A Review on OCR Post Processing Error Correction Algorithm

Mayur Burhan, Mimoh Samarth, Mrunal Talokar, Nikhil Gaikwad
Department of Computer Science & Engineering, Rajiv Gandhi College of Engineering and Research, Nagpur, Maharashtra, India

ABSTRACT

Building an effective method to detect characters from images with less error rate is the great task. Our aim is to give such an algorithm that will be generated error-free recognition of text from the given image. For this purpose, OCR was developed to translate scanned text into editable computer text. Unfortunately, OCR is still imperfect as it falsely identifies scanned text leading to misspellings and errors in the OCR output. Hence a post-processing technique used for detecting and correcting OCR non-word and real-word errors. The aim of our project is to develop an “Android app for translating text into a known language which is named as TransLocator.”

Keywords : Optical Character recognition, Post processing, Text mining, Error Correction

I. INTRODUCTION

The drastic change in computer technology that changes the life of the people to communicate, learn and sharing information. In the 21st era, people use digital format over tradition one. The present-day every person have their own computers and handheld devices. It is easy to use digital format. Anything that requires computer processing must be converted into the digital format. The rapid development of Internet and computer technology has much-improved people’s information acquisition ability. Routine works in various application domains (from market analysis and scientific research to propaganda and intelligence analysis) have benefited from the vast amount of information, which were otherwise unavailable without proper data and text processing techniques [8].Main tool uses for the processing is OCR (Optical Character Recognition). OCR is tool that converts the image into the digital format for the processing. OCR could be a system that gives full alphabetical recognition of written characters at merely scanning the document.

OCR is the method of digitization of text of printed data into the digitized-retrievable form.It enables devices with a camera to see, read and transform the data into digital output. But the problem is still present in the OCR. For instance, character “B” can be improperly converted into a number “8”, character “S” into a number “5”, character “O” into number “0”, and so forth. For the correction of the OCR output, we have to study the post-processing technique [10].

II. OCR

Optical Character Recognition (OCR) is a computerization approach which converts images of typed, handwritten or printed text into machine-encoded text. OCR is the method of digitization of hand cum typewritten or text of printed data into the digitized-retrievable form. OCR is the most active interesting evaluation of text cum character processing recognition and pattern-based image recognition. With present life, OCR has been successfully using in finance, legal, banking, healthcare and home need appliances. Character Recognition classified in two ways, online and offline. The OCR consists the different levels of processing methods like as Image Pre-Acquisition, Classification, Post-Acquisition, Pre-Level processing, Segmented

Processing, Post-Level processing, Feature Extraction [3]

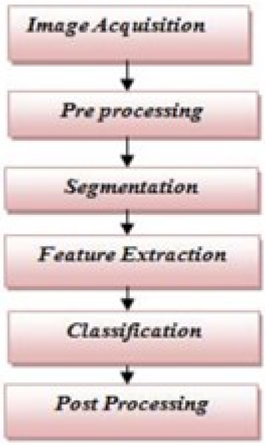

Figure 1. Traditional Steps of Character Recognition [9]

III. OCR POST-PROCESSING

OCR post-processing is the later stage of the image to text conversion (OCR). The aim of the post-processing is to correct the linguistic misspellings in the OCR output text after the image has been scanned and completely processed. OCR post-processing can be done in two ways that are offline or online. Once an image is taken, an OCR engine applies various pre-processing techniques such as de-skewing, de-speckling, finalization, and line removal, etc. to improve the chances of successful recognition. There are different techniques for OCR post processing like manual error correction, dictionary based error correction, context based error correction, etc. As these technique is not so effective. Therefore, countless post processing approaches and algorithms were proposed in an attempt to detect and correct OCR errors. Some of them are as follows

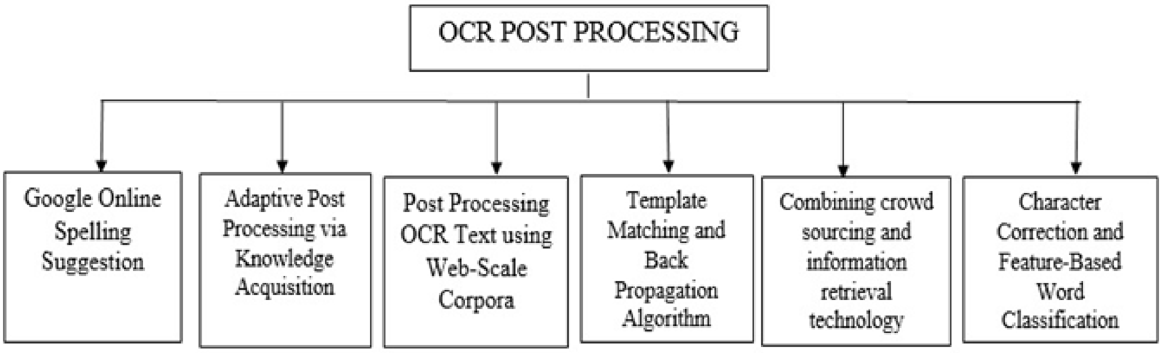
OCR POST PROCESSING
Google Online Spelling Suggestion
Adaptive Post Processing via Knowledge Acquisition
Post Processing OCR Text using Web-Scale Corpora
Template Matching and Back Propagation Algorithm
Combining crowd sourcing and information retrieval technology
Character Correction and Feature-Based Word Classification

Figure 2. OCR Post Processing Technique

3.1 OCR Post Processing Error Correction Algorithm using Google Online Spelling Suggestion

A new post-processing technique for OCR is error correction algorithm using Google online spelling suggestion [10]. This approach uses Google's giant warehouse of indexed real-world pages, articles, blogs, forums, and other sources of text, it can suggest common spellings for words not found in standard dictionaries. In this algorithm, they use Google's "did you mean" technology. In this experiment, there is a large improvement in the OCR error correction. As the higher number of misspellings and linguistic errors were detected and corrected using the proposed method compared with other traditional existing ones.

This algorithm doesn't support multiprocessing platforms to run in a parallel fashion over a bunch of concurrent processors or even over a bunch of distributed computing machines. The main drawback of this algorithm requires users connected to the internet.

3.2 Adaptive Post Processing of OCR Text via Knowledge Acquisition

In this post-processing, OCR technique they have [3]solve the problem of conversion map, character pair matrix, most error-prone characters and the most likely unrecognized characters employed, and knowledge acquisition scheme is used so that errors can be detected and corrected more accurately and efficiently. An intelligent post-processing technique using knowledge acquisition and self-learning can be very efficient and effective [11].

3.3 OCR Post Processing **using template matching and back propagation algorithm**

In this technique they have discussed complete methodology to implement character recognition, it has been observed that template matching a feature extraction method when applied with ANN (artificial neural network) i.e. the back-propagation despite low convergence rate of back-propagation and pattern dependency offer several advantages in pattern recognition rate, we observed that feature extraction is probably the single most important factor for gaining high accuracy [4].

3.4 Post-Processing OCR Text using Web-Scale Corpora

There is one technique introduced for an OCR post-processing system for automatically correcting OCR-generated errors in the text. Web-scale corpora applied to candidate generation and linguistic analysis for each feature. By integrating different linguistic analysis results in a regression process, to select and rank high-quality results. Evaluated on an OCR-generated natural history book, our system can correct 61.5% of the errors in a fully automatic way and 76.62% by interactive selection from among the top 10 candidates [12].

3.5 User-driven correction of OCR errors. Combining crowdsourcing and information retrieval technology

A new approach for correction noisy OCR text which combines the power of crowdsourcing with information retrieval technology. The actual implementation uses the standard information retrieval engine Lucerne which is a powerful tool for searching and retrieving large amounts of text. There are many issues with this approach like [13].

1. How to recruit and retain users?
2. What contributions can users make?
3. How to combine user contributions to solve the target problem?

3.6 OCR Error Correction Using Character Correction and Feature-Based Word Classification

This technique explores the use of a learned classifier for post-OCR text correction. This approach, which integrates a weighted confusion matrix and a shallow language model, improves the majority of segmentation and recognition errors, the most frequent types of error on our dataset [14].

IV. COMPARATIVE STUDY OF OCR POST PROCESSING TECHNIQUE

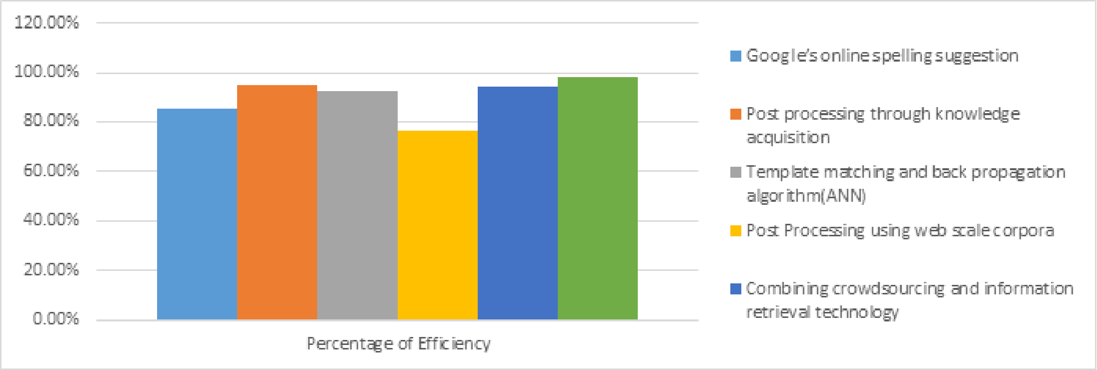

Figure 3

V. CONCLUSION

This paper gives the detailed survey on post-processing techniques used in OCR (Optical Character Recognition). OCR is a technique that came across from a long time but it still has an error to overcome this error there are several post-processing techniques have developed. In this paper, we have studied six different methods for post-processing. The most efficient post-processing technique still required some future work. In character correction and feature based word classification there is acquisition ranking can be achieved by building a domain specific language model giving a considerable gaining strength for the language dependent features suggestion and In Post processing through knowledge Depending on situations every method has its own advantages but still crowdsourcing has good efficiency compare to five other post-processing technique.

VI. REFERENCES

[1]. Journal of Emerging Trends in Computing and Information Sciences, ISSN 2079-8407, Vol. 3, No. 1, January 2012 http://www.cisjournal.org/journalofcomputing/archive/vol3no1/vol3no1_7.pdf

[2]. An Overview of the Tesseract OCR Engine Ray Smith Google Inc. theraysmith@gmail.com Proc. International Conference on Document Analysis and Recognition, ICDAR'2007, Curitiba, Brazil, Sep. 2007

[3]. ADAPTIVE POST-PROCESSING OF OCR TEXT VIA KNOWLEDGE ACQUISITION Lon-Mu Liu, Yair M. Babad, Wei Sun, and Ki-Kan Chan Department of Information and Decision Sciences University of Illinois at

[4]. Chicago (M/C 294) Box 4348, Chicago, Illinois 60680

[5]. A novel OCR approach based on document layout analysis and text block classification Weiheng ZHU1 Yuanfeng LIU2 Liang HAO1 1.Department of Computer Science 2. Information Technology Research Institute Jinan University Guangzhou, China

[6]. Optical character recognition using template matching and back propagation algorithm Swapnil Desai CSED, Thapar University Patiala (Punjab), India swapnil.innovative@gmail.com Ashima Singh Assistant Professor, CSED, Thapar University Patiala (Punjab), India

[7]. Review of the Character Recognition System Process and Optical Character Recognition Approach Jaswinder Kaur, Mrs. Rupinder Kaur CSE & PTU,

[8]. OCR Post-processing Using Weighted Finite-State Transducers Rafael Llobet, J.Ramon Navarro-Cerdan, Juan-Carlos Perez-Cortes, JoaquimArlandisInstitutoTecnologico de Informatica Universidad Politecnica de Valencia Camino de Vera s/n, 46071 Valencia, Spain {rllobet, jonacer, jcperez, arlandis}@iti.upv.es ∗

[9]. A novel OCR approach based on document layout analysis and text block classification

[10]. Optical character recognition using template matching and back propagation algorithm Swapnil Desai Ashima Singh

[11]. OCR POST-PROCESSING ERROR CORRECTION ALGORITHM USING GOOGLE'S ONLINE SPELLING SUGGESTION Youssef Bassil, Mohammad Alwani

[12]. Comparative Study with Analysis of OCR Algorithms and Invention Analysis of Character Recognition Approached Methodologies Santosh Kumar Hengel and B. Rama2 1 , 2Department of Computer Science, Kakatiya University, Warangal, Telangana State,

[13]. Post-Processing OCR Text using Web-Scale Corpora Jie Mei† , Aminul Islam , Abidalrahman Moh'd† , Yajing Wu† , Evangelos Milios† †Faculty of Computer Science, Dalhousie University {jmei,amohd,yajing,eem}@cs.dal.ca School of

Computing and Informatics, University of Louisiana at Lafayette

[14]. GAijnter MAijhlberger, Johannes Zelger, and David Sagmeister. 2014. User-driven Correction of OCR Errors: Combining Crowdsourcing and Information Retrieval Technology. In Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage (DATeCH '14). ACM, New York, NY, USA, 53-56. https://doi.org/10.1145/2595188.2595212

[15]. OCR Error Correction Using Character Correction and Feature-Based Word Classification Ido Kissos School of Computer Science, Tel Aviv University Ramat Aviv, Israel Nachum Dershowitz School of Computer Science