# UNSUPERVISED MERGE OF OPTICAL CHARACTER RECOGNITION RESULTS

*Ioana Monica DICHER*[1]
*Ana-Georgia ȚURCUȘ*[2]
*Eduard-Marius COJOCEA*[3]
*Patricia-Steliana PENARIU*[4]
*Ion BUCUR*[5]
*Marcel PRODAN*[6]
*Eduard STĂNILOIU*[7]

**Abstract:** *This paper explores an innovative Optical Character Recognition (OCR) method that aggregates the results from different methods. Due to the fact that we have to our knowledge the typical characteristics of each OCR approach in any possible situation, a decisive operation can be issued between the outcomes. The proposed method aims to use a voting-based system, apply different preprocessing operations on the input image document, in order to enhance various text characteristics and expects to retrieve the "best text" in the image where it can be "read" more confidently by the OCR engine. The obtained results proved that the proposed approach delivered robust OCR reading in all kinds of processing scenarios, thus enabling the current method to be used, alongside other voting-based techniques in an unsupervised document image processing and information extraction pipeline.*

**Keywords:** *Optical Character Recognition, voting technologies, unsupervised machine reading, Tesseract OCR engine.*

## 1. Introduction

### *A. Previous work*

Optical Character recognition is a Computer Vision technology that enables machines to retrieve humanly-readable text from regular images [1], the accuracy

[1] Engineer, Politehnica University of Bucharest, Bucharest 060042, Romania, ioana_monica.dicher@stud.acs.upb.ro
[2] Engineer, Politehnica University of Bucharest, Bucharest 060042, Romania, ana_georgia.turcus@stud.fils.upb.ro
[3] PhD Student Eng., Research and Development Department, OpenGov Ltd., Bucharest 011054, Romania; marius.cojocea@opengov.ro
[4] PhD Student, Eng., Politehnica University of Bucharest, Bucharest 060042, Romania, patricia.penariu@stud.acs.upb.ro, patriciapenariu@gmail.com
[5] Associate Professor, PhD Eng., Politehnica University of Bucharest, Bucharest 060042, Romania, ion.bucur@cs.pub.ro
[6] PhD Student, Eng., Politehnica University of Bucharest, Bucharest 060042, Romania, marcel.prodan@stud.acs.upb.ro
[7] Teaching Assistant, PhD Student, Eng., Politehnica University of Bucharest, Bucharest 060042, Romania, Eduard.staniloiu@cs.pub.ro

of the method being, in most cases, dependent on text preprocessing and segmentation algorithms [2]. One of the most popular, and widely employed OCR engines, mostly due to its open-source nature, is Tesseract, an engine developed between 84' and 94' by HP, starting from a Ph.D. research project in HP Labs [3]. At that time, commercial OCR engines were primitive and typically failed on anything except for best quality print, which is why HP Labs Bristol decided that this project may become a good product for their company [4]. After more improvements, the OCR engine was the subject of a contest in the 1995 annual test specially dedicated to accuracy in machine reading text [4] and obtained the best results, by far ahead of the performances acquired by the existing machines in that competition [4]. HP ultimately released in 2005 the Tesseract OCR engine as an open-source solution [6].

The Tesseract engine has a pipeline-based architecture in the following serial order: image thresholding [13][14], connected components retrieval from the Boolean thresholded image, blobs creation, basic character recognition, several text aggregation forms in order to build words, text lines [14], text regions or paragraphs and to detect the occurrence of small capitals [7] (Fig. 1.).

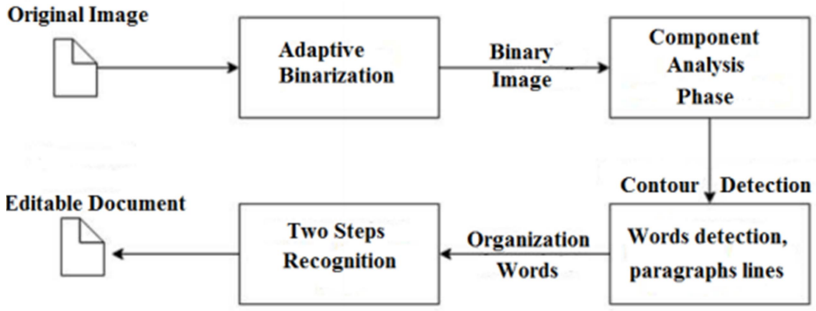


Fig. 1. The architecture of Tesseract OCR system; image taken from [7].

The first code version of Tesseract has improved over time, changes such as conversion to Unicode and retraining, contributing to the increase of its performance. R. Smith presents in [4], a comparison between the Tesseract 2.0 (2007) version and the original HP's version (1995). Table 1 presents a detailed performance comparison onto various types of document sets [4].

| Version | Set | Errors | | %Error rate | | %Change | |
|---|---|---|---|---|---|---|---|
| | | Character | Word | Character | Word | Character | Word |
| HP | Bus | 5959 | 1293 | 1.86 | 4.27 | | |
| | Doe | 36349 | 7042 | 2.48 | 5.13 | | |
| | Mag | 15043 | 3379 | 2.26 | 5.01 | | |
| | News | 6432 | 1502 | 1.31 | 3.06 | | |
| 2.0. | Bus | 6449 | 1295 | 2.02 | 4.28 | 8.22 | 0.15 |

| Version | Set | Errors | | %Error rate | | %Change | |
|---|---|---|---|---|---|---|---|
| | | Character | Word | Character | Word | Character | Word |
| | Doe | 29921 | 6791 | 2.04 | 4.95 | -17.68 | -3.56 |
| | Mag | 14814 | 3133 | 2.22 | 4.64 | -1.52 | -7.28 |
| | News | 7935 | 1284 | 1.61 | 2.62 | 23.36 | -14.51 |
| | Total | 59119 | 12503 | | | -7.31 | -5.39 |

Table 1. Results obtained with the 2007 version of Tesseract versus original 1995 Tesseract; table taken from [4].

### *B. Problem motivation*

Voting-based systems are not something unfamiliar, these being found in different approaches such as [7], [8], [9], [10] and, the most recent, [11]. Similarly to [7], in this paper, the "voting-based" component refers to the following approach: distinct image filters are applied over the input data and partial results with the best confidence are aggregated by the voting process, to optimize the final result.

In general, it can be said that OCR engines are usually very powerful tools but they exhibit several weaknesses. They are very sensitive to the quality of the image presented at the input, especially when dealing with problems like uniform or non-uniform noise, variable illumination across the entire acquired image page, variable contrast determined by an inconsistency in the quality of the paper and/or ink. Physical support degradation over time, inconsistent lines with gaps and cracks and especially thinned characters.

Unfortunately, the aforementioned problems come with a huge cost: in terms of the retroconversion effort of library databases, the errors introduced in the textual information are the most difficult ones to correct, they need the most time and, subsequently, the more allocated resources (both people and money). Layout and hierarchy errors, for example, are much easier to correct, alongside with other common page defects like skew induced by the acquiring image mechanism.

## 2. Proposed method

Our proposed solution in a voting-based one using various filters to trigger various image characteristics. By comparing the words confidence, the most suitable text version is selected as the final output result. Therefore, this method proposes as a solution to compare the confidence of each word and keep the highest value as the final result.

The diversity of the input data, using filters, determines variations in terms of contrast, sharpening, morphological operations like dilation and erosion performed with different kernels on the same image.

The first step is to apply filters on the input image, then send the image to the OCR engine. The outputted text is memorized with its detected confidence and tagged for later processing. We apply this step multiple times on the initial image in order to have more partial results. The final step assumes comparing and combining the partial results from the different OCR runs, as in the case in a proper voting mechanism.

Image binarization [13][14] is employed to increase the contrast between elements and to ensure unambiguous connected components retrieval.

Instead of a more commonly-used per-character OCR confidence, in order to increase the precision, but also have enough granularity for the voting process, word-based confidence (as an average of the individual letters' confidences) is generated and used in candidate selection.

After each run, a comparison based on each word's confidence is performed, and the best confidence word is promoted and selected in the output.

## 3. Performance measurements

Regarding the technologies used, the engine in the implementation is Tesseract 3.02 (Fig. 2) alongside EmguCV 3.0 library for various OpenCV-powered various image manipulations and processing tasks.

The combination of Tesseract and EmguCV is powerful enough to generate and compare a lot of image preprocessing tasks that can be used in order to underline several text characteristics in several image areas.

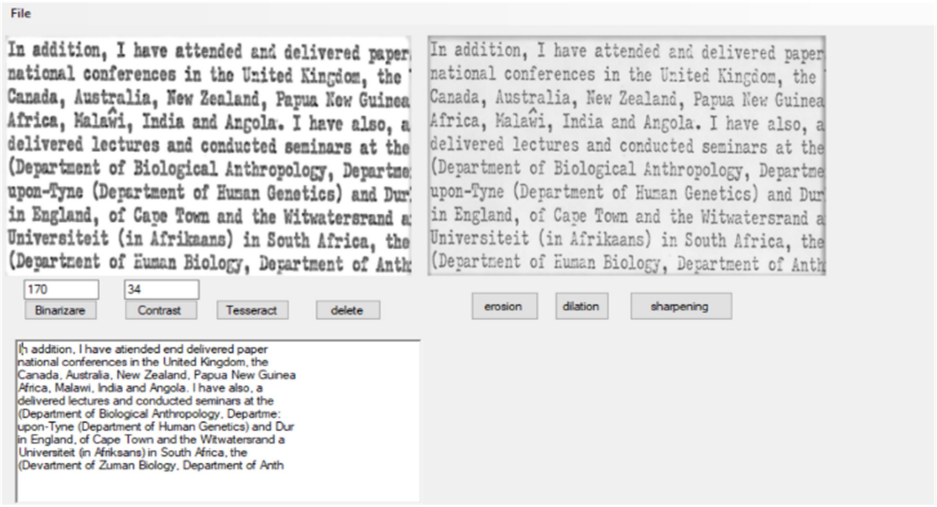


Fig. 2. Tesseract engine version 3.02.

Unfortunately, like any other OCR engine, Tesseract is disturbed by the presence of random noise, improper illumination, and variable contrast across the page, etc. In general, images are distorted locally, meaning that performing several

preprocessing might enable some areas to be confidently "read" by the OCR at the expense of worsening other areas. Capturing the best results from all the runs is the goal of the voting-based technology that was employed in this paper.

There was used a set of test scenarios in order to prove the validity of the voting system. In figure 3 it is illustrated the first scenario, where the low image quality requires applying filters in order to enhance it, and then to be ready for OCR processing.

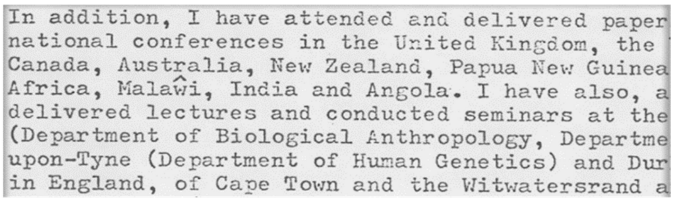


Fig. 3. Scenario 1, "Test 1" image.

All the below figures are generated using the Diffchecker [12], in order to visually present the word-level differences at every image preprocessing step and to illustrate the behavior of the proposed approach at every stage.


Fig. 4. Comparison between image "Test 1" (first row) and the proposed voting-method result (second row)


Fig. 5. Comparison between image "Test 1" (first row) and the image obtained by erosion and dilation (second row).

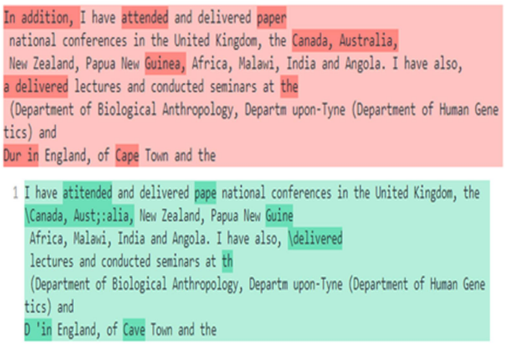


Fig. 6. Comparison between image "Test 1" (first row) and the image obtained by sharpening and erosion (second row)

The results obtained in figure Fig. 7, represents the last sequence applied to the image, namely image sharpening. There are 9 words that were not extracted correctly. In figure 8, 6 words were not extracted correctly by using Tesseract without any filter. The quality of the text was improved by using the partial results from the three sequences and, at the same time, the number of wrong words was reduced from 6 to 4.

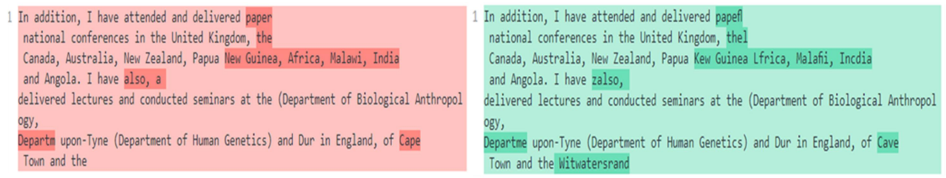


Fig. 7. Comparison between image "Test 1" (first row) and the image obtained by sharpening (second row).

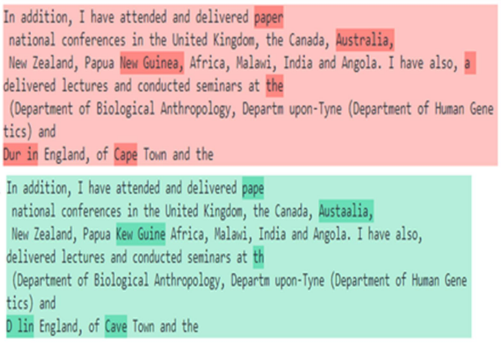


Fig. 8. Comparison between image "Test 1" (first row) and the image obtained by applying Tesseract (second row).

One can observe from the results that every small preprocessing step tends to enable the correct detection of some words, at the expense of the introduced erroneous detection of others, thus the multi-stage OCR process is a slow one but ultimately managing to reduce the number of wrong words.

## 4. Conclusion

The proposed OCR voting-based technology proved to increase the accuracy of text detection in basically any text retrieval scenario. It is robust but at the expense of the extra time needed to perform the various preprocessing tasks and the subsequent OCR runs. The combination of the obtained results into delivering the final solution is a fast operation though.

By employing a solution like this, the major problem of text correction in the large-scale mass-digitization projects is significantly dampened.

The main future development of the presented technology will be the integration in an ensemble of other voting-based methods [7-10] to increase the accuracy of a retroconversion system and to minimize the amount of supervision and correction work that often occurs in the case of image document analysis.

Ultimately it is expected that the current research to ensure a better OCR detection accuracy during the Lib2Life research project [16] aimed at obtaining good quality digital versions from the on-paper documents of several Romanian libraries.

## Acknowledgement

This work was supported by a grant of the Romanian Ministry of Research and Innovation, CCCDI - UEFISCDI, project number PN-III-P1-1.2-PCCDI-2017-0689/„Lib2Life – Revitalizarea bibliotecilor și a patrimoniului cultural prin tehnologii avansate” / "Revitalizing Libraries and Cultural Heritage through Advanced Technologies", within PNCDI III.

## 6. References

[1] A.A. Shinde, D.G.Chougule, *Text Pre-processing and Text Segmentation for OCR*, in International Journal of Computer Science Engineering and Technology, volume 2, issue 1, pp. 810-812, 2012.

[2] C. Patel, A. Patel and D. Patel, *Optical Character Recognition by Open Source OCR Tool Tesseract: A Case Study*, in International Journal of Computer Applications, volume 55, issue 10, pp. 50-56, DOI: 10.5120/8794-2784, October 2012.

[3] R.W. Smith, *The Extraction and Recognition of Text from Multimedia Document Images*, Ph.D. Thesis, University of Bristol, Bristol ITH, United Kingdom, November 1987.

[4] R. Smith, *An Overview of the Tesseract OCR Engine*, in Ninth International Conference on Document Analysis and Recognition (ICDAR 2007), Parana,

Brazil, IEEE, volume 2, pp. 629-633, DOI: 10.1109/ICDAR.2007.4376991, 2007.

[5] S.V. Rice, F.R. Jenkins, T.A. Nartker, *The Fourth Annual Test of OCR Accuracy*, volume 3, Technical Report 95, Information Science Research Institute, University of Nevada, Las Vegas, July 1995.

[6] Tesseract open source, available at http://code.google.com/p/tesseract-ocr, accessed March 20th, 2020.

[7] C.A. Boiangiu, R. Ioanitescu and R.C. Dragomir, *Voting-Based OCR System*, in Journal of Information Systems & Operations Management (JISOM), volume 10, pp. 470-486, 2016.

[8] C. A. Boiangiu, M. Simion, V. Lionte, Z. Mihai, *Voting-Based Image Binarization*, in Journal of Information Systems &Operations Management (JISOM), volume 8, pp. 343-351, 2014.

[9] C. A. Boiangiu, R. Ioanitescu, *Voting-Based Image Segmentation*, in Journal of Information Systems &Operations Management (JISOM), volume 7, issue 2, pp. 211-220, 2013.

[10] C. A. Boiangiu, P. Boglis, G. Simion, R. Ioanitescu, *Voting-Based Layout Analysis*, in Journal of Information Systems & Operations Management (JISOM), volume 8, issue 1, pp. 39-47, 2014.

[11] R. Petrescu, S. Manolache, C.A. Boiangiu, G.V. Vlăsceanu, C. Avatavului, M. Prodan, I. Bucur, *Combining Tesseract and Asprise results to improve OCR text detection accuracy*, in Journal of Information Systems & Operations Management (JISOM), pp. 57-64, 2019.

[12] DiffChecker, available at https://www.diffchecker.com/, accessed March 20th, 2020.

[13] Costin-Anton Boiangiu, Ion Bucur, Andrei Tigora, *The Image Binarization Problem Revisited: Perspectives and Approaches*, The Journal of Information Systems & Operations Management, Vol. 6, No. 2, 2012, pp. 419-427.

[14] Costin-Anton Boiangiu, Andrei Iulian Dvornic, Dan Cristian Cananau, *Binarization for Digitization Projects Using Hybrid Foreground-Reconstruction*, Proceedings of the 5th IEEE International Conference on Intelligent Computer Communication and Processing (ICCP), Cluj-Napoca, August 27-29, 2009, pp.141-144.

[15] Costin Anton Boiangiu, Mihai Cristian Tanase, Radu Ioanitescu, *Handwritten Documents Text Line Segmentation based on Information Energy*, International Journal Of Computers Communications & Control (IJCCC), Vol 9, No 1, 2014, pp. 8-15.

[16] „Lib2Life – Revitalizarea bibliotecilor și a patrimoniului cultural prin tehnologii avansate” / "Revitalizing Libraries and Cultural Heritage through Advanced Technologies", A grant of the Romanian Ministry of Research and Innovation, CCCDI - UEFISCDI, project number PN-III-P1-1.2-PCCDI-2017-0689/, within PNCDI III, available at http://lib2life.ro/, accessed March 20th, 2020.