OCR Accuracy Improvement Techique

Jyoti Goyal, Mrs.Shailja
Student, CSE, CDLU, Sirsa, India
Asst Professor, CSE, CDLU, Sirsa, India

Abstract: This paper is about Performance rate of OCR .OCR is not 100% accurate. This paper gives the best file format suggestion on which human recognition error rate can be reduced. With help of OCR, different CAPTCHA is studied, in different file formats as GIF, BMP, TIFF, PHP.This paper proposed the best file format for CAPTCHA which give least error rate in human recognition and OCR Performance.Tessnet algorithm is used for CAPTCHA generation. Various types of CAPTCHA is used for check error rate.OCR is used for recognition of CAPTCHA. This paper is helpful in reduce the weakness of OCR up to a limit.

Keywords: Captcha, OCR, GIF, BMP, Tesseract OCR engine

I. INTRODUCTION

The basic challenge in designing these CAPTCHA's is to make them easy enough that users are not dissuaded from attempting a solution, yet still too difficult to solve using available computer vision algorithms. As Modern technology grows this gap however becomes thinner and thinner. It is possible to enhance the security of an existing text CAPTCHA by systematically adding noise and other distortions, or arranging characters more tightly. These measures, however, would also make the characters harder for humans to recognize, resulting in a higher error rates and higher Network load. With advances of segmentation and Optical Character Recognition (OCR) technologies, the capability gap between humans and bots in recognizing distorted and connected characters becomes increasingly smaller. This trend would likely render text CAPTCHA's eventually ineffective.

This paper is about Modern OCR Technologies and various types CAPTCHA performance with OCR.

OCR have weakness that all text CAPTCHA is not properly read. In this paper different image file formats of CAPTCHA is used as:-

1. BMP
2. PHP
3. TIFF
4. GIF
5. JPEG

In this paper BMP images files are studied, in which error rate is observed approximately 37%,similar in PHP error rate is higher. But in case of GIF, the error rate is reduced approximately 10%.

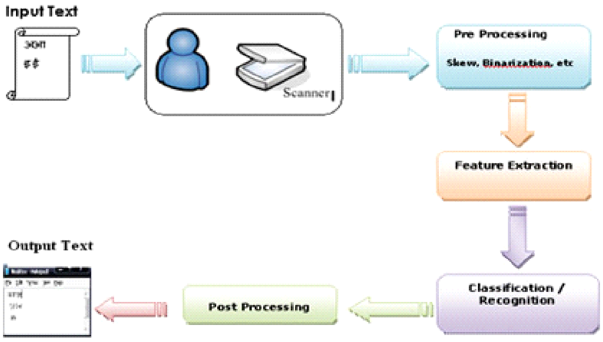


Fig 1 Graphical Representation Of image Recognition Process

II. LITERATURE REVIEW

In today's Research scenarios, there are many techniques, which have been discussed for security of CAPTCHA. But due to enhance in security, error rate is increased in human recognition. In last years, so much research was done.

The OCR Weakness is given by every researcher, when we studied Captcha with OCR recognition process, the CAPTCHA is generated and read through OCR 100% accuracy is not achieved.

Early versions were programmed with
Images of each character, and works for one font at a time.
.
. Some systems are capable of reproducing formatted page including images, columns and other non textual Components.
. In Optical Character Recognition process,
the image is converted in ASCII code. This code is matched with already stored bits, If its matched then no error is zero given, otherwise according to number of different bits, Error rate is given. To enhance the performance
of optical Character recognition this research is necessary. To give an idea, that by which way we can reduce Weakness of OCR.

III. OBJECTIVES

In the research scenario, the different image file format recognition will be implemented. This implementation will very helpful for reduce the weakness of OCR. The Working is described as:-

1. The different Captcha is generated by Tessnet algorithm.
2. With JOCR or GOCR library file, Captcha is read
3. OCR read image text is also given through Note-Pad.
4. Finally Error rate is given according to used formula:
If s is "test" and t is "test", then Error(s,t) = 0, because no transformations are needed. The strings are Equal hence no Error. If s is "test" and t is "tent", then Error(s,t) = 1, because one substitution (change "s" to "n") is sufficient to transforms into t.
Finally Result is given, GIF image format give less error rate than another image file formats.

IV.PROPOSED METHODOLOGY

CAPTCHA is used for security purpose, whereas OCR is used for read these CAPTCHA. If OCR have perfect Accuracy, then several image recognition system increase the performance quality. In this research, Tesseract engine is used with some modification in software, so that every images file format can be read. The CAPTCHA TESTER works as:-

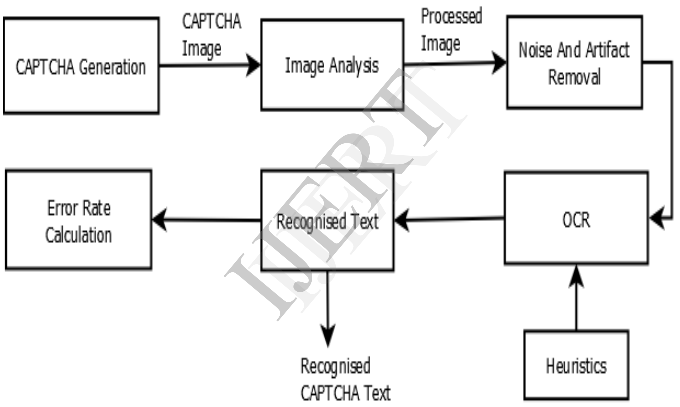
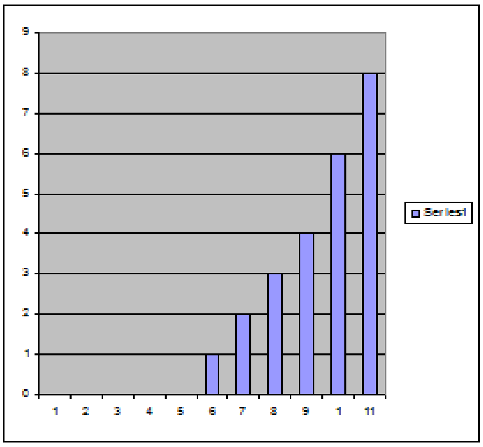
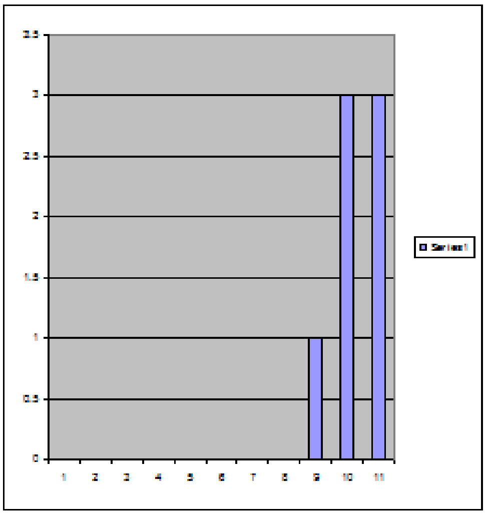


Fig2 Captcha Tester Diagram

Two image file format example is given here:-



Graph 1:-In case of BMP images upto 8 errors.



Graph 2:-In case of GIF images up to 3 errors.

V. APPLICATION

1. To reduce the human error rate in recognition Process of Optical Character Recognition.

2. To move forward to reduce the weakness of OCR.

3. The most important application is that Performance Time is reduced due to less re-generation Captcha.

VI. CONCLUSION AND FUTURE WORK

In this paper, we have been proposed the Method through which OCR Weakness can be reduced. In Future another method for more enhance the accuracy can be done by adding sorting methods in OCR library file

.

REFERNCES

[1] S.V. Rice, F.R. Jenkins, T.A. Nartker, The Fourth Annual Test of OCR Accuracy, Technical Report 95-03, Information Science Research Institute, University of Nevada, Las Vegas, July 1995.

[2] R.W. Smith, The Extraction and Recognition of Text from Multimedia Document Images, PhD Thesis, University of Bristol, November 1987.

[3] R. Smith, “A Simple and Efficient Skew Detection Algorithm via Text Row Accumulation”, Proc. of the 3rd Int. Conf. on Document Analysis and Recognition (Vol. 2), IEEE 1995, pp. 1145-1148.

[4] P.J. Rousseeuw, A.M. Leroy, Robust Regression and Outlier Detection, Wiley-IEEE, 2003.

[5] S.V. Rice, G. Nagy, T.A. Nartker, Optical Character Recognition: An Illustrated Guide to the Frontier, Kluwer Academic Publishers, USA 1999, pp. 57-60.