International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor:6.887
Volume 5 Issue IX, September 2017- Available at www.ijraset.com

A Study to Recognize Printed Gujarati Characters Using Tesseract OCR

Milind Kumar Audichya[1], Jatinderkumar R. Saini[2]
[1, 2] *Computer Science, Gujarat Technological University*

Abstract*: Optical Character Recognition (OCR) is a widely-known technique to recognize the printed text using computer with the help of various peripheral devices. Research works for OCR of many languages scripts is in process and many languages are still far away. Gujarati script is one of the least focused script in research area of OCR as compared to other scripts. A well-known Open Source OCR Engine called Tesseract which is already used for the recognition of numerous scripts, can be used to recognize printed Gujarati characters from digital images. This paper is trying to enlighten the use of Tesseract to recognize Gujarati characters with the help of already available trained data for Gujarati Script.

Keywords*: OCR, Optical Character Recognition, Gujarati OCR, printed Gujarati characters OCR, Tesseract OCR

I. INTRODUCTION

Gujarati is one of the regional languages of India which has a huge number of speakers making it to 26th most spoken language worldwide [1]. Thus, the recognition of Gujarati characters is necessary which is least focused and an attempt is made to recognize Gujarati characters from the digitally printed straight normal images using Tesseract - an Optical Character Recognition (OCR) engine.Printed Gujarati characters OCR is still a challenging task as it depends upon different parameters like: font type, font style, font size and quality of images taken or captured. Still there are several things remains to be managed some of them are disconnected characters, conjuncts and diacritics.Tesseract is Open Source and it supports many languages to recognize characters using its own trained data. It can be used to recognize Gujarati characters from the digitally printed images. Tesseract was initially developed in between the year 1985 and 1994 at Hewlett-Packard (HP) Laboratories Bristol and at HP Co, Greeley Colorado. In 2005, HP has converted Tesseract into Open Source and since 2006, Google is developing it till date [6].

Tesseract is based on 8-bit Unicode Transformation Format (UTF-8) and it can recognize 100+ languages [6]. It also has the potential to recognize different languages if another language's customized trained data is provided. This paper presents a study of "Recognition of Gujarati characters using Tesseract OCR"

Fig. 1. Gujarati Numerals

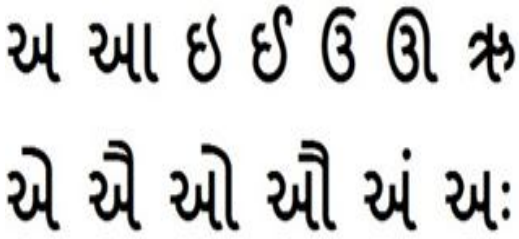


Fig. 2 Gujarati Vowels

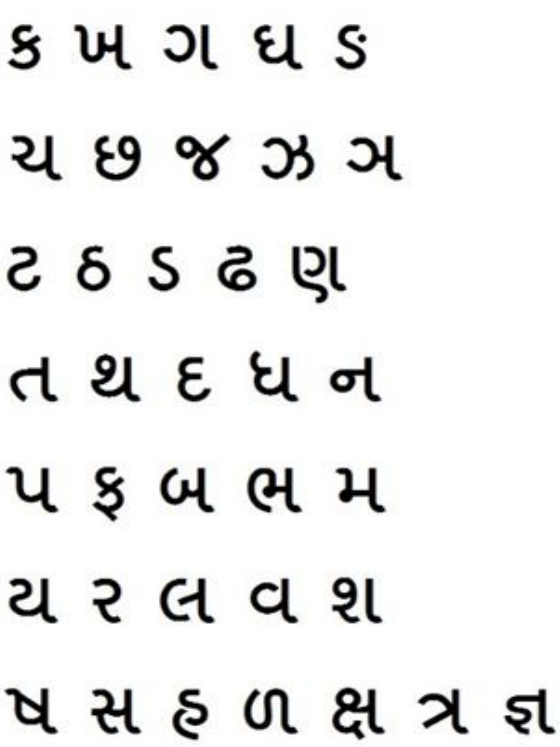


Fig. 3. Gujarati Consonants

There are 10 numerals, 13 vowels and 37 consonants in Gujarati language which are shown in Fig. 1,2 and 3. Trained data to recognize this all numerals, vowels and consonants is available along with Tesseract and thus, it can recognize the Gujarati language characters from digitally printed images.

II. LITERATURE REVIEW

Gujarati characters OCR had been a challenging and tedious task due to the complexity of script either we are working on digitally typed printed Gujarati characters or handwritten Gujarati characters.

A systematic approach using hybrid method based upon binary tree classifier and k-Nearest Neighbor (kNN) for Gujarati handwritten character had been introduced by Chhaya Patel and Apurva Desai and obtained overall 63.1% set wise average accuracy [8].

Due to complexity of script researchers limits their scope to either certain digits, characters, or characters sets. Shailesh Chaudhari and Ravi Gulati tried for the Separation as well as Identification of Mixed bilingual (English-Gujarati) Digits using kNN Classifier ad gained 99.20%, 99.26% average accuracy for English and Gujarati digits respectively [9].

Jignesh Dholakia, Atul Negi and S. Rama Mohan has presented an effective method for zone detection in digitally printed Gujarati characters [10].

Swital J. Macwan and Archana N. Vyas worked for the classification of offline Gujarati handwritten characters using three different methods combinations and achieved 87.22% accuracy for total 39 Gujarati characters [11].

Tesseract is a very well maintained and it had been used by several researchers for the OCR of multiple language scripts. Chirag Patel had done a comparative analysis between a propriety OCR tool with Tesseract OCR using English script and found Tesseract is faster as it takes approximately 1 second for processing colored images and 0.82 seconds for grey scale images [4].

Tesseract is having capabilities of getting trained for the languages which are not supported currently or exist but different fonts which are not supported now can be trained with customized trained data for the better recognition. Md. Abul Hasnat tried to integrate Bangla script in Tesseract OCR which was not available earlier using the customized trained data for Bangla Script [2].

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor:6.887
Volume 5 Issue IX, September 2017- Available at www.ijraset.com

Nitin Mishra also experimented on the Shirorekha Chopping for enhanced Hindi language recognition in integration with Tesseract OCR [3].

III.PROPOSED APPROACH

A. Why Tesseract?

Tesseract is a prominent OCR engine in comparison with other OCR engines as it is open source and registered under Apache license and support 100+ languages. It can be used on all major platforms of computer operating systems.

Apart from this It produces its output in different formats like Text, (Portable Document Format) PDF and other formats using own or already available GUIs or (Application Programming Interface) APIs. It is regularly updated and renovated under the supervision of highly expert team of Google Inc.

B. How Tesseract Works

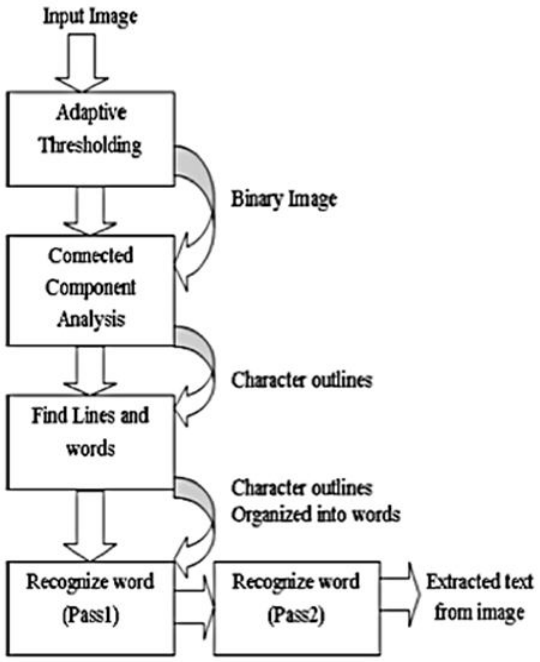


Fig. 4. How Tesseract Works [7]

Tesseract is a very popular and stable open source Optical Character Recognition (OCR) engine today which was initially started as a Doctorate of Philosophy (PhD) research work in Hewlett-Packard (HP) labs, Bristol [5]. Further it was led by Google from 2006 after the Open Source release of Tesseract in 2005. Tesseract can be obtained from the official repositories undertaken by Ray Smith, Google Inc.The repositories are having vast support for various programming languages which can be integrated along with the any programming language. With the use of these Tesseract OCR engine repositories one can test the core features of Tesseract as tesseract has no built-in Graphical User Interface (GUI) support. GUI needed to be developed or can be integrated with any other open source GUI repositories which are compatible and integration can be done with the use of their GUI.

Figure 4. Illustrating the actual working of the Tesseract OCR engine. In very first step an image is inputted on which OCR is needed to be performed after that the image pass through the “Adaptive Thresholding” step in which it gets converted into Binary Image. Later Binary Image further processed into “Connected Component Analysis” in which text/words splitting has been performed with character outlines.

Outlined Characters further sent to 2-way pass to recognize words. In Pass-1 recognized text/words are further processed in to an adaptive classifier which considers the data as trained data. Now text will be again recognized for the second time but now it will use adaptive classifier for recognition.

The reason behind the recognition for the second time is to know the context of the text from Pass-1 and later in the second, third and so on times it can be recognized with ease.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor:6.887
Volume 5 Issue IX, September 2017- Available at www.ijraset.com

C. Using Tesseract to recognize Gujarati Script

To demonstrate the uses of Tesseract OCR to recognize Gujarati Script we have used Visual Studio as a tool and C# as a programming language.

Fig. 5. Sample Web GUI to interact with Tesseract OCR

Fig5. Representing the GUI to interact with the Tesseract OCR for Gujarati Character Recognition. As a prerequisite, majorly three things are necessary to carry on this experiment was –

1) A Tesseract dll to support the programming portion to access Tesseract OCR engine.
2) A trained data.
3) A digitally printed image which contains Gujarati characters.

In this scenario for the testing purpose, the image contains Gujarati characters printed in different fonts and size with black text on white background as sample data.

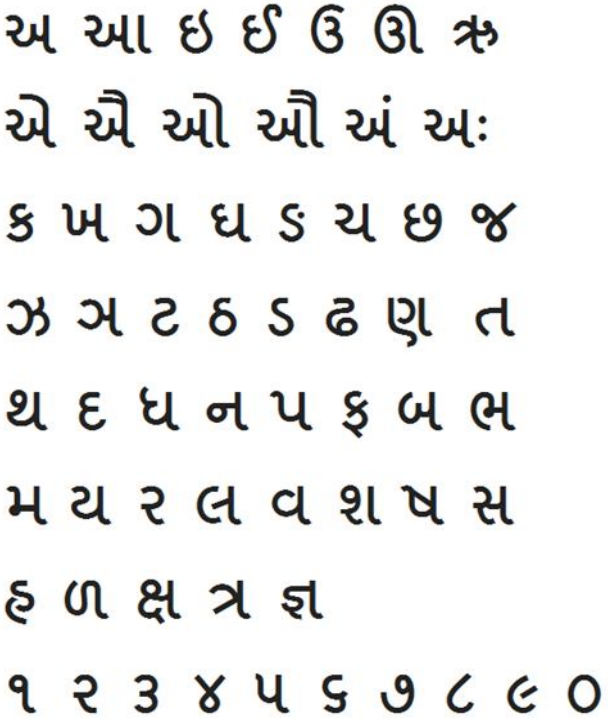
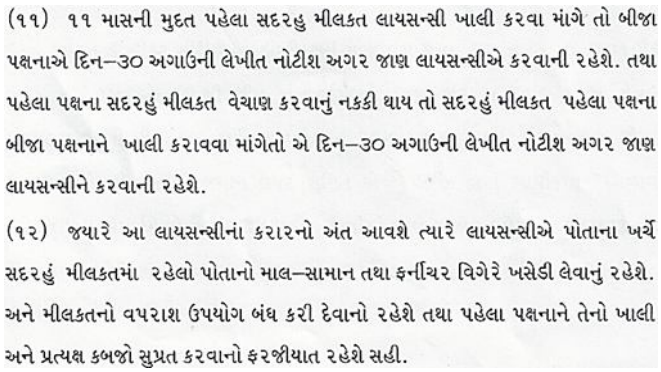


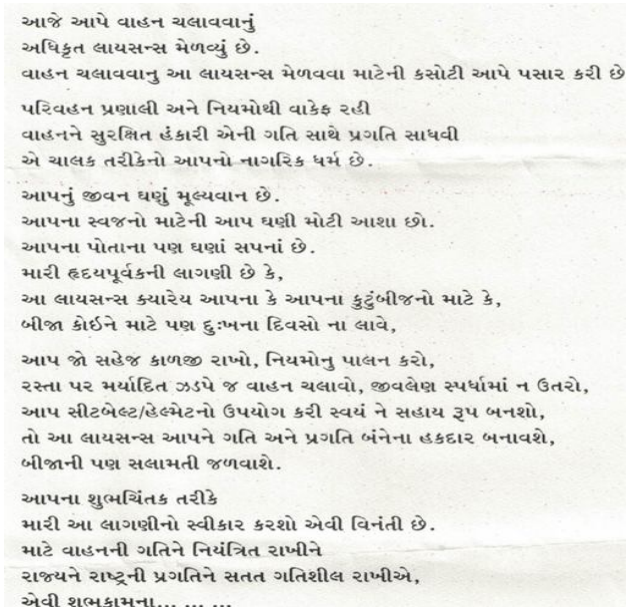
Fig. 6. Image with sample data (1)

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor:6.887
Volume 5 Issue IX, September 2017- Available at www.ijraset.com


(૧૧) ૧૧ માસની મુદત પહેલા સદરહુ મીલકત લાયસન્સી ખાલી કરવા માંગે તો બીજા પક્ષનાએ દિન–૩૦ અગાઉની લેખીત નોટીશ અગર જાણ લાયસન્સીએ કરવાની રહેશે. તથા પહેલા પક્ષના સદરહું મીલકત વેચાણ કરવાનું નકકી થાય તો સદરહું મીલકત પહેલા પક્ષના બીજા પક્ષનાને ખાલી કરાવવા માંગેતો એ દિન–૩૦ અગાઉની લેખીત નોટીશ અગર જાણ લાયસન્સીને કરવાની રહેશે.

(૧૨) જયારે આ લાયસન્સીનાં કરારનો અંત આવશે ત્યારે લાયસન્સીએ પોતાના ખર્ચે સદરહું મીલકતમાં રહેલો પોતાનો માલ–સામાન તથા ફર્નીચર વિગેરે ખસેડી લેવાનું રહેશે. અને મીલકતનો વપરાશ ઉપયોગ બંધ કરી દેવાનો રહેશે તથા પહેલા પક્ષનાને તેનો ખાલી અને પ્રત્યક્ષ કબજો સુપ્રત કરવાનો ફરજીયાત રહેશે સહી.

Fig. 7. Sample image with input data (2)


આજે આપે વાહન ચલાવવાનું
અધિકૃત લાયસન્સ મેળવ્યું છે.
વાહન ચલાવવાનુ આ લાયસન્સ મેળવવા માટેની કસોટી આપે પસાર કરી છે

પરિવહન પ્રણાલી અને નિયમોથી વાકેફ રહી
વાહનને સુરક્ષિત હંકારી એની ગતિ સાથે પ્રગતિ સાધવી
એ ચાલક તરીકેનો આપનો નાગરિક ધર્મ છે.

આપનું જીવન ઘણું મૂલ્યવાન છે.
આપના સ્વજનો માટેની આપ ઘણી મોટી આશા છો.
આપના પોતાના પણ ઘણાં સપનાં છે.
મારી હૃદયપૂર્વકની લાગણી છે કે,
આ લાયસન્સ ક્યારેય આપના કે આપના કુટુંબીજનો માટે કે,
બીજા કોઈને માટે પણ દુઃખના દિવસો ના લાવે,

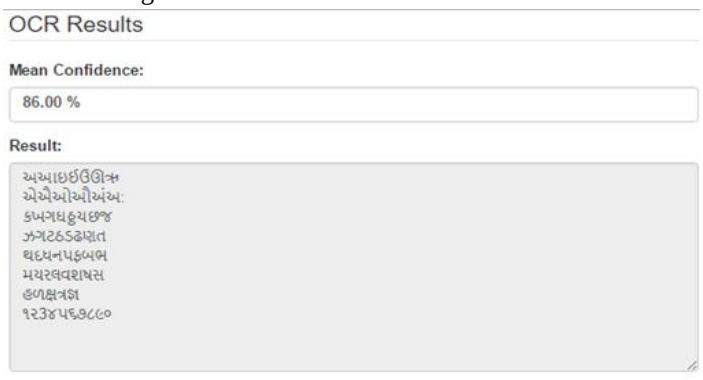
આપ જો સહેજ કાળજી રાખો, નિયમોનુ પાલન કરો,
રસ્તા પર મર્યાદિત ઝડપે જ વાહન ચલાવો, જીવલેણ સ્પર્ધામાં ન ઉતરો,
આપ સીટબેલ્ટ/હેલ્મેટનો ઉપયોગ કરી સ્વયં ને સહાય રૂપ બનશો,
તો આ લાયસન્સ આપને ગતિ અને પ્રગતિ બંનેના હકદાર બનાવશે,
બીજાની પણ સલામતી જળવાશે.

આપના શુભચિંતક તરીકે
મારી આ લાગણીનો સ્વીકાર કરશો એવી વિનંતી છે.
માટે વાહનની ગતિને નિયંત્રિત રાખીને
રાજ્યને રાષ્ટ્રની પ્રગતિને સતત ગતિશીલ રાખીએ,
એવી શુભકામના... ... ...

Fig. 8. Sample image with input data (3)

IV. EXPERIMENTAL RESULT

The output of the experiment conducted is given as below.


OCR Results

Mean Confidence:

86.00 %

Result:

અઆઇઈઉઊઋ
એઐઓઔંઅઃ
કખગઘઙચછજ
ઝઞટઠડઢણત
થદધનપફબભ
મયરલવશષસ
હળક્ષત્રજ્ઞ
૧૨૩૪૫૬૭૮૯૦

Fig9. Output derived from the given input

We have taken various images as sample data input all of 300 dpi. This images are taken from several sources such as images captured from newspaper, books, official documents, etc. which are in different font style, size, and color. Every time after processing through Tesseract OCR we got more than 80% of accuracy which may be enhanced by some preprocessing on images and font specific customized trained data.

V. CONCLUSIONS

This research paper tries a representation of a systematic procedure to printed Gujarati Character OCR with use of the open source OCR engine Tesseract OCR. We used the built-in trained test data provided by Tesseract OCR to recognize the characters from the digitally typed printed images.

We tried to observe the results in various situation using the different font, font style and font size etc. We found it performing satisfying apart from some complex characters and similar looking characters ambiguity.

A detailed study is done for Gujarati Character Recognition using Tesseract OCR Engine which can be used as an initial footprint for upcoming future research work.

VI. ACKNOWLEDGMENT

The authors thank SJD International Surat, Gujarat, India and Narmada College of Computer Application (NCCA), Bharuch, Gujarat, India for the contribution of all the resources access to carry out this research study.

REFERENCES

[1] "Gujarati language - wikipedia," 16 February 2017. [Online]. Available: https://en.wikipedia.org/wiki/Gujarati_language.
[2] "Tesseract OCR GitHub," 16 02 2017. [Online]. Available: https://github.com/tesseract-ocr.
[3] M. A. Hasnat, M. R. Chowdhury, M. Khan and others, "Integrating Bangla script recognition support in Tesseract OCR," BRAC University, 2009.
[4] N. Mishra, C. Patvardhan, V. C. Lakshimi and S. Singh, "Shirorekha Chopping Integrated Tesseract OCR Engine for Enhanced Hindi Language Recognition," International Journal of Computer Applications, vol. 39, no. 6, pp. 19-23, 2012.
[5] C. Patel, A. Patel and D. Patel, "Optical character recognition by open source OCR tool tesseract: A case study," International Journal of Computer Applications, vol. 55, no. 10, 2012.
[6] R. Smith, "An overview of the Tesseract OCR engine," Document Analysis and Recognition, 2007. ICDAR 2007. Ninth International Conference on, vol. 2, pp. 629-633, 2007.
[7] C. Verstraeten, "How to train Tesseract 3.01 - Cédric Verstraeten," 16 02 2017. [Online]. Available: https://blog.cedric.ws/how-to-train-tesseract-301.
[8] Patel, C., & Desai, A. (2013). Gujarati Handwritten Character Recognition Using Hybrid Method Based On Binary Tree-Classifier And K-Nearest Neighbour. International Journal of Engineering Research & Technology (IJERT), 2(6), 2337-2345.
[9] S. A. Chaudhari and R. M. Gulati, "An OCR for separation and identification of mixed English — Gujarati digits using kNN classifier," International Conference on Intelligent Systems and Signal Processing (ISSP), pp. 190-193, 2013.
[10] J. Dholakia, A. Negi and S. R. Mohan, "Zone Identification in the Printed Gujarati Text," International Conference on Document Analysis and Recognition (ICDAR'05), vol. 1, pp. 272-276, 2005.
[11] S. J. Macwan and A. N. Vyas , "Classification of Offline Gujarati Handwritten Characters," InternationalConference on Advances in Computing, Communications and Informatics (ICACCI), pp. 1535-1541, 2015.