International Journal of Innovations in Engineering and Technology (IJIET)
http://dx.doi.org/10.21172/ijiet.91.04

An Improved scheme of Optical Character Recognition Algorithm

T.Kranthi
Department of Computer Science and Engineering
Anil Neerukonda Institute of Technology and Sciences,
Sangivalasa, Visakahapatnam, India

Jagadish Gurrala
Department of Computer Science and Engineering
Anil Neerukonda Institute of Technology and Sciences,
Sangivalasa, Visakahapatnam, India

G. Santhoshi
Department of Computer Science and Engineering
Anil Neerukonda Institute of Technology and Sciences,
Sangivalasa, Visakahapatnam, India

Abstract- In this paper, the researchers are recognized on hand written character efficiently on computer with input is either a trained image[11] or currently provided scanned image, convert into a text document and also convert text document into trained hand written. Character recognition, usually abbreviated to optical character recognition or shortened OCR, is the mechanical or electronic translation of images of handwritten, typewritten or printed text (usually captured by a scanner) into machine editable text. It is a field of research in pattern recognition, artificial intelligence and machine vision. Though academic research in the field continues, the focus on character recognition has shifted to implementation of proven techniques. Optical character recognition is a scheme which enables a computer to learn, understand, improvise and interpret the written or printed character in their own language. Optical Character Recognition uses the image processing technique to identify any character computer/typewriter printed or hand written. A lot of work has been done in this field. But a continuous improvisation of OCR techniques is being done based on the fact that algorithm must have higher accuracy of recognition, higher persistency in number of times of correct prediction and increased execution time. The objective of this paper is to identify handwritten characters with the use of OCR. The improved scheme of algorithm should be able to extract the characters one by one and map the target output for training purpose.

Keywords – Support vector machine, Maximally Stable Extremal Regions(MSER)

I. INTRODUCTION

The idea of this research is to device efficient algorithms which get input in digital image format. After that it processes the image for better comparison. Later the processed image is compared with already available trained dataset. The last step gives a prediction of the character in percentage accuracy.

After automatic processing of the image, the training dataset has to be used to train "classification engine" for recognition purpose. The program code has to be written in MATLAB. To solve the defined handwritten character recognition problem of classification authors used in MATLAB computation software[10] with computer vision Toolbox and Image Processing Toolbox add-on. The computation code is divided into the next categories:

Pre-processing of the image, Region Detection, Creating a data template, Training, Testing and Recognition.

The rest of this paper is organized as follows. In section 2 discuss literature review on machine learning concept were presented. In section 3 focuses on related work and hand written characters pattern recognition and their schemes and running a scenarios were discussed in section 4. In section 5 the results of the performance evaluation are discussed. Conclusion & Future work is given in section 6.

II. EXISTING ALGORITHM

A. Machine Learning algorithm –

Machine simulation of human functions has been a very challenging research field since the advent of digital computers. In some areas, which entail certain amount of intelligence, such as number crunching or chess playing, tremendous improvements are achieved. On the other hand, humans still outperform even the most powerful computers in the relatively routine functions such as vision. In machine learning, we have supervised and unsupervised learning.

1. Supervised Learning algorithm-

If you are training your machine learning task for every input with corresponding target, it is called supervised learning, which will be able to provide target for any new input after sufficient training. Your learning algorithm seeks a function from inputs to the respective targets. If the targets are expressed in some classes, it is called classification problem. Alternatively, if the target space is continuous, it is called regression problem.

2. Unsupervised Learning algorithm-

If you are training your machine learning task only with a set of inputs, it is called unsupervised learning, which will be able to find the structure or relationships between different inputs. Most important unsupervised learning is clustering, which will create different cluster of inputs and will be able to put any new input in appropriate cluster. However unsupervised learning also encompasses many other techniques that seek to summarize and explain key features of the data. In this Overview, Character Recognition (CR) is an umbrella term, which has been extensively studied in the last half century and progressed to a level, sufficient to produce technology driven applications. Now, the rapidly growing computational power enables the implementation of the present CR methodologies and also creates an increasing demand on many emerging application domains, which require more advanced methodologies.

B. Handwritten Character Recognition:

It is enhanced version of machine learning approaches. Handwriting recognition inherited a number of technologies from Optical Character Recognition (OCR). The main difference between handwritten and typewritten characters is in the variations that come with handwriting. It is also worth noticing that OCR deals with offline recognition while handwriting recognition may be required for both on-line and off-line signals. (On-line means that data is captured as it is written. For offline, all the data is collected before processing starts). On-line handwriting recognition involves the automatic conversion of text as it is written on a special digitizer or PDA, where a sensor picks up the pen-tip movements X(t),Y(t) as well as pen-up/pen-down switching. It involves the automatic conversion of text in an image into letter codes, which are usable within computer and text processing applications. The data obtained by this form is regarded as a static representation of handwriting. The method we chosen are comparatively difficult for Offline handwriting Recognition. Because Recognition of any handwritten characters with respect to any language is difficult, since, the handwritten characters differ not only from person to person but also according to the state of mood of the same person.

The process of handwriting recognition involves extraction of some defined characteristics called features to classify an unknown handwritten character into one of the known classes. A typical handwriting recognition system consists of several steps, namely: preprocessing, segmentation, feature extraction and classification. Several types of decision methods, including statistical methods, neural networks, structural matching (on trees, chains and more) and stochastic processing (Markov chains and more) have been used along with different types of features. Many recent approaches mix several of these techniques together in order to obtain improved reliability, despite wide variation in handwriting. So we proposed a new approach, which solves the problems faced by existing approaches. The main advantage of the proposed system is which overcomes the problem of slant correction.

III. PROPOSED ALGORITHM

in the paper authors are decided to perform improved scheme of optical character recognition in 3 phases.

C. First Phase:

Pre-processing of the sample image involves few steps that are mentioned as follows:

A.1 Grey-scaling of RGB image:

Grey-scaling of an image is a process by which an RGB image is converted into a black and white image.

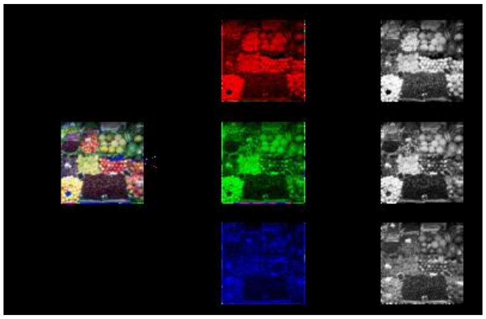


Figure 1. Converting RGB image palatte into Gray Scale images

In this process is important for Binarization as after gray-scaling of the image, only shades of gray remains in the image, binarization of such image is efficient.

A.2 Binarization:

Binarization of an image converts it into an image which only have pure black and pure white pixel values in it. Basically during binarization of a grey-scale image, pixels with intensity lower than half of the full intensity value gets a zero value converting them into black ones. And the remaining pixels get a full intensity value converting it into white pixels.

D. Related Work

Handwritten Character Recognition is a process of transforming handwritten text into machine executable format. There are mainly three steps in pattern recognition:

1. Observation,
2. Pattern Segmentation
3. Pattern Classification.

Recognition of character has become very interesting topic in pattern recognition for the researchers during last few decades. In general, handwritten recognition is classified in to two types as on-line and off-line recognition methods. Offline handwriting recognition involves the automatic conversion of text into an image into letter codes which are usable within computer and text-processing applications. The data obtained by this form is regarded as a static representation of handwriting. But, in the on-line system, the two dimensional coordinates of successive points are represented as a function of time and the order of strokes made by the writer are also available. Offline character recognition is comparatively more challenging due to shape of characters, great variation of character symbol, different handwriting style and document quality. Several applications including mail sorting, bank processing, document reading and postal address recognition require offline handwriting recognition systems. As a result, the off-line handwriting recognition continues to be an active area of research towards exploring the newer techniques that would improve recognition accuracy.

C. Proposed System

We proposed converting handwritten to text and text to handwritten by using OCR algorithm. This can be done in 3 phases

1. Pre-processing
2. Region detection
3. Matching

1. Pre-processing

First we scan the handwritten document and give that image as input to the pre- processing step .In pre-processing some operations are performed on the input image. First convert the RGB image into gray scale and remove noise from that image after that detect the edges using sobel operator and perform dilation and erosion on that image after that detect each character in the form of block and display it. Finally save the images in a folder.

2. Region Detection

Second phase is region detection in this phase the region detection algorithm that is MSER is used. MSER(Maximally Stable Extremal Regions) MSER is a method for blob detection in images. The MSER algorithm extracts from an image a number of co-variant regions, called MSERs: an MSER is a stable connected component of some gray-level sets of the image. MSER is based on the idea of taking regions which stay nearly the same through a wide range of thresholds. All the pixels below a given threshold are white and all those above or equal are black. If we are
Throughput is the ratio of number of packets received to time in seconds shown a sequence of threshold images[5] It with frame ‘t’ corresponding to threshold we would see first a black image, then white spots corresponding to local

intensity minima will appear then grow larger. These white spots will eventually merge[6], until the whole image is white. The set of all connected components in the sequence is the set of all extremal regions. Optionally, elliptical frames are attached to the MSERs by fitting ellipses to the regions. Those regions descriptors are kept as features. The word extremal refers[7] to the property that all pixels inside the MSER have either higher (bright extremal regions) or lower (dark extremal regions) intensity than all the pixels on its outer boundary.

This operation can be performed by first sorting all pixels by gray value and then incrementally adding pixels to each connected component as the threshold is changed. The area is monitored. Regions such that their variation with respect to the threshold is minimal are defined maximally stable: Let's make all the pixels below a threshold white. The others black considering a sequence of thresholded images with increasing thresholds sweeping from black to white we pass from a black image to images where white blobs appear and grow larger by merging, up to the final image. Over a large range of thresholds the local binarization[4-8] is stable and shows some invariance to affine transformation of image intensities and scaling.

MSER Processing : The MSER extraction implements the following steps:

1. Sweep threshold of intensity from black to white, performing a simple luminance thresholding of the image.

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[\,n\sum x^2 - (\sum x)^2\,][\,n\sum y^2 - (\sum y)^2\,]}}$$

2. Extract connected components ("Extremal Regions")

3.Find a threshold when an extremal region is "Maximally Stable", i.e. local minimum of the relative growth of its square. Due to the discrete nature of the image, the region below / above may be coincident with the actual region, in which case the region is still deemed maximal.

4.Approximate a region with an ellipse (this step is optional) Keep those regions descriptors as features.

However, even if an extremal region[10] is maximally stable, it might be rejected if:

a) It is too big (there is a parameter MaxArea).
b) It is too small (there is a parameter MinArea).
c) It is too unstable (there is a parameter MaxVariation).
d) It is too similar to its parent MSER .

Margin = the number of thresholds for which the region is stable

3. Matching

Finally matching is done based on correlation. Correlation and regression analysis are related in the sense that both deal with relationships among variables. The correlation coefficient is a measure of linear association between two variables. Values of the correlation coefficient are always between −1 and +1. A correlation coefficient of +1 indicates that two variables are perfectly related in a positive linear sense, a correlation coefficient of −1 indicates that two variables are perfectly related in a negative linear sense, and a correlation coefficient of 0 indicates that there is no linear relationship[8][9] between the two variables. For simple linear regression, the sample correlation coefficient is the square root of the coefficient of determination, with the sign of the correlation coefficient being the same as the sign of b1, the coefficient of x1 in the estimated regression equation.

Correlation formula:

Example for correlation:

Table 1.The following numerical table shows about how this formula is being used:

X	Y
1	2
3	5
4	5
4	8

$\sum X$ = 1+3+4+4=12

$\sum X\,Y$=(1)(2)+(3)(5)+(4)(5)+(4)(8)=69

$\sum Y=2+5+5+8= 20$

$\sum X2 = 12+32+42+42=42$

$\sum Y2=22+52+52+82=118$

Now Substitute these values in correlation formula

r=(69-(20)(12)/4)/(√ (〖(42〗 ^ - 〖(12)〗 ^2/4)) √ (〖118〗 ^ -(20)^2/4))=.866

Proposed Algorithm:

This is processing steps to convert the handwritten details into text form in the following way:

1. Take the handwritten document as a input and scan it.
2. Pre-processing: convert RGB image into gray scale. Remove noise from the image and convert into binary. Store the output in a file.
3. Give that file as input to the feature extraction.
4. After that give labels to each feature.
5. Finally matching is done.This is processing steps to Converting text details in to handwritten form:
6. Take input as text.
7. Replace that text with the hand writing of the trained data.

D. METHODOLOGIES

In this paper, we implemented three modules to convert handwritten details into text.

1.Pre-processing
2.Feature extraction
3. Classification

PRE-PROCESSING

1.Scan the handwritten document
2.Convert that RGB image into grayscale
3.Convert grayscale into binary-im2bw()
4.Detect edges using- edge()
5.Segmentation is done using –regionprops()

E. PROPOSED SYSTEM

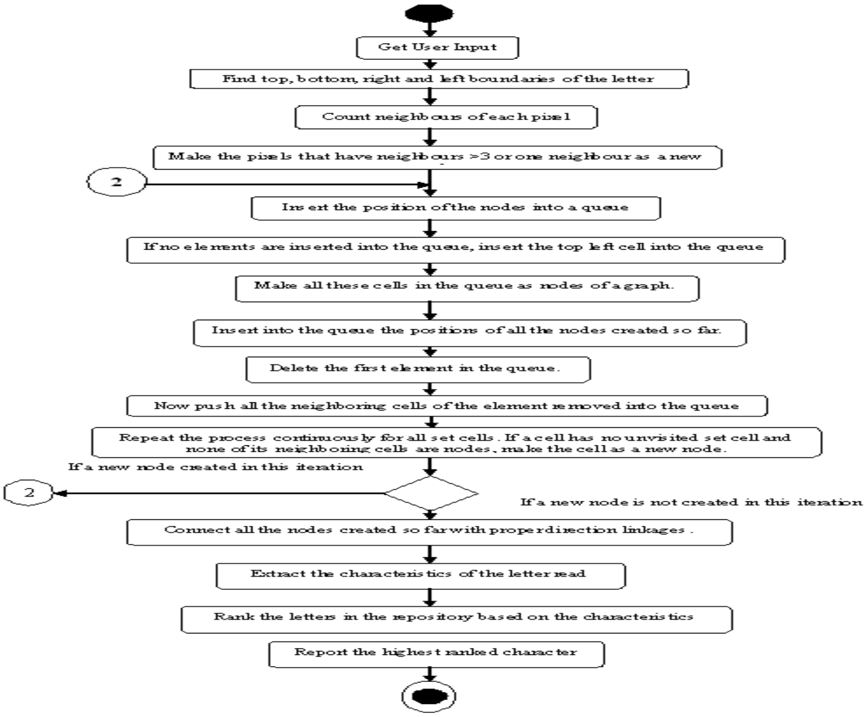

Get User Input
Find top, bottom, right and left boundaries of the letter
Count neighbours of each pixel
Make the pixels that have neighbours >3 or one neighbour as a new
2
Insert the position of the nodes into a queue
If no elements are inserted into the queue, insert the top left cell into the queue
Make all these cells in the queue as nodes of a graph.
Insert into the queue the positions of all the nodes created so far.
Delete the first element in the queue.
Now push all the neighboring cells of the element removed into the queue
Repeat the process continuously for all set cells. If a cell has no unvisited set cell and none of its neighboring cells are nodes, make the cell as a new node.
If a new node created in this iteration
2
If a new node is not created in this iteration
Connect all the nodes created so far with proper direction linkages .
Extract the characteristics of the letter read
Rank the letters in the repository based on the characteristics
Report the highest ranked character

Figure 2. Detailed Activity Diagram for proposed algorithm

F. Detailed Analysis:

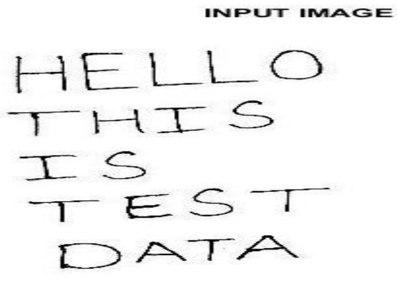
In the above algorithm decribed in fig1, the system accepts user handwritten details and finding the top, bottom, left and right of handwritten image. Activity diagram is to describe dynamic aspects of the system. Activity diagram is basically a flow chart to represent the flow from one activity to another activity. The activity can be described as an operation of the system. So the control flow is drawn from one operation to another.

G. *Experimental Details:*

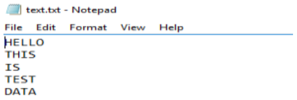
Sample Input:

(a)

Sample Output:



(b)



```
HELLO
THIS
IS
TEST
DATA
```

Figure 3. (a) Hand Written input image (b) Text output

H.Results and Discussions

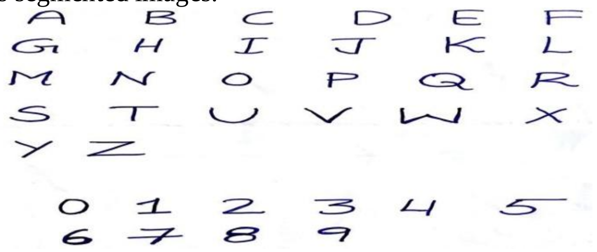
First Scenario- Test Case 1:

Testing and Results:

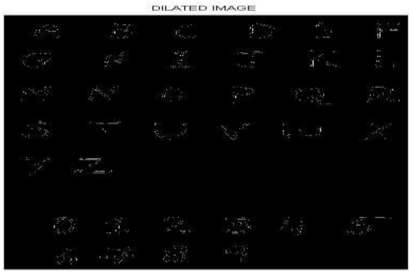
Testing itself may be defined at various levels of SDLC[3]. The testing process runs parallel to software development. Before jumping on the next stage, a stage is tested, validated and verified. Testing separately is done just to make sure that there are no hidden bugs or issues left in the software. Software is tested on various levels like Unit testing, Integrated testing, etc.

Pre-Processing Results:

In pre –processing we take the scanned image and perform noise removal ,edge detection and morphological operations. The output we get is segmented images.



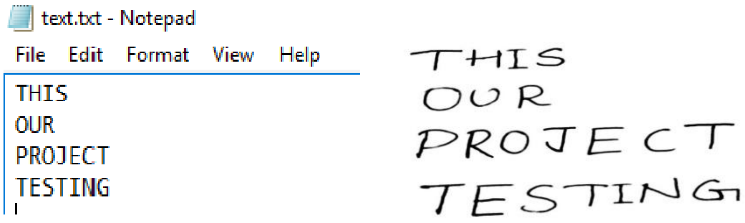
(a) Text input



(b)

Figure 4. (a) Text image (b) Hand written character approximate manner

Second Scenario- Test Case 2:



```
THIS
OUR
PROJECT
TESTING
```

(a) (b)

Figure 5. (a) Given Text Document (b) Generated Hand Written Character from text.

IV.CONCLUSION

In this paper, we proposed an OCR system for handwritten recognition .We have taken only 10 datasets and trained. Classification is done based on maximum correlation result for each and individual character. The accuracy of the system can be increased by collecting more datasets. In the future work, the MS concept is alleviated to improve version of OCR using enhanced techniques.

REFERENCES

[1] Jayashree R Prashad, Dr. U V kulkarni "Trends in handwriting recognition" IEEE, International Journal of Engineering and Innovative Technology (IJEIT) Volume 2, Issue 7, January 2013.

[2] Ratnashil N Khobragade, Dr. Nitin A. Koli and Mahendra S Makesar, A Survey on Recognition of Devnagari Script, International Journal of Computer Applications & Information Technology, Vol. 2, Issue 1, 2013.

[3] G.Santoshi, et al., "Novel approach of DNA sequencing algorithm to image security" published in IEEE conference , SCOPES DOI: 10.1109/SCOPES.2016.7955498, 3-5 Oct. 2016,Page(s):1501- 1505

[4] Chih-Chung Chang and Chih-Jen Lin, LIBSVM: a library for support vector machines, 2001. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm

[5] Mahesh Jangid, Kartar Singh, Renu Dhir, Rajneesh Rani, "Performance Comparison of Devanagari Handwritten Numerals Recognition", Internation Journal of Computer Applications (IJCA), Vol. 22, No.1, May 2011.

[6] Kartar Singh Siddharth, Mahesh Jangid, Renu Dhir, Rajneesh Rani, "Handwritten Gurmukhi Character Recognition Using Statistical and Background Directional Distribution Features", International Journal of Computer Science and Engineering (IJCSE), Vol. 3, No. 6, June 2011.

[7] Chih-Chung Chang and Chih-Jen Lin, LIBSVM: a library for support vector machines, 2001. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[8] Chih-Wei Hsu, Chih-Chung Chang, and Chih-Jen Lin, "A Practical Guide to Support Vector Classification", [Online]. Available:

[9] http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf

[10] http://www.isical.ac.in/~ujjwal/download/database.html. Gonzalez, Woods and Eddins,"a book on Digital Image Processing Using MATLAB.

[11] P. Sanyasi Naidu and Jagadish Gurrala.," "A Study of Low Power Consumable Frameworks for Hiding Audio signals in Image Container Using Different Steganography techniques Using ARM Controller Devices For Smart Living", No.9 (2016) Issue No. :32 (2016),pages : 261-269 Serial Publications