© 2018 IJSRSET | Volume 4 | Issue 2 | Print ISSN: 2395-1990 | Online ISSN : 2394-4099
National Conference on Advanced Research Trends in Information and
Technologies (NCARTICT-2018), Department of IT, L. D. College of Engineering,
Ahmedabad, Gujarat, India In association with
International Journal of Scientific Research in Science, Engineering and Technology

Design and Implementation of OCR to identify English Characters and Numbers

Rachit Adhvaryu*, Rachana Parikh, Komil Vora
Information Technology, V.V.P. Engineering College, Rajkot, Gujarat, India

ABSTRACT

Optical character recognition has become one of the most successful applications of technology in the field of pattern recognition and artificial intelligence. At present, there is great demand of such softwares which can identify the characters from scanned documents or images. Optical Character Recognition deals with the problem of recognizing optically processed characters. Optical recognition is performed off-line after the writing or printing has been completed, as opposed to on-line recognition where the computer recognizes the characters as they are drawn. This paper presents the system which identifies characters from the images. The objective of this system prototype are to develop a prototype for the Optical Character Recognition (OCR) system and to implement the Template Matching algorithm in developing the system prototype.

Keywords: OCR, Template Matching, Gray Scale Images, Color Images, Fixed Size Templates

I. INTRODUCTION

Optical character recognition (OCR) is the process of classification of optical patterns contained in a digital image corresponding to alphanumeric or other characters [1]. OCR has attained much popularity in the both academics as well as in industry. Over the last few years, machine reading has grown rapidly through the development of much more sophisticated and easy OCR systems. OCR Technology allows us to convert scanned documents, pdf files and images from digital camera to editable and readable form [1]. Optical character recognition belongs to the family of techniques performing automatic identification [2]. OCR has become most prominent and successful technological applications in the field of pattern recognition and artificial intelligence [2]. Below we discuss these different techniques and define OCR's position among them.

Automatic Identification

The traditional way of entering data into a computer is through the keyboard. However, this is not always the best nor the most efficient solution. In many cases automatic identification may be an alternative. Various technologies for automatic identification exist, and they cover needs for different areas of application. Below a brief overview of the different technologies and their applications is given [1].

a. Speech recognition

In systems for speech recognition, spoken inputs from predefined library of words are recognized. Such systems should be speaker-independent and may be used for instance for reservations or ordering of goods by telephone. Another kind of such systems are those used to recognize the speaker, rather than the words, for identification [1].

b. Radio frequency

This kind of identification is used for instance in connection with toll roads for identification of cars. Special equipment on the car emits the information. The identification is efficient, but special equipment is needed both to send and to read the information [2]. The information is also inaccessible to humans.

c. Vision systems

By the use of a TV-camera objects may be identified by their shape or size. This approach may for instance be used in automatons for recirculation of bottles [2]. The type of bottle must be recognized, as the amount reimbursed for a bottle depends on its type.

d. Magnetic stripe

Information contained in magnetic stripes is widely used on credit cards etc. Quite a large amount of information can be stored on the magnetic stripe, but specially designed readers are required and the information cannot be read by humans [2].

e. Bar code

The bar code consists of several dark and light lines representing a binary code for an eleven digit number, ten of which identify the particular product. The bar code is read optically, when the product moves over a glass window, by a focused laser beam of weak intensity which is swept across the glass window in a specially designed scanning pattern. [2] The reflected light is measured and analysed by a computer.

f. Magnetic ink

Printing in magnetic ink is mainly used within bank applications. The characters are written in ink that contains finely ground magnetic material and they are written in stylized fonts which are specifically designed for the application. Before the characters are read, the ink is exposed to a magnetic field. This process accentuates each character and helps Simplify the detection. The characters are read by interpreting the waveform obtained when scanning the characters horizontally [2]. Each character is designed to have its own specific waveform. Although designed for machine reading, the characters are still readable to humans. However, the reading is dependent on the characters being printed with magnetic ink.

g. Optical Mark Reading

This technology is used to register location of marks. It may be used to read forms where the information is given by marking predefined alternatives. Such forms will also be readable to humans and this approach may be efficient when the input is constrained and may be predefined and there is a fixed number of alternatives [2].

h. OCR

Optical character recognition is needed when the information should be readable both to humans and to a machine and alternative inputs can’t be predefined. In comparison with the other techniques for automatic identification, optical character recognition is unique in that it does not require control of the process that produces the information [1].

II. OPTICAL CHARACTER RECOGNITION

Optical Character Recognition deals with the problem of recognizing optically processed characters. Optical recognition is performed off-line after the writing or printing has been completed, as opposed to on-line recognition where the computer recognizes the characters as they are drawn. Both hand printed and printed characters may be recognized, but the performance is directly dependent upon the quality of the input documents [2, 3].

The more constrained the input is, the better will the performance of the OCR system be. However, when it comes to totally unconstrained handwriting, OCR machines are still a long way from reading as well as humans. However, the computer reads fast and technical advances are continually bringing the technology closer to its ideal.

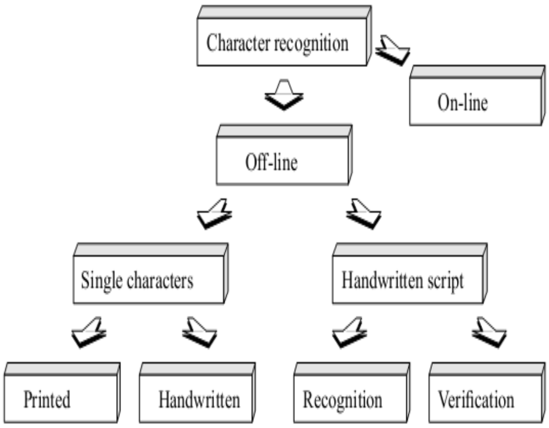


Figure 1. Types of character recognition

A. Online Character Recognition

Today, many tools are available over internet which identifies the given characters in the fraction of time [3]. These tools are mainly used in Multiple Choice Questions examinations conducted using OMR Sheets.

B. Offline Character Recognition

Offline character recognition deals with developing the systems as per the requirement of the user or as per the requirement of desired output. This type of character recognition can be done on the following:

i. Single Character

In this, the system is created in such a way that it identifies one single character at one time [3]. The single characters can be in 2 ways:

a) Printed Characters

In this, the system tries to recognize the printed characters either in the form of text file or an image.

b) Handwritten Characters

In this, the system tries to recognize the handwritten characters which is in the form of scanned document and image.

ii. Handwritten Scripts

The system developed for handwritten scripts are mainly used for 2 purposes:

a) Recognition

This type of system mainly recognizes a particular set of characters from the given handwritten scripts. These systems split the recognized characters in a single character or the group of characters depending on the requirement of system [3, 4].

b) Verification

Once the characters are recognized, the correctness of the characters is required to be validated. The verification deals with the correctness of shapes, diagonals and curves of characters [3]. Also using various Natural Language Processing techniques, these characters once verified can classified in many different classes.

III. METHODS OF OCR

There are 2 methods of OCR:

A. Matrix Matching

The Matrix Matching is technique in which the library of characters is created. The system then compares the scanned characters with the library character matrices. This system works best when the characters to be scanned and the library characters have very little or no variation in style [4].

B. Feature Extraction

Feature Extraction generally deals with the features of characters like shape, closed areas, diagonal lines, line interaction and curves. It is more effective and flexible methods as it has a wide scope to identify the same character with different shapes and dimensions [4].

In this paper, we are focus on Matrix Matching (Template Matching) method.

IV. COMPONENTS OF OCR

A typical OCR system consists of several components. In figure 2 a common setup is illustrated. The first step in the process is to digitize the analog document using an optical scanner. When the regions containing text are located, each symbol is extracted through a segmentation process [4, 5]. The extracted symbols may then be pre-processed, eliminating noise, to facilitate the extraction of features in the next step.

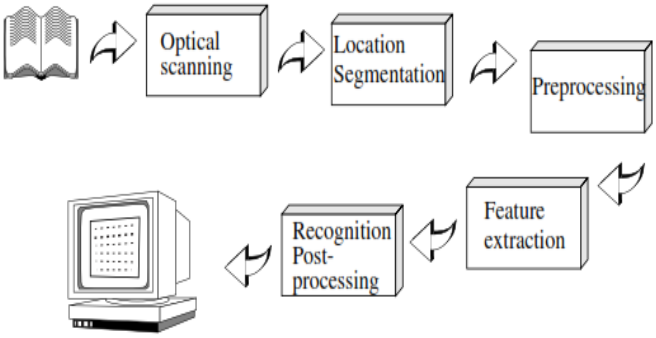


Fig.2 Components of OCR

A. Optical Scanning

It converts multilevel colored image to bi-level image. It uses “Thresholding” method. The quality of recognition depends on bi-level image. Thresholding can be fixed or Variable depends on contrast and brightness of bi-level image [5].

B. Location and Segmentation

It deals with isolation of characters or words. It segments the lines into words or characters. A problem occurs if the characters are connected or fragmented and consists of several parts. It is hard to distinguish noise from text [5, 7].

C. Pre-processing

Image resulting from scanning may contain noise. Pre-processing smoothens the image of characters. It uses Filling and Thinning and also uses Normalization [5, 8].

D. Feature Extraction

It captures essential characteristics of characters. Extraction of features divided into three groups [5, 6]:

i. Distribution of Points
ii. Based on statistical distribution of points.
iii. Includes Zoning, Moments, Crossing and Distance

E. Recognition Post Processing

This phase reduces the dimensionality of feature vector. It extracts feature by deformation like rotation and translation. Transformation can be Fourier, Walsh, Haar, Hough or Karhunen-Loeve and based on the curve describing the contour of the characters. It also deals with Error detection and correction [6].

The identity of each symbol is found by comparing the extracted features with descriptions of the symbol classes obtained through a previous learning phase. Finally contextual information is used to reconstruct the words and numbers of the original text. In the next sections these steps and some of the methods involved are described in more detail.

V. IMPLEMENTATION

A. Implementation Steps

The template-matching algorithm implements the following steps [10]:

a. Firstly, the character image from the detected string is selected.
b. After that, the image to the size of the first template is rescaled.
c. After rescale the image to the size of the first template (original) image, the matching metric is computed.
d. Then the highest match found is stored. If the image is not match repeat again the third step.
e. The index of the best match is stored as the recognized character.

B. Flow control of Template Matching

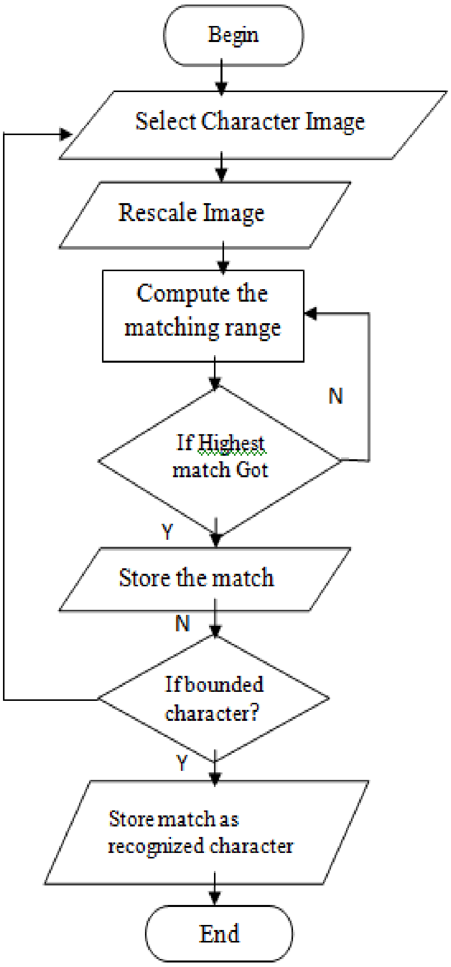


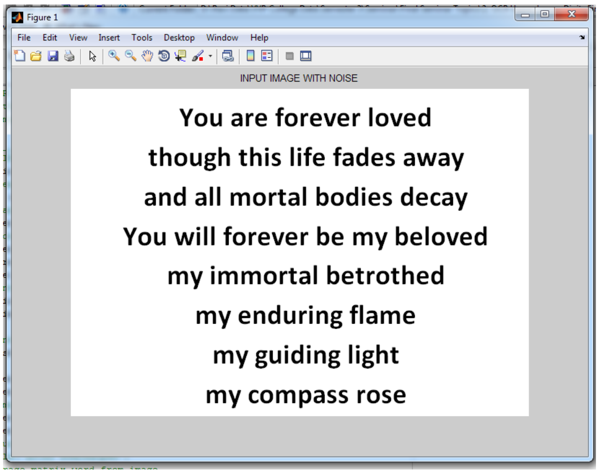
Figure 4.2.1. Shows the Flowchart.

VI. RESULTS

The system was developed using Template Matching method of OCR. 2 Gray Scale and 1 Color Images were taken as an input [9] and the following results were found:

A. Input 1:

Gray Scale Image with medium sized characters were given input to the system.



INPUT IMAGE WITH NOISE

You are forever loved
though this life fades away
and all mortal bodies decay
You will forever be my beloved
my immortal betrothed
my enduring flame
my guiding light
my compass rose

Output 1:

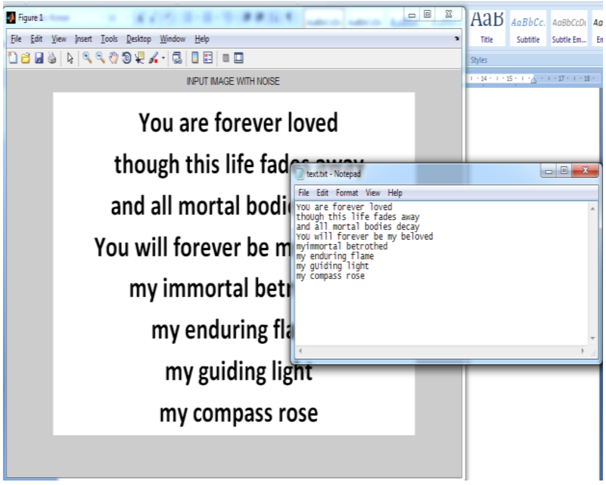


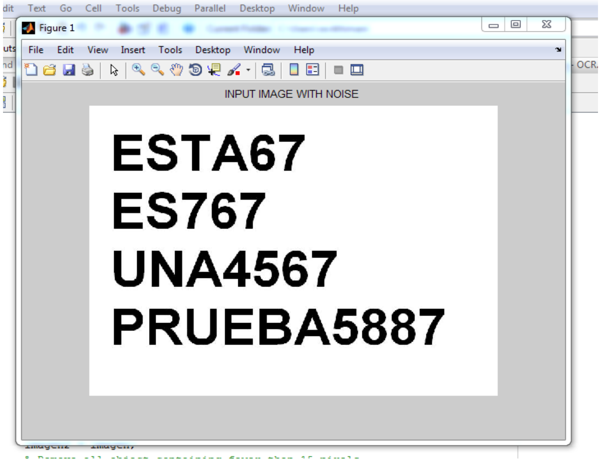
Figure 1
File Edit View Insert Tools Desktop Window Help
INPUT IMAGE WITH NOISE
text.txt - Notepad
File Edit Format View Help

```
YoU are forever loved
thoUgh this life fades away
and all mortal bodies decay
YoU will forever be my beloved
myimmortal betrothed
my endUring flame
my gUiding light
my compass rose
```

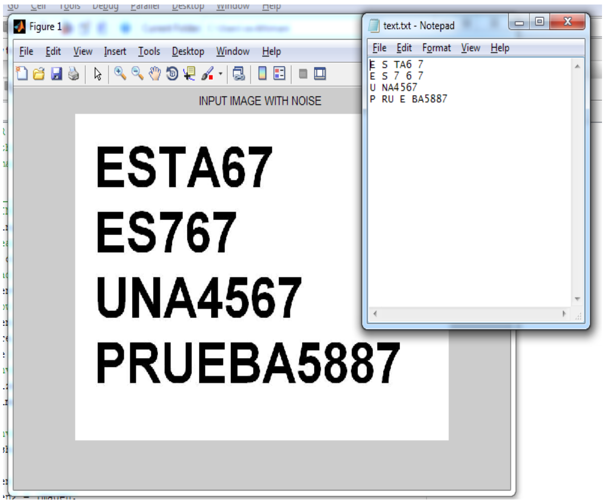
All the texts are clearly visible and are as per the size of library characters, so each character of whole string is recognized properly and generates output without any error.

B. Input 2:

Gray Scale Image with large sized characters was given as an input:



Output 2:

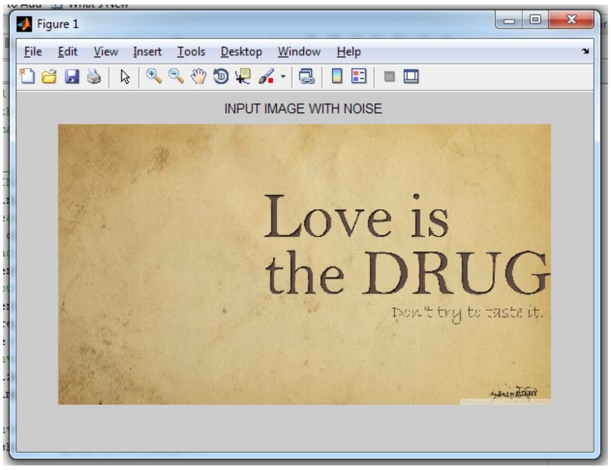


```
E S TA6 7
E S 7 6 7
U NA4567
P RU E BA5887
```

OCR recognizes all the character and digits but the additional spaces are padded in between in resultant text because the library characters are smaller in size as compared to the given input characters.

C. Input 3:

A colored image with some text was given as an input



Output 3:

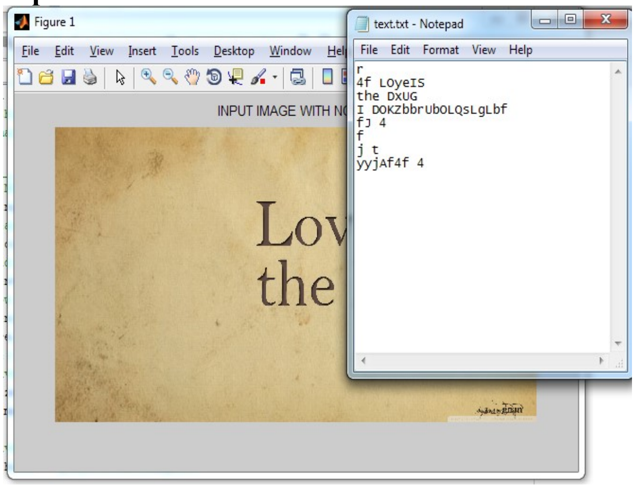


Figure 1
File Edit View Insert Tools Desktop Window Help
INPUT IMAGE WITH NO
Lov
the
text.txt - Notepad
File Edit Format View Help

```
r
4f LOyeIS
the DxUG
I DOKZbbrUbOLQsLgLbf
fJ 4
f
j t
yyjAf4f 4
```

It fails to identify the characters from the image, and generates in-correct output as the system is developed for Gray Scale Image only.

VII. CONCLUSIONS & FUTURE WORK

As an overall view of the system prototype, it could be conclude that this system prototype has been developed by using the technique that has mentioned and elaborated which is the Template Matching approach to recognize the character image. Besides, the interface of the system prototype looks user-friendly and makes the user of this system prototype easier to use it. As a result, the recognition process of this system become smoothly because of the steps that used in this system while recognizing the character. Even though this system prototype could gives several advantages to the users, but this system prototype are still facing a number of limitations like:

1. The system prototype has some limitation related to performance.
2. It works only with stored templates of alphabets and numbers with fixed sized templates.
3. It works only for gray scale image only.

In future, an effective system can be developed which works on Feature Extraction method and also colored image. This new system may be a motivation for enormous and fruitful researches in future.

VIII. REFERENCES

[1]. Maninder Kaur, "A Brief Review on Optical Character Recognition Techniques", International Journal of Computer Science and Mobile Computing, Vol.6 Issue.2, February-2017, pg. 95-100.

[2]. A. Chaudhuri, "Optical Character Recognition Systems for Different

[3]. Languages with Soft Computing", Studies in Fuzziness and Soft Computing 352, Springer International Publishing AG 2017.

[4]. Chowdhury Md Mizan, Tridib Chakraborty and Suparna Karmakar, "Text Recognition using Image Processing", International Journal of Advanced Research in Computer Science, ISSN No. 0976-5697, Volume 8, No. 5, May-June 2017.

[5]. Ashima Singh and Swapnil Desai, "Optical character recognition using template matching and back propagation algorithm", 2016 International Conference on Inventive Computation Technologies (ICICT) Vol-3.

[6]. M. Shen, "Improving OCR Performance with Background Image Elimination", 2015 12th Int. Conf. Fuzzy Syst. Knowl. Discov., pp. 1566-1570, 2015

[7]. E. N. Bhatia, "Optical Character Recognition Techniques : A Review", IJARCSSE, vol. 4, no. 5, pp. 1219-1223, 2014

[8]. Kartar Singh Siddharth, Mahesh Jangid, Renu Dhir, Rajneesh Rani (2011), "Handwritten Gurmukhi Character Recognition Using Statistical and Background Directional Distribution Features", International Journal on Computer Sceince and Engineering, Vol. 3(6), Pages 2332-2345.

[9]. Kirill Safronov, Dr.-Ing. Igor Tchouchenkov, Dr.-Ing Heniz Worn (2007), "Optical Character Recognition Using Optimisation Algorithm", Workshop on Computer Science and Information Technology, Pages 1-5.

[10]. Aparna Vara Lakshmi Vemuri, T.V.Sai Krishna, Atul Negi, "Dataset Generation for OCR" [OCR_Datasheet_Generation.pdf]

[11]. Jesse Hansen, "A Matlab Project in Optical Character Recognition (OCR)".