International Journal of Technical Innovation in Modern Engineering & Science (IJTIMES)
Impact Factor: 5.22 (SJIF-2017), e-ISSN: 2455-2585
Volume 4, Issue 10, October-2018

A Comparative study of GOCR, Tesseract and Improved Tesseract for Character Recognition

Priyanka Kumari [1], Arvind Kalia [2]

[1,2]*Department of Computer Science, Himachal Pradesh University*

Abstract— Optical Character Recognition is a technique which is used to change printed document or handwritten character to readable text format. If the text is in printed form against a complex background, it is usually very difficult to perform the detection method. OCR localization approach is used to detect the horizontally aligned text automatically. Tesseract and GOCR are most widely used tools in optical character recognition to recognize the characters from the given input. Tesseract provides better accuracy with low error rate however, in some cases it is unable to recognize text written in image, so there is a need for improvisation of Tesseract tool. This paper introduces a new approach to improvise Tesseract tool by using Artificial Neural Network.

Keywords— Tesseract, GOCR, ANN, OCR.

I. INTRODUCTION

Optical Character Recognition is popularly known as OCR. It is digital image processing method used for decryption of text embedded in an image or object. This method is sensitive to both handwritten and electronic text [1]. Optical Character Recognition (OCR) is usually considered as the most efficient and economical way of regulating a task which is used for obtaining data from the documents stored [2]. The utility based on OCR has found its applications in various fields such as printing, publishing, libraries, law and government offices, etc. Presently, such agencies prefer documents to be stored electronically with the help of image scanning process saving its space and providing easy assessing property [3]. The OCR software further simplifies the searching of text electrically. It basically operates on pattern recognition methodology by scanning the given text based image. The image formed is compared with bitmap images before the process of recognition, in order to yield best suitable results [4]. These results basically rely over the original image quality and then over the scanner. For the sensitive or complicated tasks, the OCR can be developed more on intelligent basis. The ability of OCR helps to recognize the handwritten text that is used to secure the secret information and to check the authenticity [5]. In addition, it can detect the accuracy and spelling errors of the document. It is also feasible with distinct languages and fonts [6]. The process of Optical character Recognition can be broadly classified into four stages which are as follows Pre-processing, Feature Extraction, Segmentation and Classification [7,8]. Some of the major applications of OCR are data entry, text automation [9], multimedia system design, process automation [10] and language processing [11].

Optical Character Recognition tool is used to change different type of files, such as scan paper document, PDF file into searchable data [7]. In this paper two OCR tools has been used namely Tesseract and GOCR. Tesseract is open source optical character identification tool. It was developed under the Apache license [20]. Tesseract is a command based tool and it was written in C++ language [13]. Tesseract works in stages, after adaptive thresh holding which convert the digital image into binary image [20]. Firstly it got outlined by connected component analysis. Than organized those outline to blobs and blobs to text line [22]. The second tool is GOCR tool. It is a command based tool. It was written in C language [14]. It is a fast and simple engine that does not require any trained set of data [15].

The remaining paper is organized into four sections. In section 2 the survey of Optical Character Recognition along with tools and techniques are discussed. Section 3 consists of Proposed Method for recognizing the character from image. Section 4 discusses test investigation and examinations of GOCR, Tesseract and proposed method. Section 5 gives conclusions and future scope.

II. RELATED STUDY

Shivani Dhiman [20] described a comparative case study of Tesseract and Gocr tool on the basis of accuracy and various parameters such as image type, resolution, brightness and font type. It was concluded that Tesseract provides better accuracy than GOCR but in some cases GOCR provides better cases. The main limitation of this study was as it combines rules of fuzzy logics like minmax and maxmin which sometimes not able to recognize text from images robustly.

Patel et al. [3] described a case study on optical character recognition by applying Tesseract and Transym tool by using Vehicle number plate as input. Both tools have been used to extract vehicle numbers from vehicle number plates and these tools were compared on various parameters. The Artificial Neural Network technique was used. The conclusion was drawn that transym is more correct in extracting text from the vehicle number plates, and transym provide better accuracy as compared to the Tesseract tool. The main flaws of this study was as Tesseract works on small character set but there is need of tool which can work on large character set.

Munish et al. [16] proposed an offline handwritten Gurumukhi character recognition system. A Sample of offline Gurumukhi character was collected from one hundred different writers. In this paper, diagonal features and transition features were computed using K-nearest Neighbours classifier. It is concluded that the diagonal feature provides a better accuracy of 94.12%. However, this study has not been done by using the larger character set.

Sahil Badla [17] describes improvement of ability of Tesseract OCR engine. It improved the ability of the Tesseract OCR system to make it run on the mobile devices. The focus was done on enhancement of the Tesseract OCR ability for Hindi language to run on the mobile devices. The ability of the Hindi text extraction was improved for the mobile phone contributed toward Hindi OCR. But this approach has not been used for translation of the hindi word.

Abul Hasnat et al. [18] describes the integrating bangla handwriting recognition base in tesseract OCR. In order to make tesseract identify some other script, the engine can be trained with significant data. It was concluded with a procedure to bangla printed text, tesseract OCR engine which is open source series of test were conducted to decide the amount of training data required, and to understand tesseract requirements. This study has the limitation as it works with only one kind of languages.

S.Vijayarani and A.Sakila [19] describes performance comparison of OCR Tools. A comparison of OCR tool was done on two factors accuracy and error rate. The performance of eight different types of OCR tools was analyzed. The k-mean clustering algorithm has been used. It was concluded that existing OCR tools produced good result for converting character. However, this study has produced unsatisfactory results for mathematical equations and symbols from the text image and generates unsuitable result.

Anitha Mary et al. [23] describes relative study of different feature extraction techniques for offline Malayalam character recognition. A relative study of Malayalam characters uses four different feature sets which were performed by zonal feature, project histograms, chain code histogram and histogram of oriented gradients. A set of 5 Malayalam vowels were used where 5 consonants were evaluated in feed forward neural network. It was concluded that the best recognition correctness of 94% was obtained by using histogram of oriented gradient feature. This study has the limitation as it works with only one kind of languages.

III. THE PROPOSED METHOD

In this section, the proposed approach and the methodology used to achieve the results is discussed. The proposed framework involves different step to obtain better results. Firstly, the neural network and bias values are defined. After that set of files are loaded which further changes the character to its Unicode where ‘N’ defines the number of character. It further calculates the output. The back error gets propagated in the next step. If the number of characters equal to zero than compute the average error which further involves entry of an output in the form of test image which further analyzes the character of an image.

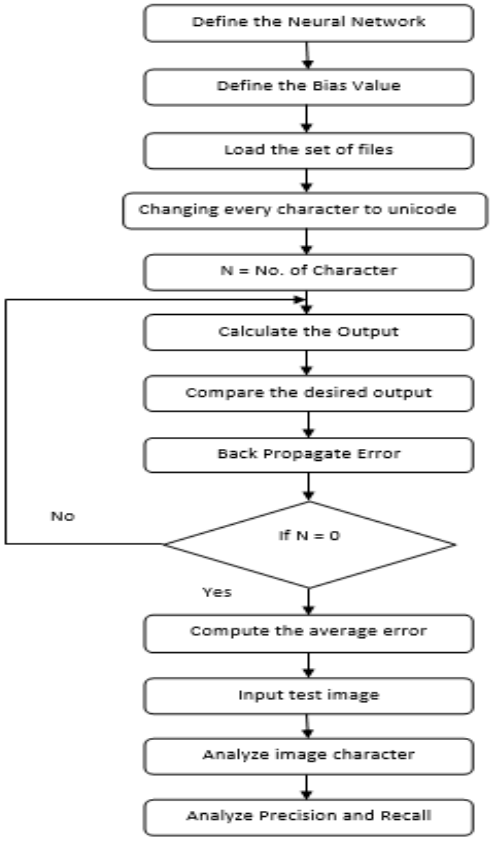

Figure1. Proposed Flowchart

The mechanism of precision and recalling is performed. If the value of 'N' is not equal to zero, then the process again goes back to sixth step and calculates the output and the process is repeated again until it reaches a valid calculation or result.

A. *Proposed Algorithm*

1. Input Image (I) and I_P ← Preprocessor
2. Initialize neural network (Layers (H), Learning rate (n))
3. Randomly Initialize weights $w = \{w_0 \dots \dots \dots w_i\}$
4. Input PixelsD $= \{x_k, y_k\}$; x_k = feature and y_k = Label
5. repeat
6. for all $\{x^{(i)}y^{(i)}\}$ ε D do
7. Compute $y^{(i)}$ according to parameter
8. Compute w and ŋ
9. end for
10. Until achieve parameter use in Tesseract
11. Tesseract with learning model (L) ← test
12. Extract test and analysis accuracy, precision, and recall.

IV. ANALYSIS AND RESULTS

In this section the Tesseract, GOCR and Proposed Method are compared based on their Accuracy, Precision and Recall. The analysis has been done using the brightness, font type, image type and resolution.

A. Parameters

The criterion used for comparison is:

1) *Accuracy:* Accuracy is the parameter used for testing the samples which are correctly classified. A accuracy is used for comparing different approaches, considering the experimental results.
2) *Precision:* Precision is the depiction of errors. It is also called as positive imagine value. Precision tells how many selected items are relevant.
3) *Recall:* Recall is also called as the sensitivity or true positive rate. The recall the highlights the number of relevant items selected.

B. Results

The tools were applied on the brightness feature of image. Three images were taken with brightness of twenty five, fifty and hundred. The results of Tesseract, Gocr and Proposed Method for brightness are given below:

TABLE I

BRIGHTNESS (TWENTY-FIVE, FIFTY & HUNDRED)

Brightness	**Accuracy (tesseract)**	**Accuracy (Gocr)**	**Accuracy (proposed)**	**Precision (tesseract)**	**Precision (Gocr)**	**Precision (proposed)**	**Recall (tesseract)**	**Recall (Gocr)**	**Recall (proposed)**
twenty five	67	57	74	66	52	70	62	50	80
Fifty	68	60	72	64	54	77	65	51	82
Hundred	69	62	76	59	53	72	67	52	83

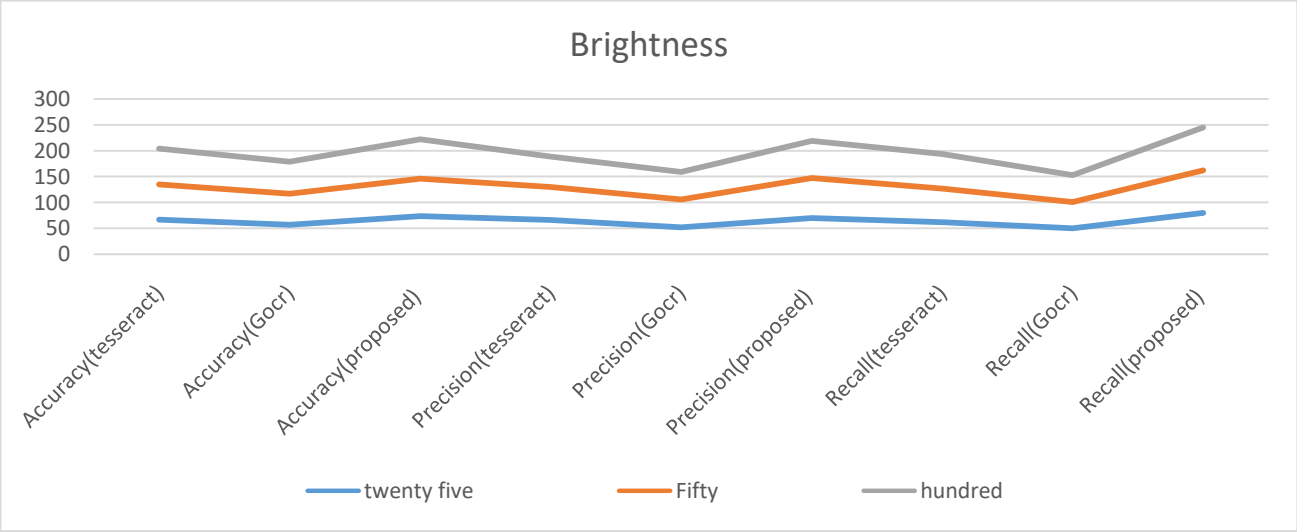

Fig. 2 Brightness (twenty-five, fifty & hundred)

The tools were applied on the font type feature. Three font types were taken named Arial, Roman and Tahoma. The results of Tesseract, GOCR and Proposed Method for font type are given below:

TABLE 2

FONT TYPE (ARIAL, ROMAN, TAHOMA)

Font Type	Accuracy (Tesseract)	Accuracy (Gocr)	Accuracy (proposed)	Precision (Tesseract)	Precision (Gocr)	Precision (proposed)	Recall (Tesseract)	Recall (Gocr)	Recall (proposed)
ARIAL	66	57	74	68	53	78	66	66	77
ROMAN	69	61	72	64	57	77	70	64	80
TAHOMA	71	63	76	62	52	70	71	63	81

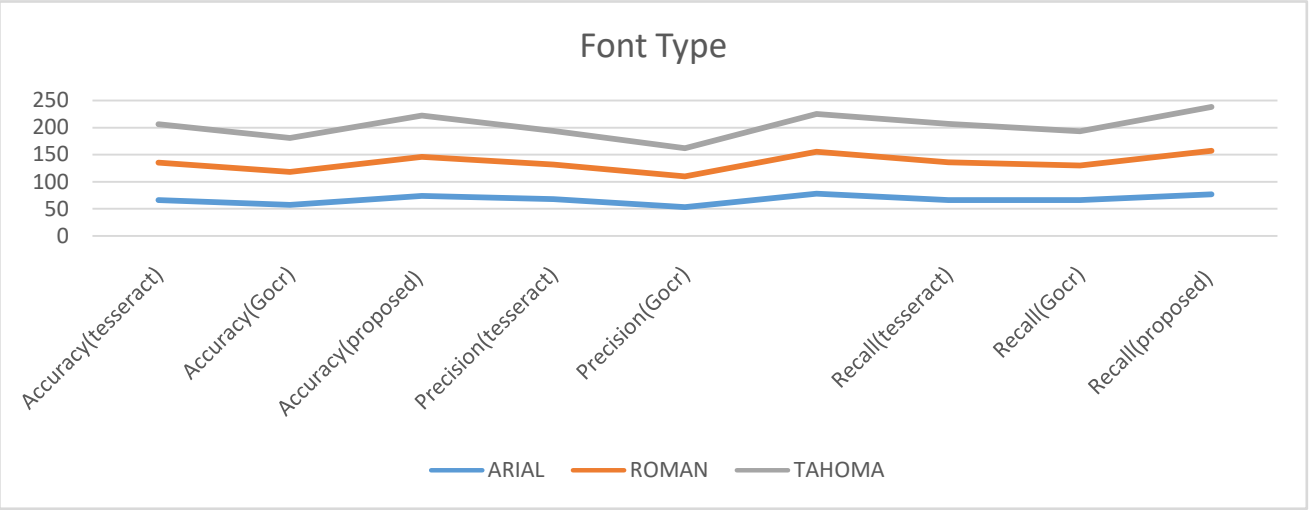

Fig. 3 Font Type (Arial, Roman, Tahoma)

The tools were applied on the image type feature. Three images were taken for each image type that is Color, black and white and noisy. The results of Tesseract, Gocr and Proposed Method for Image type is given below:

TABLE 4

IMAGE TYPE (3 COLOR, 3 BLACK AND WHITE, AND 3 NOISY)

IMAGE TYPE	Accuracy (Tesseract)	Accuracy (Gocr)	Accuracy (proposed)	Precision (Tesseract)	Precision (Gocr)	Precision (proposed)	Recall (Tesseract)	Recall (Gocr)	Recall (proposed)
COLOR	62	56	77	66	60	78	66	46	77
BLACK AND WHITE	61	50	70	69	58	72	62	47	66
NOISY	67	55	71	63	55	69	65	50	69

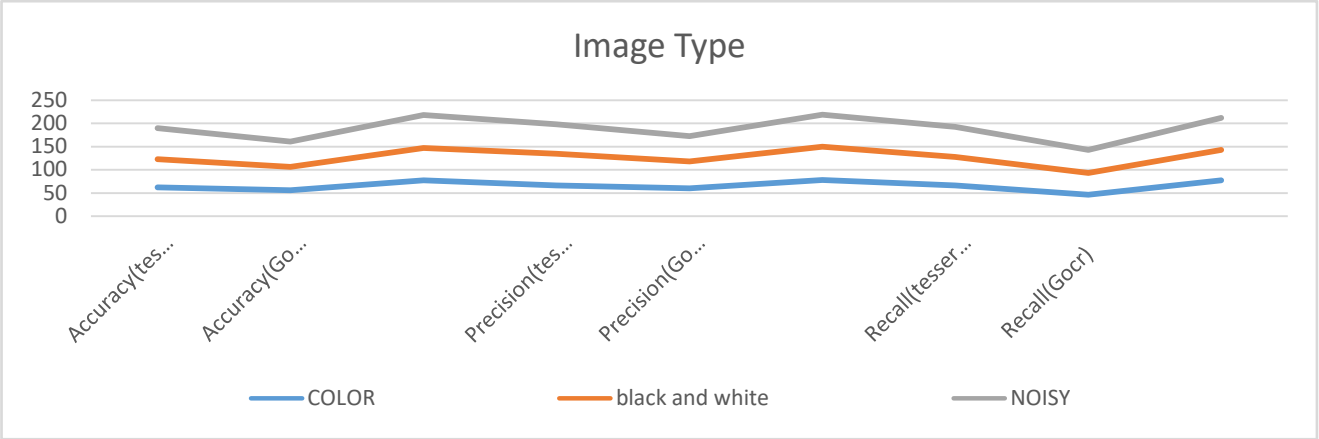

Fig. 4 Image Type (color, black and white, and noisy)

The tools were applied on the resolution feature. Three images were taken for each image type that is Color, black and white and noisy. The results of Tesseract, Gocr and Proposed Method for Resolution are given below:

TABLE 4

RESOLUTION (3 COLOR, 3 BLACK AND WHITE, AND 3 NOISY)

IMAGE TYPE	**Accuracy (tesseract)**	**Accuracy (Gocr)**	**Accuracy (proposed)**	**Precision (tesseract)**	**Precision (Gocr)**	**Precision (proposed)**	**Recall (tesseract)**	**Recall (Gocr)**	**Recall (proposed)**
COLOR	66	56	75	56	45	66	56	60	89
COLOR	62	54	70	60	40	60	69	56	74
COLOR	60	52	72	63	42	56	63	58	76
Black and white	56	56	69	56	48	69	64	52	77
Black and white	63	53	65	63	49	58	66	50	60
Black and white	52	55	63	64	50	69	67	46	67
Noisy	66	58	65	69	52	68	62	47	69
Noisy	67	59	66	66	56	70	66	50	70
Noisy	60	45	69	61	57	71	61	53	66

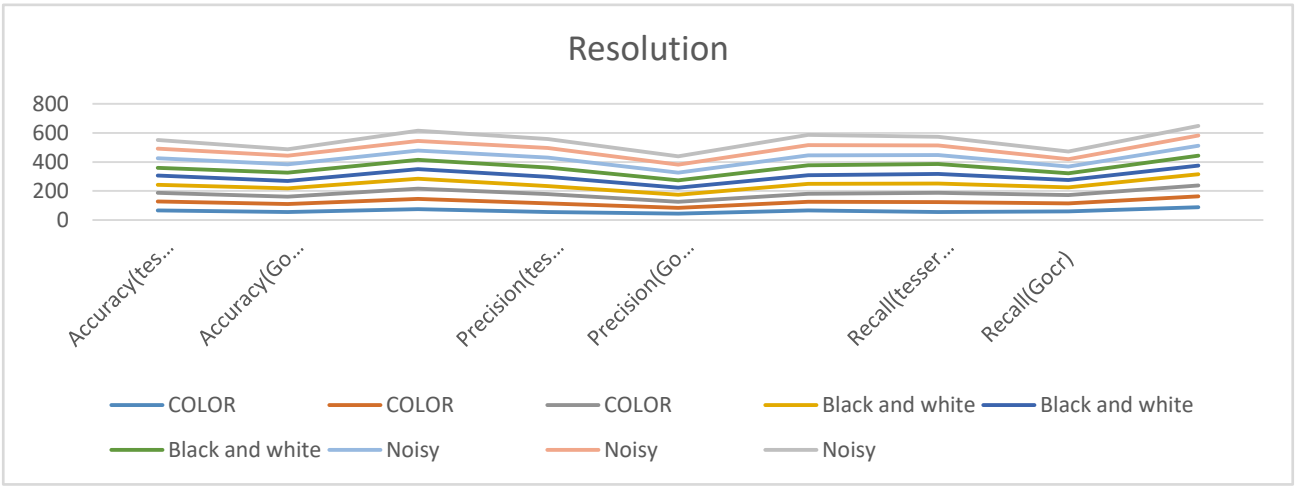

Fig. 5 Resolution (3 color, 3 black and white, and 3 noisy)

V. CONCLUSIONS AND FUTURE WORK

Recognization of the characters from the scanned images is a complex process. To perform the various operations on the data which is present in digital form, record maintenance is necessary. In this research paper a new algorithm has been proposed. This algorithm is based on Artificial Neural Network concepts. Results show that proposed methodology increases the accuracy of Tesseract tool.

In future this work will be enhanced by adding activation function of neural network using deep learning.

REFERENCES

[1] Umal Patel, "An introduction to the process of optical character recognition," International journal of science and Research (IJSR), vol. 2, pp. 155-158, 2013.

[2] G. Vamvakas, Ed., A Complete Optical Character Recognition Methodology for Historical Documents, The Eighth IAPR Workshop on Document Analysis Systems. pp. 525-532, Sep. 2008.

[3] Chirag Patel Ed, " Optical Character Recognition by Open Source OCR Tool Tesseract: A case study. International Journal of computer Application, vol. 55, pp. 50-56, 2012.

[4] Mello Carlos. (2003, March). "A comparative study on OCR tools" Available: https://www.researchgate.net/publication/2555664_A_Comparative_Study_on_OCR_Tools.

[5] Sumit Sharma and Ritik Sharma, "Character recognition using Image Processing. International Journal Of Advancement In Engineering technology, Management and Applied science, vol. 03, pp. 115-122, 2016

[6] Ravina Mithe Supriya Indalkar and Nilam Divekar, "Optical Character Recognition. International Journal of Recent Technology and Engineering (IJRTE) *vol.29*, pp 72-75, 2013.

[7] Kaur Mandeep, "A Recognition system for Handwitten Gurumukhi Characters. International Journal of Engineering Research & Technology, vol. 1, pp. 1-5, 2012.

[8] Enis Bilgin, "Road sign recognition system on Raspberry Pi. In Systems," Applications and Technology Conference (LISAT), IEEE Long Island, pp. 1-5, 2016.

[9] Do Hung Ngoc, Ed. Automatic license plate recognition using mobile device." In Advanced Technologies for Communications (ATC), 2016 International Conference on, vol. 4, pp. 268-271, 2016.

[10] Pal U. and Sarkar Anirban, "Recognition of Printed Urdu scrip. Proceedings of the Seventh International Conference on Document Analysis and Recognition, vol. *2*, pp.1183, 2003.

[11] Shivani Dhiman and A.J. Singh, "Tesseract Vs Gocr A Comparative Study. International Journal of Recent Technology and Engineering (IJRTE), vol. 2, pp. 80-83, 2013.

[12] Ray smith, "An overview of the Tesseract OCR Engine, Proceedings of Document analysis and Recognition, IEEE Ninth International Conferences ICDAR , pp. 1-5, 2007.

[13] Vithlani Purna and Kumbharana C.K. (2015).Comparative Study of Character Recognition Tools. International Journal of Computer Application ,vol.118, *pp. 31-36*, 2015.

[14] Munish, Ed., "K-Nearest Neighbor Based Offline Handwritten Gurumukhi Character Recoognition," Proceeding of the International conference on Image Information Processing, IEEE Ninth Internatonal Conference, 2011

[15] Badla Sahil, "Improving the efficiency of Tesseract OCR engine. Master Theses and Graduate Research, University of San Jose State, vol 12, pp. 269-298, *2014*.

[16] Hasnat Abul, Chowdhury Rahman Muttakinur, "Integrating Bangala script recognition support in Tesseract OCR. Proceedings of the Conference on Language & Technology, pp.108-112, 2009.

[17] Vijayarani S. and Sakila A, "Performance Comparision of OCR Tools," International Journal of Ubi Comp(IJU) vol.6, pp. 19-30, 2011.

[18] Mary Anitha, "A Comparative Study of Different Feature Extraction Techniques Techniques for offline Malayalam Character Recogntion," Proceddings of the International Conference on CIDM, vol. 2, 9-18, 2014.

[19] "GOCR open source OCR engine" July 2013 http:jocr.sourceforge.net. [Accessed: March 2, 2018].

[20] "Tesseract Tool" http://en.wikipedia.org/wiki/FreeOCR#User_interfaces [Accessed: March 2].

[21] "Architecture of tesseract" November 2011, http://www .freewaregenius. [Accessed: March 2, 2018].

[22] "Tesseract Architecture" July 2015 http://www.ocronline.com/ [Accessed: March 2, 2018].