ISSN:0975-9646

Optical Character Recognition Implementation Using Pattern Matching

Faisal Mohammad, Jyoti Anarase, Milan Shingote, Pratik Ghanwat

ISB&M School of Technology, Nande, Pune- 412115

***Abstract*-This paper represent an algorithm for implementation of Optical Character Recognition (OCR) to translate images of typewritten or handwritten characters into electronically editable format by preserving font properties. OCR can do this by applying pattern matching algorithm. The recognized characters are stored in editable format. Thus OCR make the computer read the printed documents discarding noise.**

***Keywords*- Character recognition, feature extraction, pattern matching, training.**

1. INTRODUCTION

Optical character recognition (OCR) is a process of converting a printed document or scanned page into ASCII characters that a computer can recognise. Computer systems equipped with such an OCR system improve the speed of input operation, decrease some possible human errors and enable compact storage, fast retrieval and other file manipulations. The range of applications include postal code recognition, automatic data entry into large administrative systems, banking, automatic cartography and reading devices for blind.

Accuracy, flexibility and speed are the main features that characterise a good OCR system. Several algorithms for character recognition have been developed based on feature selection**.** Some of them have been found commercially viable and have gone into production like OmniPage, Wordscan, TypeReader etc. The performance of the systems have been constrained by the dependence on font, size and orientation.

The recognition rate in these algorithms depends on the choice of features. Most of the existing algorithms involve extensive processing on the image before the features are extracted that results in increased computational time.

In this paper, we discuss a pattern matching based method for character recognition that would effectively reduce the image processing time while maintaining efficiency and versatility. The parallel computational capabilities of neural networks [1, 3, 5] ensures a high speed of recognition which is critical to a commercial environment. The key factors involved in the implementation are: an optimal selection of features which categorically defines the details of the characters, the number of features and a low image processing time.

2. OCR SYSTEM DESIGN

The main functional modules in our OCR systems are: image acquisition module, pre-processing module, and feature extraction module and pattern generation. The main task of image acquisition module is to obtain text image from a scanner or a pre-stored image file. It is called 'image' because scanner inherently scans pixel of the text and not characters when patterns are scanned and digitised, the data may carry some unwanted noise. For example, a scanner with low resolution may produce touching line segments and smeared images. A pre-processor [3,4] is used to smooth the digitised characters. Moreover, the system must be able to handle touching characters, proportional spacing, variable line spacing and change of font style in the scanned text, in addition to the problems of multi-fonts.

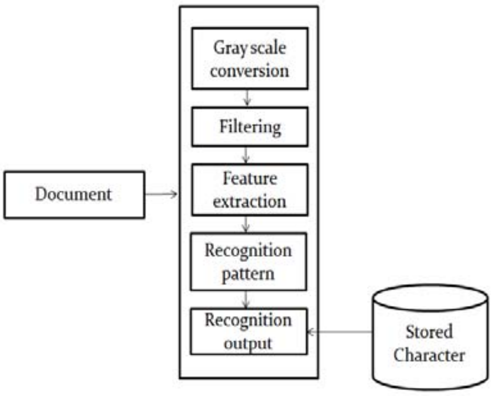


Figure 1. System Block Diagram

2.1 Grayscale

Grayscale images have many shades of gray. Grayscale images is result of measuring intensity of each pixel. For achieving accuracy input document should be grayscaled. To convert a colour from a colourspace based on an RGB colour model to a grayscale representation following function is used

$$Y = 0.2126R+0.7152G+0.0722B$$

Figure 2 and 3 shows an image before and after grayscaling respectively.

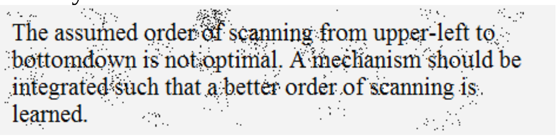

The assumed order of scanning from upper-left to bottomdown is not optimal. A mechanism should be integrated such that a better order of scanning is learned.

Figure 2 Before Grayscaling

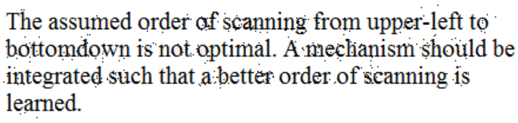

The assumed order of scanning from upper-left to bottomdown is not optimal. A mechanism should be integrated such that a better order of scanning is learned.

Figure 3 After Grayscaling

2.2 Feature Extraction

Feature extraction [4] is the process of getting information about an object or a group of object in order to facilitate classification. This is an important part in our system.

The input document may contain several lines of text that needs to be categorized into single character for recognition. For this purpose the following steps are to be applied:

1. The document is to be scanned for the initial darker pixel to be named as top of the row.
2. Now for bottom the next blank line is detected. The area between this is row of characters in image.

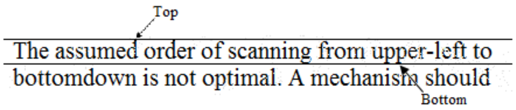

The assumed order of scanning from upper-left to bottomdown is not optimal. A mechanism should

Figure 4 Row Detection

3. Now each character is to be identified for the row obtained earlier. This is done by scanning the row vertically from top to bottom, the first darker pixel detected is the leftmost (left) pixel of character.
4. Now if all pixel are found to be blank then this is right of character.

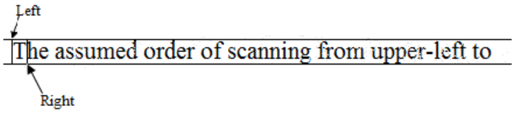

The assumed order of scanning from upper-left to

Figure 5 Boundary

5. The character from the scanned image is normalised from any pixel size to 15 X 15 pixel. It cropped the image by using top, left, right, and bottom boundaries as in figure 6.

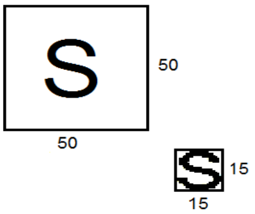
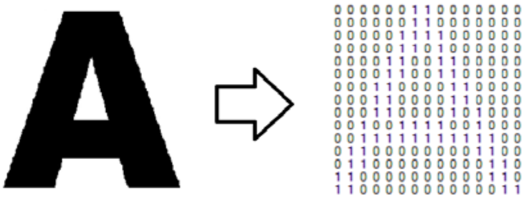


Figure 6 Scaling

6. Now the cropped image of 15 X 15 can be binarized into array of 15 X 15, where black representing 1 and white representing 0 as shown in figure 7.



```
0 0 0 0 0 0 1 1 0 0 0 0 0 0 0
0 0 0 0 0 0 1 1 0 0 0 0 0 0 0
0 0 0 0 0 1 1 1 1 0 0 0 0 0 0
0 0 0 0 0 1 1 0 1 0 0 0 0 0 0
0 0 0 0 1 1 0 0 1 1 0 0 0 0 0
0 0 0 0 1 1 0 0 1 1 0 0 0 0 0
0 0 0 1 1 0 0 0 0 1 1 0 0 0 0
0 0 0 1 1 0 0 0 0 1 1 0 0 0 0
0 0 0 1 1 0 0 0 0 1 0 1 0 0 0
0 0 1 0 0 1 1 1 1 0 0 1 0 0 0
0 0 1 1 1 1 1 1 1 1 1 1 1 0 0
0 1 1 0 0 0 0 0 0 0 0 1 1 0 0
0 1 1 0 0 0 0 0 0 0 0 0 1 1 0
1 1 0 0 0 0 0 0 0 0 0 0 1 1 0
1 1 0 0 0 0 0 0 0 0 0 0 0 1 1
```

Figure 7 Binarization of character

2.3 Recognition of Pattern

Pattern based recognition require matching of generated binary format with the existing template for this purpose the binary has been divided into 5 tracks and each track subdivided into 8 sectors. A corresponding track-sector matrix is to be generated, identifying number of pixels in each region.

This procedure is shown in figure 8.

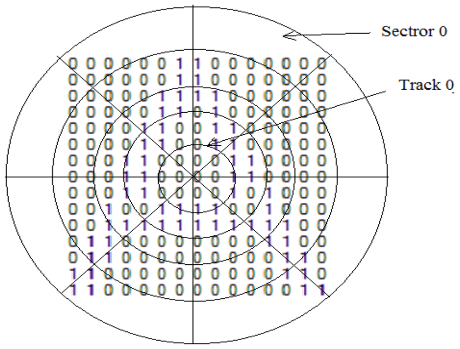


Figure 8 Division into tracks and sectors

This can be done using following procedure

1. Identify center of matrix
2. Calculate radius say rad by finding pixel with maximum distance from center using distance formulae.
   Dist= $\sqrt{((y2-y1)^2+(x2-x1)^2)}$
3. Perform (rad ÷ 5) to identify size of each imaginary track.
4. Identify imaginary sectors.
5. Generate track- sector matrix by calculating number of 1's in each intersection of sector and track.

2.4 Recognition of Output

The track-sector matrix generated above is then matched with existing template. The existing template consist of each track-sector intersection value, each track value and each sector value. If all these parameters are found to match with the template values then the resultant is the character identified. The resultant matrix contain unique value for each font and thus makes it easy to identify each font separately.

3. Research Results

The recognition rate for character images of same font used of up scaling is almost 100%.However, for down scaling the recognition rate reduces. Algorithm was tasted for handwritten characters where two observation affects the recognition rate.

1. People tend to use different fonts than the algorithm has been trained on.
2. Characters may have been written in bad handwriting.

4. Conclusion

We have shown that Pattern Matching can be implemented successfully in optical character recognition. The system has image pre and post processing modules for text image. The experiment result shows recognition rate is 70% for noisy data to up to 75%. Further work is initiated for multiple font and size characters and hand written character recognition.

REFERENCES

[1] Rokus Arnold, Poth Miklos," Character Recognition Using Neural Networks", *IEEE Computer 978-1-4244-9280-0/10.*

[2] S. T. Kahan, T. Pavlidis, and W. Baird,"On recognition of printed characters of any font and size", *IEEE Transactions* of *Pattern Recognition and Machine Intelligence,PAMI-91987,pp.274-285.*

[3] B. Hussain, and M. R. Kabuka, "A novel feature recognition neural network and its application to character recognition", *IEEE Transactions* of *Pattern Recognition and Machine Intelligence,* Vol. 16, No. 1, 1994,pp.98-106.

[4] Nallasamy Mani and Bala Srinivasan," Application of Artificial Neural Network Model for Optical, Character Recognition",*IEEE Computer 0-7803-4053-1/97*

[5] H.I. Avi-Itzhak, T.A. Diep, and H. Garland, "High accuracy optical character recognition using neural networks with centroid dithering", *IEEE Transactions* of *Pattem Recognition and Machine Intelligence,* Vol. 17, No.2, 1995, pp.218-224.