IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-6, Issue-6, Jun 2018 E-ISSN: 2347-2693

Implementation of Optical Character Recognition Using Machine Learning

Vishal Chourasia[1], Sanjay Silakari[2], Rajeev Pandey[3]

[1,2,3] Department of Computer Science Engineering, University Institute of Technology, RGPV, Bhopal (India)

Available online at: www.ijcseonline.org

Accepted: 21/Jun/2018, Published: 30/Jun/2018

Abstract : With the passing of time, the realm of human knowledge is ever expanding. Further, with each passing day, we witness the explosion of information which is evident in life style, social events and breakthrough in medical science. The human beings from time memorial have attempted to preserve the information for posterity by adopting various forms starting with pictorial forms in stone carvings and subsequently recorded in palm leafs, metal sheets, as well as leather sheets. With the invention of paper and subsequent electronics, the information is recorded with ease and could be transferred to any corner of world within seconds, but modern technology, facilitating electronic preservation of information faced a challenging task of gigantic and herculean proportion while it preserving information, voluminous in quantity, recorded on papers, from preceding centuries, into electronic form. The same became more difficult with numerous languages spoken and written by people from every corner of the world. Adoption of Optical Character Recognition (OCR), producing editable text out of text image documents, has reduced the problem to a great extent. Even though, the OCR is fairly advanced in major languages like English, French etc. Various random images are taken for simulation then accuracy is measured to conclude the efficiency of the OCR system.

Keywords – Optical Character Recognition (OCR), Editable Text, Modern Technology, Feature Extraction

I. INTRODUCTION

Interpretation of textual contents in document images through computer systems is the major inclination of optical technologies like Optical Character Recognition (OCR). The transformation of text document images to its equivalent editable format through computing machines is motivation behind the development and evolution of OCR systems.

Initial attempts on OCR development have been successful in automatic reading and data entry of various Roman and Latin language scripts. Further, the progressions towards various Indian language scripts are accorded to extend the functionality of OCR to read many scripts. Even though there exist many successful investigations on OCR towards Indian scripts, there are still many challenging research issues to be addressed in this regard. One of the critical barriers that hinder the path to reach the successful recognition rate is the complexity involved in script. Especially, this is true with many of the Indian Language scripts due to their wide character set and structural diacritics. Therefore, it is essential to address the complexities of simple issue language (SIL) scripts in order to reach higher recognition rates.

This research is focused on exploring solutions to the various research challenges involved in interpretation and recognition of one of the SIL script. The diverse business needs and smart technological advancements require the simulation of human activities to be accomplished through computing machines efficiently. OCR is one of such software that simulates the vision function performed by humans to read or interpret via computing machines. Basically, the activity of reading is associated with the knowledge of language and the ability to interpret its script. The knowledge of the script by processing images can be instilled as functionality to OCRs' through the vast knowledge domain of Digital Image Processing (DIP) and Machine Learning (ML) techniques.

At present, Main Motivation is higher recognition accuracies by OCR technologies can be assured only with restrictions like absence of broken/complex compound characters and good resolution of document images etc. Comparatively techniques employed for handwritten character recognition are still flourishing with several additional instructions to process text image data correctly.

The task of development of a complete OCR system for any script is still in its infancy. Different categories of character recognition methods are shown in Figure 1.1

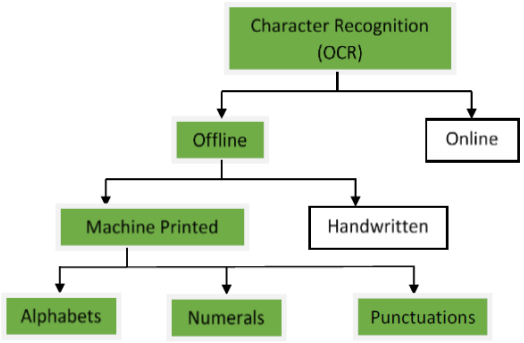

Figure 1 : Categories of character recognition system

According to the image acquisition, the OCR systems can be of two types such as:

a. Off line systems and
b. Online systems.

Off line OCR systems are needed to be installed on computers, which accept image documents and convert it into editable text document of the type of language required.

However, the Online OCR systems are applications that save the input at the time of writing. The user of the system gives the direct input on a digital device using a special stylus. The system records the point sequence with respect to the direction of movement of the stylus.

Again, based on the type of script accepted as input by the OCR system, it can be further categorized into two types:

a. Printed OCR systems;
b. Handwritten OCR systems;

Input given to the printed OCR system accepts hard copy documents such as Printed, Scanned and Xerox documents of any language. Thus, it helps to preserve the old ancient historical, mythological copies in digital form. The recognition character is extremely high as sizes of the printed characters are consistent and the spaces between characters are also uniform. The only difficulty is to collect the dataset for training the system.

Furthermore, the handwritten system accepts the image documents written by people. Thus, it is more complicated and recognition of accuracy is very less as a single character can be written differently by different persons. It is very easy to create a dataset for the system that need only involvement of few persons.

II. BACKGROUND

Optical Character Recognition (OCR) is a piece of software that converts printed text and images into digitized form such that it can be manipulated by machine. Unlike human brain which has the capability to very easily recognize the text/characters from an image, machines are not intelligent enough to perceive the information available in image. Therefore, a large number of research efforts have been put forward that attempts to transform a document image to format understandable for machine. OCR is a complex problem because of the variety of languages, fonts and styles in which text can be written, and the complex rules of languages etc. Hence, techniques from different disciplines of computer science (i.e. image processing, pattern classification and natural language processing etc. are employed to address different challenges. This paper enlightens the reader with the historical perspectives, applications, challenges and techniques of OCR.

The process of OCR is a composite activity comprises different phases [1]. These phases are as follows:

Image Acquisition: To capture the image from an external source like scanner or a camera etc.

Preprocessing: Once the image has been acquired, different preprocessing steps can be performed to improve the quality of image. Among the different preprocessing techniques are noise removal, thresholding and extraction image base line etc.

Character Segmentation: In this step, the characters in the image are separated such that they can be passed to recognition engine. Among the simplest techniques are connected component analysis and projection profiles can be used. However in complex situations, where the characters are overlapping, broken or some noise is present in the image. In these situations, advance character segmentation techniques are used.
inter-class variations.

Feature Extraction: The segmented characters are then processes to extract different features. Based on these features, the characters are recognized. Different types of features that can be used extracted from images are moments etc. The extracted features should be efficiently computable, minimize intra-class variations and maximizes inter-class variations.

Character Classification: This step maps the features of segmented image to different categories or classes. There are different types of character classification techniques. Structural classification techniques are based on features extracted from the structure of image and uses different decision rules to classify characters. Statistical pattern

classification methods are based on probabilistic models and other statistical methods to classify the characters.

Post Processing: After classification, the results are not 100% correct, especially for complex languages. Post processing techniques can be performed to improve the accuracy of OCR systems. These techniques utilizes natural language processing, geometric and linguistic context to correct errors in OCR results. For example, post processor can employ a spell checker and dictionary, probabilistic models like Markov chains and n-grams to improve the accuracy. The time and space complexity of a post processor should not be very high and the application of a post-processor should not engender new errors.

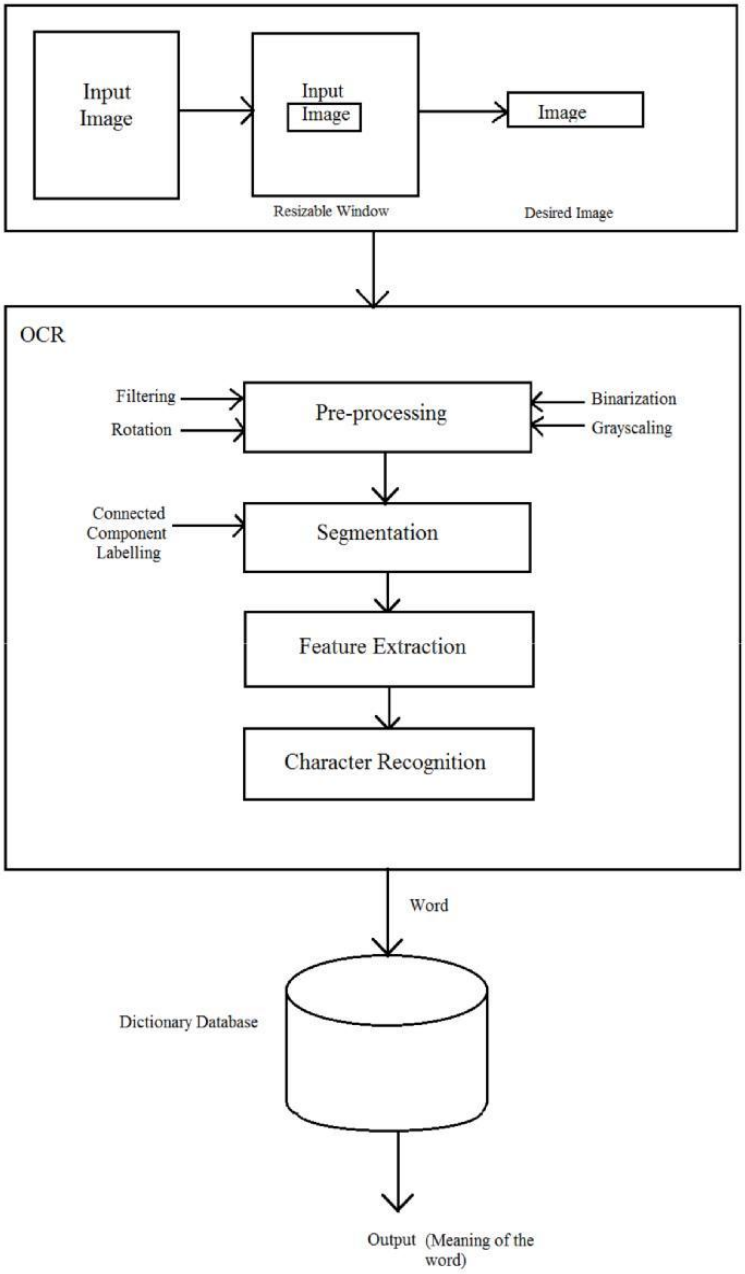

Figure 2- Flow Chart of OCR Functioning

In the research paper [2], Tao Wang *et*. al. proposed Full end-to-end content recognition in regular images is a testing issue that has gotten much attention as of late. Customary systems around there have depended on expound models fusing thoroughly hand built highlights or a lot of prior learning. In this manuscript, they take an alternate route and join the authentic power of substantial, multilayer neural networks together with late improvements in unsupervised feature learning, which enables us to utilize a typical framework to prepare highly precise text detector and character recognizer modules. At that point, utilizing just straightforward off-the-rack methods, we integrate these two modules into a full end-to-end, lexicon-driven, scene text recognition structure that achieves state-of-the-art presentation on standard benchmarks, specifically Street View Text and ICDAR 2003.

III. IMAGE PROCESSING

Accurate interpretation of contents in captured digital images is subjected to the magnitude of quality in images. Befitting the needs of maintaining the required image resolution, DIP equips various pre-processing techniques towards improving the image quality. Pre-processing is a prerequisite operation to be performed on images for transforming them to a suitable form to meet the needs of subsequent stages. Most of the investigations have been carried out on image pre-processing operations from past few decades to date.[3]

Pre-processing techniques usually vary based on the type of noise content available in an image. The selection of an appropriate pre-processing technique to process an image plays a vital role in attaining the desired outputs. Some of the major categories include medical, industrial, satellite, forensic, surveillance, biological and document images. As the current research aims at performing the interpretation of contents in document images, the investigated pre-processing techniques are focused much towards the necessities of text in document images [1,4].

Pre-processing is the most essential operating protocol desired for text document images to induce better efficiency in OCR systems. Text document images are of varied categories ranging from simple text images to documents with complex graphical notations and text, which are termed as pre-printed documents.[3] The pre-processing procedures employed for pre-printed documents differ from usual methods and require intensive processing especially in case of detection and removal of graphical components. Suggested method uses edge detection filters, rectangular structuring element and area features of connected components.

Pre-processing is a preliminary operation required for obtaining clear and noise free image leading to a suitable form for subsequent processing. In the perspective of document images, it is an obligatory processing task without which feasible recognition accuracies cannot be attained. Basically, the pre-processing operation varies from one type of document image to other. In real time variety of documents exists, the pre-processing operations befitted for one type of documents are not suitable for other type. The pre-processing of document image in fact is most intensive as well as expensive operations, especially in the case of pre-printed documents.

In this research the focus is mainly on the pre-processing required for pre-printed documents, which are composed of graphical components as well as textual components.[4] The graphical components represent the horizontal and vertical lines, photography, emblems, logos, scratched marks and artistic text etc. Detection or removal of such graphical components is more significant, as the presence of these components makes the subsequent processing stages as erroneous.

The investigations in this direction focus on three important pre-processing tasks i.e., detection and removal of horizontal and vertical lines, detection of scratched words in pre-printed documents and printed and handwritten text classification. Some of the important works on these pre-processing tasks are discussed subsequently:

1. **Binarization of Documents:**
   Document binarization is a technique of converting a gray scale or color image to a binary image. Generally, the image acquired for processing will be gray scale or color image available in any of the common image formats JPEG, PNG, BMP, GIF, TIFF [5]. As the DIP mainly focuses on characterization of text from its background, it is desirable to process the document images in binary or monochrome image format or gray scale format. The process of converting a RGB or gray scale image to binary image is known as thresholding. The thresholding techniques range from global level to multi-level thresholding techniques. The thresholding technique returns a discrete outcome to decide the intensity level that should be assigned to a target pixel in the output image. The outcome of binarization process depends on the quality of the input image and the type of thresholding technique employed for conversion.

2. **Enhancement of Documents:**
Implication of direct line elimination on pre-printed document an image leave the document with broken line fragments and is unconnected from base edges. Hence, it is very much significant to perform the enhancement of gradient details like grid lines and characters. The binarized image obtained from the thresholding process is subjected to the process of enhancement to improve the gradient quality of both characters as well as the layout of document. The enhancement is performed on a binary image by using order static filters [6] with mask of size 3x3. The median filtered image is further subjected to the process of specific pre-processing tasks such as line removal and scratch mark removal.

3. **Line Elimination:**
The enhanced image is assumed as input to the line elimination algorithm. The algorithm processes the image by identifying the rows with horizontal lines initially and then convolving the identified lines using circular structuring element subsequently. The rows with horizontal lines in document image are identified by using a distinguishing characteristic of line. The distinguishing characteristic employed in the proposed methodology is "a line is a continuous sequence of some '**n**' number of black pixels" where $20 \leq n \leq cols$ and *cols* represent the total number of columns in an image.
The rows identified in this stage are convolved with circular structuring element. The convolution of image with circular structuring element represents moving the origin of structuring element through each pixel and computing the weighted average.

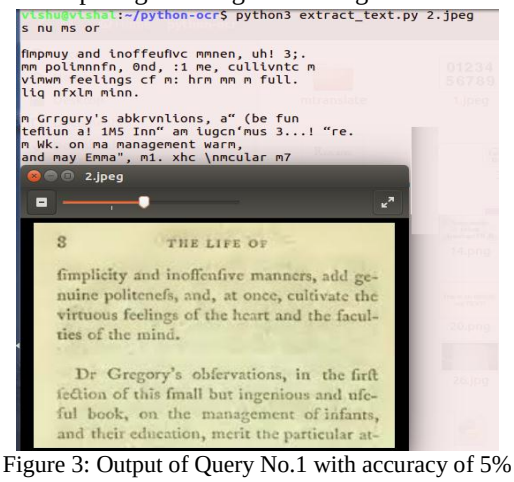

Figure 3: Output of Query No.1 with accuracy of 5%

4. **Edge detection and Filtering:**
Line detection in this methodology is carried out using edge detection operators. The edge detection and filtering is the process of identifying the rows and columns with horizontal-vertical edges and marking those rows for subsequent processing. The process of line removal includes the challenge of retaining the line pixels with text stroke crossings.
The detection of horizontal and vertical edges in the enhanced image is accomplished using the second order derivative operator, The Laplacian [7]. The process of marking the rows or columns with only edges at horizontal or vertical orientations reduces the computation time involved in detection of text strokes crossing horizontal lines.

IV. IMPLEMENTATION ENVIRONMENT AND RESULTS

In the course of recent decades Machine Learning has turned out to be one of the pillars of data innovation and with that, a somewhat focal, yet generally concealed, some portion of our life. With the regularly expanding measures of information getting to be accessible, there is good reason motivation to trust that keen data analysis will turn out to be significantly more inescapable as an important ingredient for technological advance [8].

Python is a universally useful deciphered, intelligent, object-oriented, and high-level programming dialect. It was made by Guido van Rossum amid 1985-1990. Like Perl, Python source code is additionally accessible under the GNU General Public License (GPL). This instructional exercise gives enough comprehension on Python programming dialect.

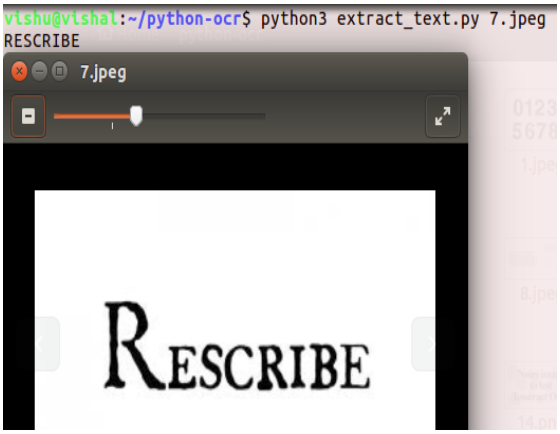

Figure 4: Output of Query No.2 with accuracy of 100%

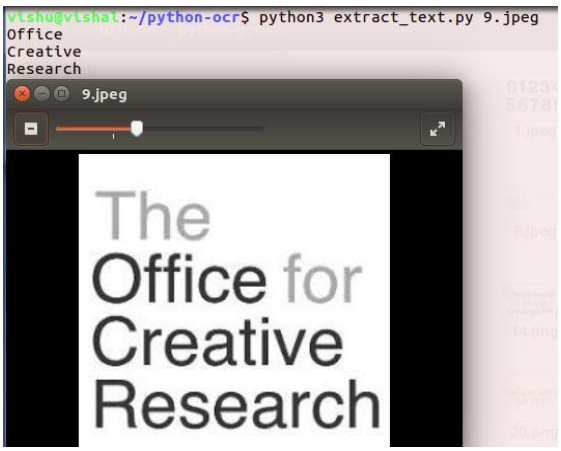

Figure 5: Output of Query No.3 with accuracy of 80%

Figure 6: Output of Query No.4 with accuracy of 50%

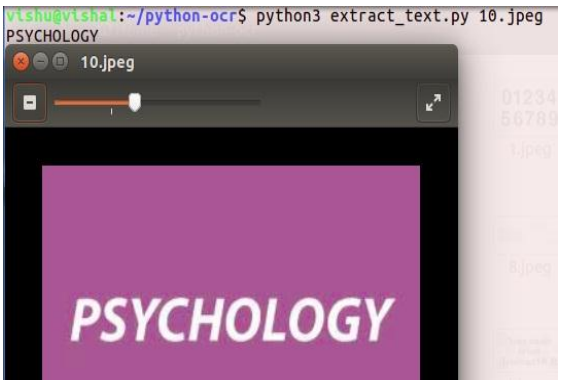

Figure 7: Output of Query No.5 with accuracy of 100%

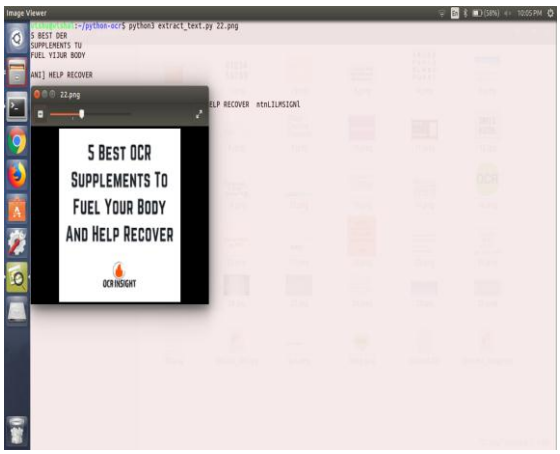

Figure 8: Output of Query No.6 with accuracy of 70%

V. CONCLUSION

Around 25 images are there for simulation, and then accuracy of system is calculated on the basis of recognized characters. On the basis of simulation of this offline optical character recognition, **we concluded that the accuracy of our proposed system is 85-90% approximately.** In this concluding section, the Researcher elucidates how the objectives of the research, set in the beginning of the research, had been met by the methods that had been propagated during the course of the research.

The future possibilities are listed and explained in the following list:

(1) A complete and robust OCR needs to be developed to accept and convert multiple pages at a time.

(2) More research is to be needed to develop a multi lingual OCR system.

(3) The work can also be extended for handwritten and postal address recognition.

(4) The recognition accuracy is achieved by increasing the number of training of samples.

(5) The rate of recognition can also be improved by adopting scanned document cleaning techniques to reconstruct the corrupted scanned texts.

REFERENCES

[1]. Vishal Chourasia, Dr. Sanjay Silakari, Dr. Rajeev Pandey, "A Survey Paper on Optical Character Recognition using Machine Learning ", International Journal of Computer Technology and Applications, ((IJCTA)Vol 9(3), 160-164 ,2018

[2]. Tao Wang, David J. Wu, Adam Coates, Andrew Y. Ng, "End-to-End Text Recognition with Convolutional Neural Networks" Stanford University, CA 94305.2013

[3]. Aayushi Jain, Nitish Joshi, MayureshKhendkar "Implementation of OCR Based on Template Matching and Integrating in Android Application(IJCSE)-04, Issue-02, 2016

[4]. R. Smith, "An overview of the Tesseract OCR Engine", Proc 9th Int. Conf. on Document Analysis and Recognition, IEEE, Curitiba, Brazil, Sep 2007, pp629-633.

[5]. Ankush Gautam, "Segmentation of Text from Image Document", International Journal of Computer Science and Information Technologies, Vol. 4, Issue 3, pp. 538-540, 2013.

[6]. P.B. Pati, S. Sabari Raju, N. Pati, A. G. Ramakrishnan, "Gabor filters for document analysis in Indian bilingual documents", International Conference Intelligent Sensing and Information Processing, pp. 123- 126, 2004.

[7]. Ranjeet Srivastava., Ravi Kumar Tewari., Shashi Kant., "Separation of machine printed and handwritten text for Hindi documents", International Research Journal of Engineering and Technology (IRJET), Vol. 2, Issue 2, pp. 704--708, 2015.

[8]. Saba, Tanzila, Amjad Rehman, Mohamed Elarbi-Boudihir, "Methods and strategies on off-line cursive touched characters segmentation: a directional review", Artificial Intelligence Review, Vol. 42, Issue 4, pp. 1047-1066, 2014.

[9]. Anil R, Arjun Pradeep, Midhun E.M, Manjusha K., "Malayalam Character recognition using singular value decomposition", International Journal of Computer applications, Vol. 92, Issue 12, pp.6-11, 2014.

[10]. Apurva A. Desai, "Gujarati handwritten numeral optical character reorganization through neural network", Pattern Recognition, Vol. 43, Issue 7, pp. 2582–2589, 2010.

