# Unicode Optical Character Recognition Using Neural Networks

**Bhaviya Rajesh Gandani, Ami Rajesh Gandani, Kushal Dharmesh Doshi, Ankur Naishadh Gandhi**

***Abstract*** **— The current probe shows how to convert the text documents into the software translated Unicode text. Unicode is nothing but the encoding standard in which each letter, number or symbol is assigned a unique numeric value which may be used with different languages. It involves scanning the document and then recognizing each and every character of the printed text so that it can be converted to Unicode. Unicode encoding technique is been used in this probe as it uses 16 bits for each character; in contrast to ASCII technique which only uses 7 bits per character.**

***Keywords*****—Unicode Character Recognition, Neural Networks, OCR, Image Processing**

## I. INTRODUCTION

Character Recognition is the method of translating an image; which can be in the form of type written, printed text or handwritten; into the format which is understood by the machines. It involves two things – scanner to scan the print image and software to analyze the characters in that printed image. Once the document is scanned, it is analyzed for the light and dark areas in order to identify each character. Unicode technique is been used to recognize each character as ASCII (American Standard Code for Information Exchange) technique supports only 128 characters whereas Unicode technique supports 65536 characters.

Fig.1 shows how the Unicode can be used in converting the scripts in different languages. In English, it's mentioned as 'I can eat glass and it doesn't hurt me' and the same sentence is been written in different languages with the help of the software that converts every character in the sentence into the machine encoded form.

***Manuscript received Sept, 2015.***

***Bhaviya Rajesh Gandani****, Dept. Of Computer Science, Theem College of Engineering Mumbai, India, +91-7875337948*

***Ami Rajesh Gandani****, PGDBM & M.Com, Rustomjee Business School/University of Mumbai Mumbai, India, +91-9096095664*

***Kushal Dharmesh Doshi****, Dept. Of Information Technology, D.J. Sanghvi of Engineering Mumbai, India, +91-7303919156*

***Ankur Naishadh Gandani****, Dept. Of Computer Science, Theem College of Engineering Mumbai, India, +91-9920797615*

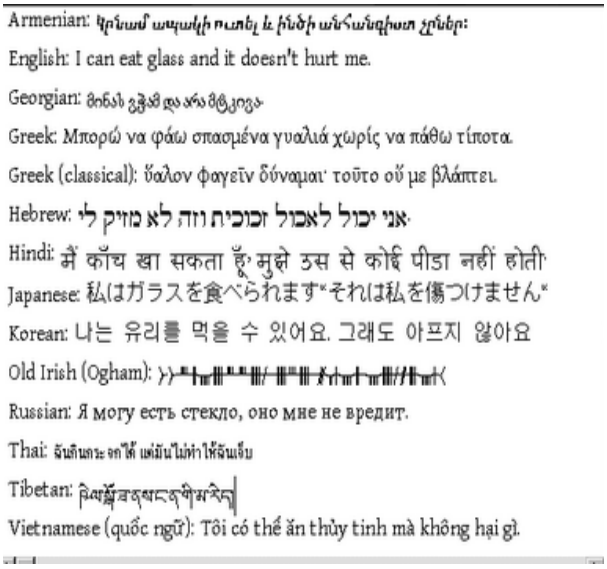


Fig.1: One sentence in different languages using Unicode

## II. METHODOLOGY

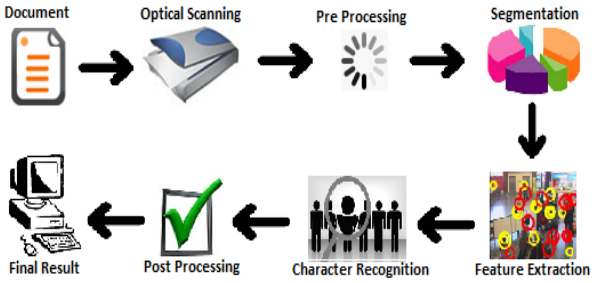


Fig.2: Step By Step Process Of Character Recognition

Fig.2 shows the step by step process of how the individual character is recognized and then the final output is generated. The process includes several stages namely Optical Scanning, Pre Processing, Segmentation, Feature Extraction, Character Recognition and Post Processing.

### a. Optical Scanning

Optical Scanning is the process by which a device called scanner can read the text printed on a paper and translates into the form that a computer can use. It actually works by digitizing an image which divides it into grid of boxes in which a box can have either value '1' or '0'. '1' means the box is filled and '0' means the box is empty. Then the matrix; called as bitmap; is created and based on that it is displayed



ISSN: 2278 – 7798 

on the screen of the computer. It is preferable to save the image in jpg, bmp or tif format.

b. Pre Processing

Next step after scanning the document is Pre Processing. Pre Processing consists of number of basic steps after which the image is suitable to be sent to segmentation. In this stage, skew detection of an image and binarization of an image is done.

b.1 Skew Detection

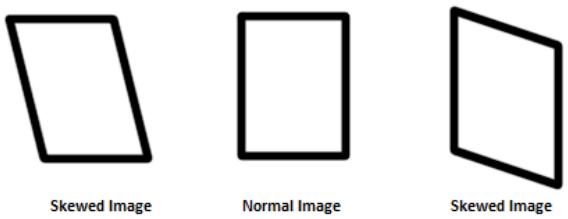


Fig.3: Skew Detection of an image

Skewing of an image means the image is not in the proper direction i.e. the image is either left tilted or right tilted. Fig.3 shows three images in which the two images are skewed images and one is normal image. There are possibilities that the image is skewed after scanning is done and that might be a major concern in case of character recognition. To avoid this, skew detection is carried out in which an image is been checked for the angle of orientation. A simple image rotation is done till the horizontal lines are exactly parallel with the horizontal axis and the vertical lines are exactly parallel with the vertical axis.

b.2 Binarization

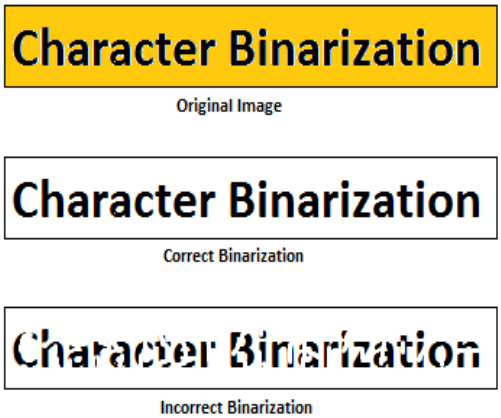


Fig.4: Binarization of an image

The scanned image is primarily in colored form. Binarization of an image means the image is first converted to grayscale image and then it agains converts to the binary image.The binary image will be in the form of black and white image. It is also necessary to have correct binarization of an image as incorrect binarization may cause problems as shown in Fig.4.

One of the best ways to use image binarization is to select a threshold value for the intensity of an image and then classify all the pixels with the values above the threshold value as 'White' and all the pixels with the values below the threshold value as 'Black'.

A(x,y)=1 if A(x,y) is less than threshold value
A(x,y)=0 if A(x,y) is greater than threshold value

Fig.5: Compare pixel with threshold

Fig.5 shows the comparision between the pixel from an image with the threshold value chosen where A(x,y) represents the pixel of an image. We obtain the binary image after applying binarization which consists of two values – 1 (black) and 0 (white).

c. Segmentation

Segmentation is the fundamental stage in character recognition process because the character recognition depends on the segmentation and incorrect segmentation may lead to incorrect character recognition. In segmentation, an image is decomposed into individual character. Segmentation stage includes two parts – Character Lines Segmentation and Individual Character Segmentation.

c.1 Character Lines Segmentation

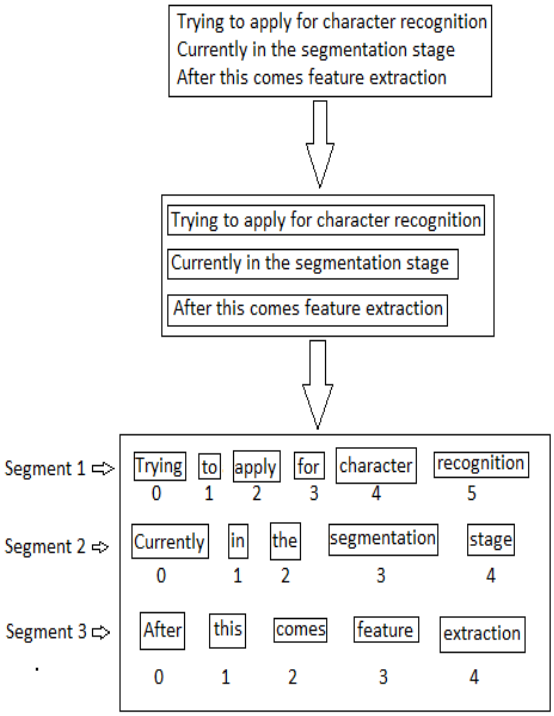


Fig.6: Character Lines Segmentation

In character lines segmentation, the character line is segmented and the words are stored in an array. Thus we achieve extracting various words from a line and then storing it into an array for future process. Fig.6 shows segmenting character lines. The first line is broken down into different words i.e. Trying, to, apply, for, character, recognition. These words are stored in an array.



ISSN: 2278 – 7798 

c.2 Individual Character Segmentation

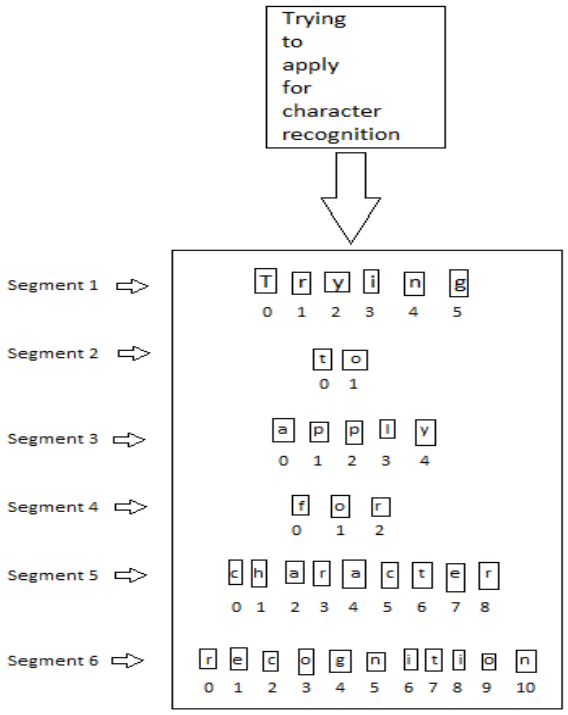


Fig.7: Individual Character Segmentation

Fig.7 shows segmenting individual character and then stored in an array. The word 'Trying' is segmented to characters and then stored in an array. Hence, the segementation is done and further proceed to next stages of character recognition.

d. Feature Extraction

Feature Extraction is the stage that works on a single character image. Here, there are various numerical features that represents the character images such as width of the character, height of the character and pixels in various regions. An individual character image is been converted to two dimentional binary matrix in which all the pixels of the character image are mapped to the matrix. The width and height of character image may vary. Therefore, sampling algorithm is adopted.

Sampling Algorithm:-

1. Find the width of character
   - Map the first (0,y) and the last (width,y) of matrix
   - Map (width/2,y) of the matrix
   - Further divide and map accordingly to the matrix
2. Find the height of character
   - Map the first (x,0) and the last (x,height) of matrix
   - Map (x,height/2) of the matrix
   - Further divide and map accordingly to the matrix
3. Reduce the matrix by sampling both height and width

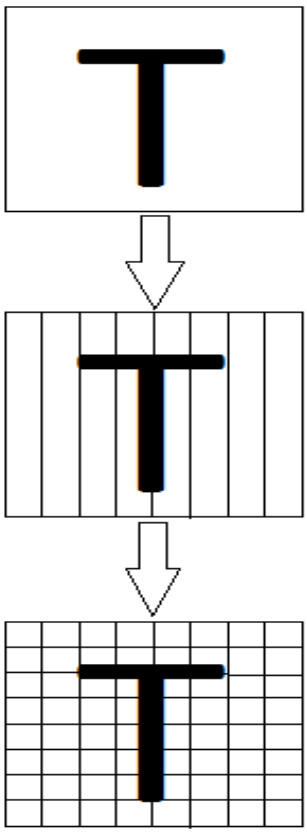

Fig.8: Feature Extraction of Character Image

Fig.8 shows the feature extraction of a character image where the character image is broken down into matrix. It means that the pixels are converted to binary numbers and are stored in the matrix.

The matrix is formed that needs to be implemented with neural network. For this, two-dimensional matrix is converted to one-dimensional matrix.

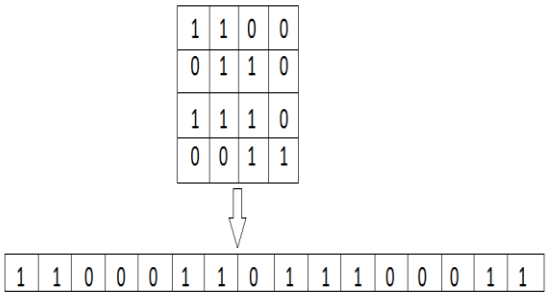


Fig.9: Example of Two Dimensional to One Dimentional

Fig.9 shows the example of two dimentional to one dimentional matrix. This one dimentional matrix will be used as an input for character recognition.



ISSN: 2278 – 7798 

e. Character Recognition

Before recognition, classification is done by using the features extracted from the character image. Multi Layer Perceptron (MLP) is been used for the classification which maps sets of input data onto the sets of outputs.

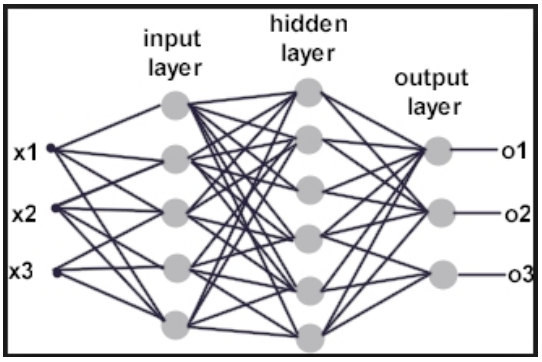


Fig.10: Multi Layer Perceptron

Fig.10 shows the multi layer perceptron where there are three layers – input, hidden, output. The one dimentional arrays which are created in the previous stage will be the input layer of MLP. Each and every layer of MLP is connected to one another. Thus, after feeding the hidden layer; Final output layer is created.

f. Post Processing

Post processing is the last stage of the process in which the characters; which are recognized; are printed in the structured form by calculating the equivalent unicode value using the standard unicode table.

Table.1: Character Unicode Values

| Character | Unicode Value |
|---|---|
| T | 0x0054 |
| R | 0x0072 |
| Y | 0x0079 |
| I | 0x0069 |
| N | 0x006E |
| G | 0x0067 |

Each and every alphabet of the word 'Trying' is been converted to their unicode values as shown in Table.1.

## III. Conclusion

Optical character recognition using neural networks involves various stages that needs to be implemented correctly else the whole process will fall back. The input document is scanned, properly adjusted, segmented according to criteria, features extracted, recognized, obtained the unicode values and thus the output document is created. Our further research will be to make the recognition more faster and considering the other symbols such as geometric symbols.

## IV. References

[1] Mohanad Alata and Mohammad Al-Shabi, "Text Detection and Character Recognition using Fuzzy Image Processing", Journal of Electrical Engineering, vol.57, No.5, 2006, p258-267

[2] Pramod J Simha and Suraj K V, "Unicode Optical Character Recognition and Translation Using Artificial Neural Network", International Conference on Software Technology and Computer Engineering (STACE-2012), 22nd July 2012, Vijayawada, Andhra Pradesh, India

[3] Madhup Shrivastava, Monika Sahu, and Dr. M.A. Rizvi, "Artificial Neural Network Based Character Recognition using Backpropagation" International Journal of Computers & Technology, vol. 3, No. 1, Aug, 2012

[4] Vivek Shrivastava and Navdeep Sharma, "Artificial Neural Network based Optical Character Recognition", Signal & Image Processing: An International Journal (SIPIJ), vol.3, No.5, October, 2012

## V. Author's Profile

**Bhaviya Rajesh Gandani**
B.E. In Computer Engineering (2013), Mumbai

**Ami Rajesh Gandani**
PGDBM & M.Com (2010), Mumbai

**Kushal Dharmesh Doshi**
S.E. In Information Technology, Mumbai


**Ankur Naishadh Gandhi**
B.E. In Computer Engineering (2013), Mumbai



ISSN: 2278 – 7798 