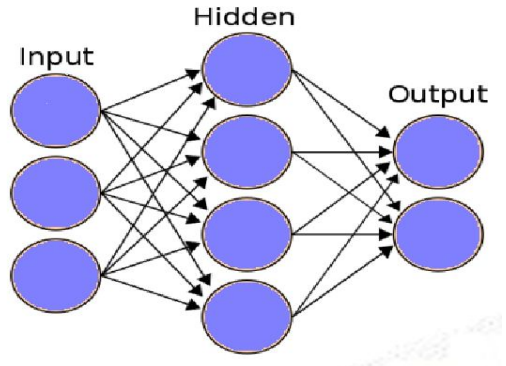


Fig 4. Artificial Neural Network

The Artificial Neural Network can be trained into two main group that are supervised and unsupervised learning. In supervised learning, the network learns by example whereas in unsupervised method no target value or example is given. Unsupervised learning is very difficult and complex to implement.

The neural network receives 35 Boolean values as 35-element input vector. It is then required to identify letter by responding with a 26-element output vector. The 26-elements of output vector each represent a letter. To operate correctly the network should respond with a '1' in position of letter being represented in network. All other values in output vectors should be '0'. In addition, the network should be able to handle noise. In practice, the network doesn't receive a perfect Boolean vector as input. Specifically, the network should make as few mistakes as possible when classifying vector with noise of mean 0 and standard deviation of 0.2 or less.

B. Architecture: The neural network needs 35 input and 26 neurons in its output layer to identify the letters. The network is a two layer "log-sigmoid" network. The log-sigmoid T.F is picked, as its output ranges from 0 to 1 is perfect for learning to output Boolean values. The hidden layer has 25 neurons. This number was picked by trial and error.

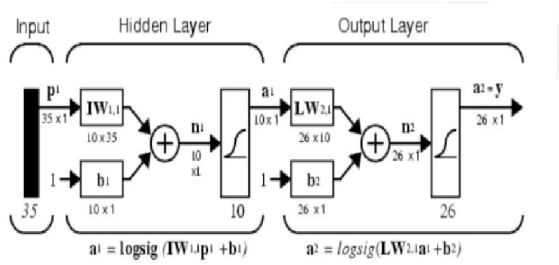


Fig 5.Architecture of back-propagation neural network.

C. Training: A two-layer network is created and training is done with and without noise. All training is done using back propagation with both adaptive learning rate and momentum.

4. OPTICAL CHARACTER USING MATLAB: Optical character recognition (OCR) is an important research area in pattern recognition. The objective of an OCR system is to recognize alphabetic letters, numbers, or other characters, which are in the form of digital images, without any human intervention. This is accomplished by searching a match between the features extracted from the given character's image and the library of image models. Ideally, we would like the features to be distinct for different character images so that the computer can extract the correct model from the library without any confusion. At the same time, we also want the

features to be robust enough so that they will not be affected by viewing transformations, noises, resolution variations and other factors. Figure 1.1 illustrates the basic processes of an OCR system.

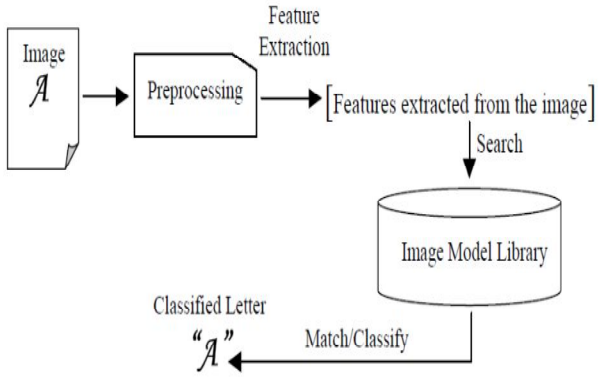


Figure1.1.Basic Processes of an OCR

A. Need For OCR using Matlab Software: There are numerous benefits to using OCR. You need OCR software to make all your scanned documents searchable. This applies especially to scanned documents that contain a lot of addresses, names and numerical values. You would not want to manually search the document as if you were scanning a physical document; that defeats the purpose of digitizing the document in the first place. You also need OCR software to reduce expenses with bookkeeping and other related activates. PDF files do not require experts; anyone can be taught to efficiently use and relay the files when needed. You also need OCR software to make various file types searchable; even TIFF files can be subsumed under a good OCR engine.

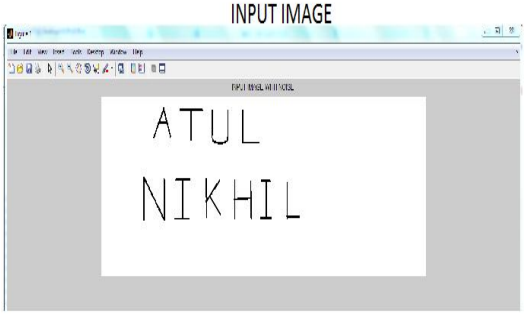
B. Matlab code for OCR:

```
clc,

all,imagen=imread('TEST_11.jpg');

imshow(imagen);

title('INPUT IMAGE WITH NOISE')

if size(imagen,3)==3

   imagen=rgb2gray(imagen);
```

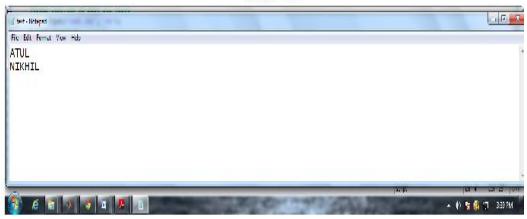
```
end
threshold = graythresh(imagen);
imagen =~im2bw(imagen,threshold);
imagen = bwareaopen(imagen,30);
word=[ ];
re=imagen;
fid = fopen('text.txt', 'wt');
global templates
load templates
num_letras=size(templates,2);
while 1
[fl re]=lines(re);
   imgn=fl;
   [L Ne] = bwlabel(imgn);
   for n=1:Ne
[r,c] = find(L==n);
n1=imgn(min(r):max(r),min(c):max(c));
img_r=imresize(n1,[42 24]);
      letter=read_letter(img_r,num_letras);
word=[word letter];
   end
fprintf(fid,'%s\n',word);%Write 'word' in text file (upper)
word=[ ];
```

```
if isempty(re)
      break
   end
end
fclose(fid);
winopen('text.txt')
clear all
```

INPUT IMAGE



OUTPUT



C. Advantages:

No more retyping. If you lose or accidentally erase an important digital file, such as a proposal or invoice, but still have a hard copy, you can easily replace it in your digital filing system by using OCR software to scan the paper original or most recent draft.

Quick digital searches.OCR software converts scanned text into a word processing file, giving you the opportunity to search for specific documents using a keyword or phrase. For example, you could effortlessly search hundreds of invoices and locate a specific name or account in moments, without having to thumb through extensive files.

Edit text. Once you've scanned your document using OCR, you have the option to edit the text within a word processing program of your choice. Scan items that may need to be updated in the future to help expedite the editing process

– Typed family recipes

– Rental agreements

– Resumes

– Contracts

Save space. free up storage space by scanning paper documents and hauling the originals off to storage. You can easily turn a filing cabinet worth of information into editable digital files, and create a backup system consisting of a single CD.

D. Comparison Table in between OCR &OMR:

Item	OCR	OMR
Handprint recognition	Y	N
Machine print recognition	Y	N
Recognition of checks and "X"s	Y	Y
Requires timing tracks/ form IDs		Y
Requires registration marks	Y	N
Electronic image storage and Retrieval	Y	N

5. CONCLUSION

A number of techniques that are used for optical character recognition have been discussed which uses correlation and neural networks. Much other advancement in Optical Character Recognition are being under development. The main research is currently going on in extending Optical Character Recognition to all the popular native languages of India like Hindi, Telugu, Tamil etc., Recognition system works well for simple language like English. It has only 26 character sets. And for standard text there are 52 numbers of characters including capital and small letters. But a complex but organized language like Telugu, OCR system is still in preliminary level. The reason of its complexities are its characters shapes, its top bars and end bars more over it has some modified, vowel and compound characters and also one of the important reasons for poor recognition in OCR system is the error in character recognition.

REFERENCES:

1. Sang Sung Park, Won Gyo Jung, Young Geun Shin, Dong-Sik Jang, Department of Industrial System and Information Engineering, Korea University, South Korea, "Optical Character System Using BP Algorithm".

2. Ahmad M. Sarhan, and Omar I. Al Helalat, "Arabic Character Recognition using Artificial Neural Networks and Statistical Analysis".

3. Arun K Pujari, Prof. C Dhanunjaya Naidu, AI Lab, University of Hyderabad "An Ad aptive Character Recognizer for Telugu Scripts using Multiresolution Analysis a Associative Memory". UNESCAP, Pop-IT project, 1997-2001

4. S. Mori, C.Y. Suen and K. Kamamoto, "Historical review of OCR research and development," Proc. of IEEE, vol. 80, pp. 1029-1058, July 1992.

5. S. Impedovo, L. Ottaviano and S. Occhinegro, "Optical character recognition", International Journal Pattern Recognition and Artificial Intelligence, Vol. 5(1-2), pp. 1-24, 1991

6. R. Plamondon and S. N. Srihari, "On-line and off- line handwritten character recognition: A comprehensive survey," IEEE. Transactions on Pattern Analysis and Machine Intelligence, vol. 22, no. 1, pp. 63-84, 2000.

7. N. Arica and F. Yarman-Vural, "An Overview of Character Recognition Focused on Off-line Handwriting", IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews, 2001, 31(2), pp. 216 – 23.