VOTING-BASED OCR SYSTEM

Costin-Anton Boiangiu [1*]
Radu Ioanitescu [2]
Razvan-Costin Dragomir [3]

ABSTRACT

Current solutions for performing Optical Character Recognition (OCR) in both academic and commercial environments have good recognition capabilities but each one of them has limitations as a consequence of the assumptions made for the defining algorithmic approach. This paper aims to define a new OCR method that combines the results from different algorithms and/or engines. Because we know in advance the specific characteristics of each OCR approach in a given context, a voting algorithm can be applied between their results. The final result of the proposed method is a combination of the different algorithms and exhibits better characteristics than any individual version taken separately. Furthermore, we propose a fully integrated solution containing voting-based approaches for all the preprocessing stages necessary in a complete OCR solution: image binarization, image segmentation, and layout analyze.

KEYWORDS: *OCR, voting, image binarization, image segmentation, layout analysis, document image analysis system, image understanding.*

1. INTRODUCTION

Optical Character Recognition (OCR) is the process that converts a digital image into editable text. An image containing text, after being processed using OCR technology will result in a text document formatted accordingly to the styles presented in the original document.

In general, any voting system starts with a basic or reference element for which certain parameters are varied, thus introducing more components with different properties, each having the potential to improve the final result. To be able to perform the voting process, there must be a quantifiable term indicating how much a given choice will improve the end result. For OCR, this quantifiable term is the detected accuracy percentage.

The “voting-based” component refers to two approaches: the first one, in which the input data is processed using different image filters, and the second one, in which two or more OCR engines are used on the same data input. The voting process chooses the partial results with the highest accuracy percentage, combining them to improve the final result. The objective of this paper is to design, implement and test a voting system based on

[1*] corresponding author, Professor PhD Eng., ”Politehnica” University of Bucharest, 060042 Bucharest, Romania, costin.boiangiu@cs.pub.ro
[2] Engineer, ”Politehnica” University of Bucharest, 060042 Bucharest, Romania, ioanitescuradu@gmail.com
[3] Engineer, ”Politehnica” University of Bucharest, 060042 Bucharest, Romania, razvan.drc@gmail.com

varying input data and merging the results according to internal heuristics, guided by the percentage of accuracy and verifying different permutations with the goal to improve the final recognition quality.

This paper presents part of the work completed in the license thesis of the third author of the paper at the faculty of Automatics and Computers from the "Politehnica" University of Bucharest [24].

2. RELATED WORK

Tesseract is a popular OCR engine [1, 2, 3] and one of the first to obtain good recognition accuracy results [4]. Its pipeline-based architecture (presented in Figure 1). It consists of the following sequential steps: preprocessing providing a binary threshold, determining the connected components and connections between them (also storing them in objects called blobs), character recognition and character aggregation to form words, lines, paragraphs and the finally solving the problem of detecting small capitals.

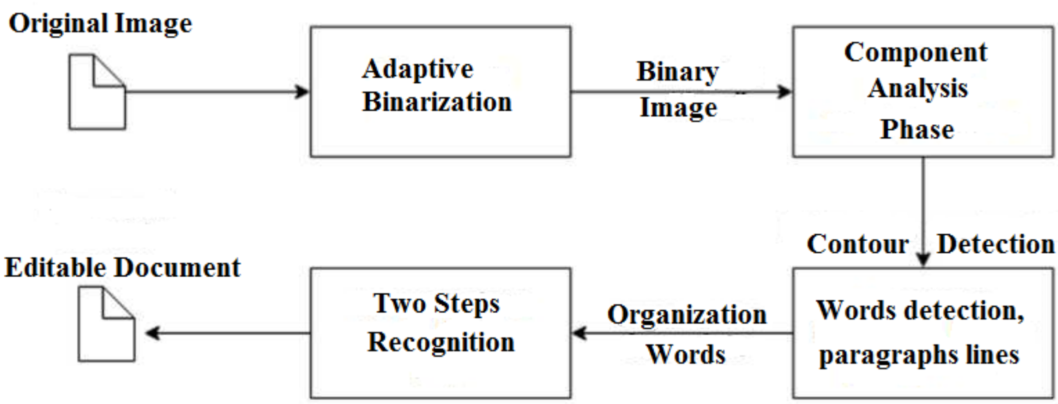


Figure 1. Tesseract OCR Architecture [3]

Voting-Based Image Binarization

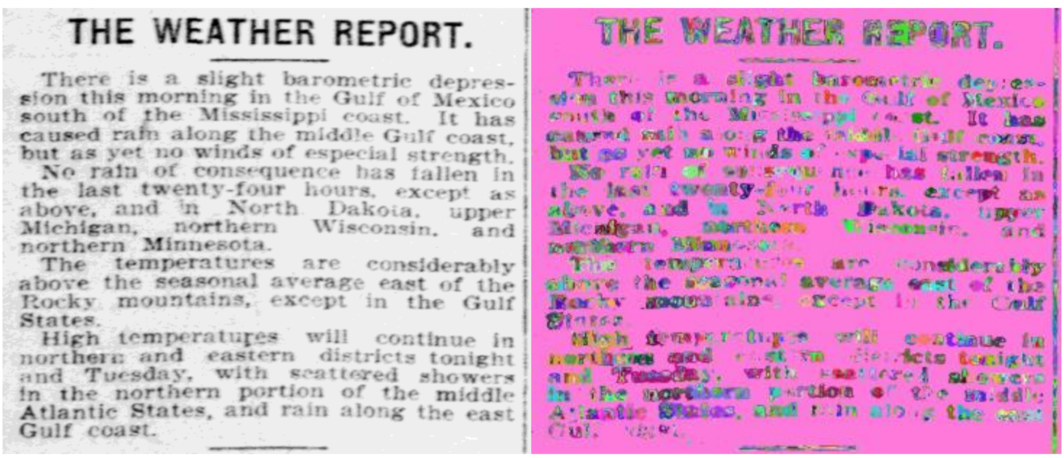
Digital images are composed of pixels, however the range of colors that can be displayed depends on the number of bits used to represent each such pixel (a term called BPP or Bits-Per-Pixel). For example, a binary image having a BPP of one and a single component means that the image will be represented using only two colors: black and white one color for each possible value of the binary representation. The thresholding operation is the processing stage that takes as input an image having a different representation and converts it to a black and white image and based on determining a computer threshold, hence the name [23].

The thresholding step is essential for an OCR engine because the analyzed picture becomes easier to process and the background noise is largely reduced being lower than the threshold limit and thus removed. Some of the best known approaches are the following: the Otsu method for selecting a fixed threshold [5, 6] or a more complex method based on a bimodal histogram analyze [8] or even using voting-based binarization process [7] (Figure 2).

Figure 2. Example of text separation using voting-based image binarization: original image (left), binarized image (right). Image taken from [7]

Voting-Based Image Segmentation

Although thresholding simplifies the input image transforming it into black and white, it cannot identify the elements of an image. Segmentation is the process of identifying the objects in an image based on certain properties (pixel color, intensity, texture) [11]. Typically, segmentation creates a mask image consisting of input pixels belonging to a zone of interest (an object image) of a certain color and/or property. Because a normal image that would be processed through OCR can contain text, graphics and complex layouts, the goal of segmentation is the identification of the areas of interest as well as the evaluation of their type. This image segmentation step is particularly useful in the detection of lines or other layout separators.



THE WEATHER REPORT.

There is a slight barometric depression this morning in the Gulf of Mexico south of the Mississippi coast. It has caused rain along the middle Gulf coast, but as yet no winds of especial strength.

No rain of consequence has fallen in the last twenty-four hours, except as above, and in North Dakota, upper Michigan, northern Wisconsin, and northern Minnesota.

The temperatures are considerably above the seasonal average east of the Rocky mountains, except in the Gulf States.

High temperatures will continue in northern and eastern districts tonight and Tuesday, with scattered showers in the northern portion of the middle Atlantic States, and rain along the east Gulf coast.

Figure 3. Example of text (as object) voting-based image segmentation in an old, variable-contrast and noisy document: original image (left), segmented image (right). Image taken from [13]

Any method of segmentation must meet the following criteria: the recombination of all regions (segments) must reconstruct the original image (i.e. segmentation must be complete) and the regions must be disjointed to avoid duplication and different from each other in the sense that each pixel region only groups based on fixed conditions. Segmentation may be performed using multiple approaches like histogram analysis [11], region growing [12], watershed [14] or even a voting-based segmentation [13] (Figure 3).

Voting-Based Document Image Layout Analysis

The next mandatory processing step in an automatic document image analysis system involves determining the document's logical structure. For example, newspapers are documents that have complex layouts. Thus, the classification process takes into account the hierarchical organization such as the fact that the text blocks will contain objects paragraph types that are composed of lines of text and these in turn are composed of words. Among the algorithms that perform this type of document analysis are: XY-cut [15], Whitespace cover [17] or even a voting-based layout analyze approach [16] (Figure 4).

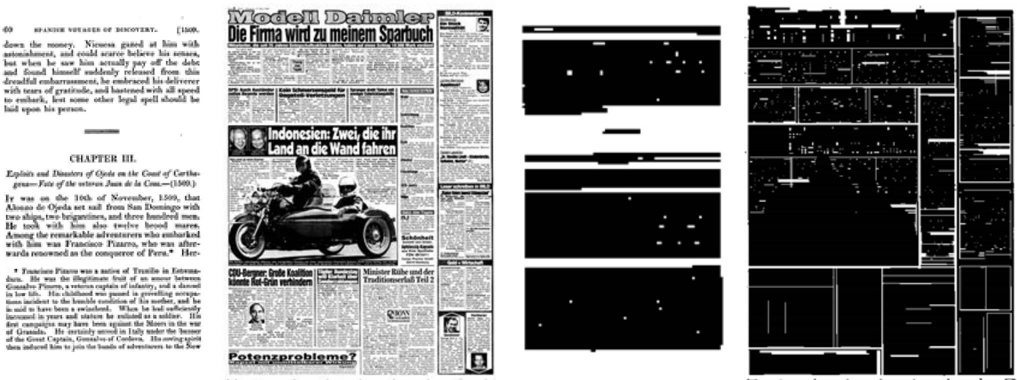


Figure 4. The result of document layout analyze: simple and complex document original images (left), voting-based layout analyze results (right). Image taken from [16]

The next processing step is optional and involves the usage of processing filters with the goal of reducing the background noise and improving the OCR engine output. Through a process called dilation [18], the elements of the resulting image are thinned, making the background image bolder and sometimes causing darker elements to be separated. Erosion [18] is the reverse process by which elements of the resulting image contours are thicker and the background image is thinned and sometimes brighter elements are combined.

3. VOTING-BASED OCR: THE PROPOSED METHOD

During the last two decades, the OCR process and its stages have been consistently developed due to rapid technological changes and the need to store information in electronic formats. The methods to improve the accuracy of the results may be divided into three categories, depending on the time with respect to the text recognition step, namely: preprocessing, processing and post-processing.

Typically, processing is encapsulated in the OCR engine and its own algorithms and methods. OCR engines are based on different approaches that yield different results, a characteristic that can be observed by comparing the results obtained from processing the same set of input data by different technologies.

In the preprocessing category, which aims to improve data entry by bringing them to a suitable quality, we will briefly present a number of approaches based on implementing a system of voting as a method of studying the results of these approaches, showing significant processed image quality improvement.

We aim to propose a fully-functional OCR that performs in a voting-based manner all the necessary steps in a document image analysis system up to the OCR process. The solutions presented by the first author of this paper in [7, 13, 16] tried to combine classical algorithms to improve the image binarization, image segmentation and layout analysis steps.

In the study [19] a detailed comparison between Tesseract v. 3.0.1 and ABBYY FineReader version 10 has been made. Test data represented historical documents dated before 1850 and printed in Polish. There are a total of 186 pages, of which 38 were used for the training and the remaining 148 pages have been used as recognition data. Moreover, the pages were processed both in a form having the noise removed as well as in their original one, were built from only Antiqua or Gothic characters (fonts) and the measured accuracy was calculated both at the level of characters and words.

Table 1. ABBYY-Tesseract OCR engines comparison results

Test number	Character type	Image type	Number of pages (training)	Character accuracy (percentage)		Word accuracy (percentage)	
				ABBYY	Tesseract	ABBYY	Tesseract
Test_Antiqua	Antiqua	processed	28	**86,97**	84,81	**65,43**	54,99
Test_Antiqua	Antiqua	original	28	83,08	69,38	60,42	42,52
Test_Gothic	Gothic	processed	4	73,98	**80,64**	36,98	**45,62**
Test_Gothic	Gothic	original	4	52,79	70,99	20,74	41,16

For "Test_Antiqua" is observed that ABBYY has a higher degree of accuracy in both cases, but for "Test_Gothic" Tesseract seems to obtain an accuracy better than ABBYY, but, perhaps, not good enough for most document preservation purposes. After the study, it was observed that Tesseract returns generally better outcomes for documents containing gothic characters, both in terms of character-level and word-level accuracy. As shown in Table 1, the results are better for processed images, which confirms that the preprocessing step is necessary to improve the accuracy, especially in the processing of historical documents.

In conclusion, given a set of data, in order to return a relatively good solution for different preprocessing techniques or after applying several OCR technologies, a voting system can be created to choose the OCR results with a maximum percentage of accuracy. The findings above show that implementing such a voting system is a viable idea.

Implementing a voting system is not a new idea, but one that has been used in a recent approach [7, 13, 16] to achieve optimum and efficient operation of thresholding segmentation and analysis of the document. The idea used in these papers consists of starting from a set of classical algorithms which are known to address a particular behavioral problem.

The proposed voting-based OCR solution involves the usage/combination of specific filters on the input image, resulting in a number of output data. This operation seeks to obtain a data set that is composed of different blocks of text that can be properly recognized relatively by the OCR process. By comparing the percentage of accuracy between similar text blocks, the best option is retained for inclusion in the final result.

A simple solution is to compare the percentage of accuracy of each partial result and promote the highest value as the final result, but in practice this approach does not lead to optimal results. The problem is that only certain regions of the input image may yield better results after applying a particular filter, in other words, each such data entry variation will cause the OCR engine to return some text blocks that are complete and correct, but omit a few sentences or words because that area of the input data is inappropriate for text recognition. As a result, it is possible that none of the preprocessing candidates contain the full text, even if the percentage of accuracy is good, and thus the final result will also not contain that part of the text.

Providing a smaller granularity voting system that will consider small regions of the input image and achieve accuracy rates based on their similarity is a better solution and a viable real-world system. The proposed solution consists of the realization of a voting system based on variations in the input data (preprocessing of the image), the application of the recognition process on each such variation, obtaining a number of partial results and combining them in a final version based on the highest accuracy percentage (using a process of voting between partial results).

The variation of the input data is represented by the use of filters as well as image dilation and erosion morphological operators of various kernel shapes and sizes. It consists of a thresholding followed by correcting the text angle, reducing artifacts that may appear in the image.

To make it possible to have a viable voting system, a common base is required to be able to choose between areas of the input image on which voting will be performed. Please note that none of the obtained results from the preprocessing stages can be used as a reference text and thus accurate comparisons cannot be made.

The common base will be computed by determining a black mask input image which will be calculated using the coordinates of the text blocks. To compute these masks we use a thresholding method and Hough probabilistic transformation [21].

The next step consists of extracting the coordinates for each region. The coordinates are sent to the OCR engine to enable it to process only the desired region with corresponding filters. The resulting text along with the percentage of accuracy returned is stored as a text-value pair and is used in the next step.

The last step is the comparison and combination of the partial results, also known as voting. At this stage the algorithm iterates through all the possible results, choosing one that has a higher percentage of accuracy and positioning it in the final result based on previously determined masks.

The general application architecture is represented in the Figure 5.

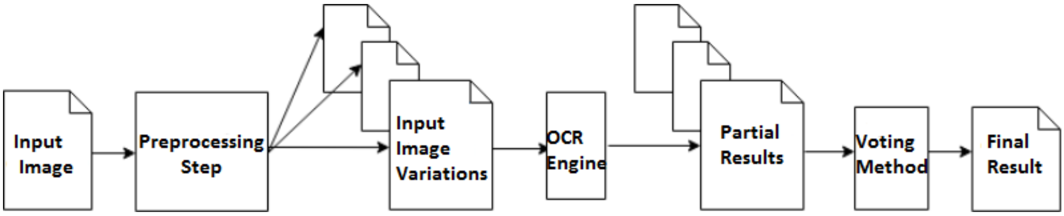


Figure 5. The generic OCR voting model

The OCR voting starts by selecting and loading the input image, followed by automatic grayscale conversion. After this, a preprocessing stage is applied, which attempts to reduce noise and apply different filters to achieve variations between the inputs controlling the voting process.

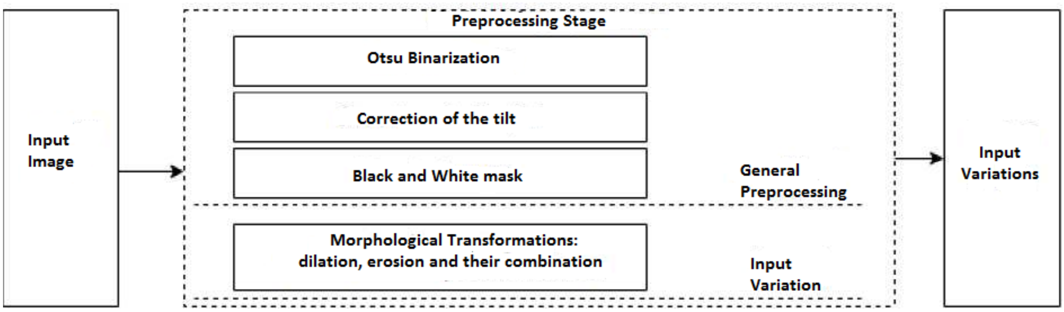


Figure 6. The preprocessing module

The preprocessing module (Figure 6) is the first step in the voting system. In order to obtain a suitable quality of data input and prepare it for the recognition, the applied filters are divided into two categories.

On one hand, a category of filters is represented by the general ones. The application of these filters is a preliminary stage to vary the input data and aims to clean the input and obtain a reference model of the page on which operators are applied to build candidates to perform comparisons between percentages of accuracy.

This includes thresholding, a simple algorithm for detection and correction of the skew of the page and a step of building a black and white mask of the input image that represents a virtual template onto which the voting will be performed.

The chosen thresholding approach is Otsu [5], a global method, because it results in fast and robust computation of a close-to-optimal global threshold value. In a written document, usually the grouped objects are similar also in terms of color intensity (usually foreground representing text is darker than the background) which is the principle underlying the chosen method.

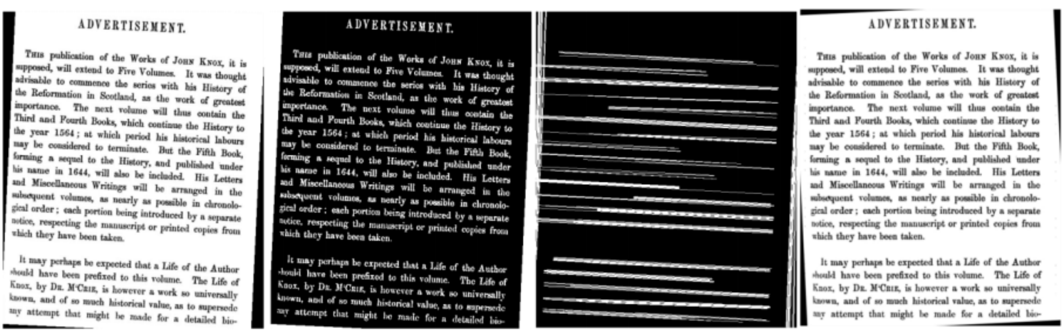
During the next processing step, the image obtained is binarized, ensuring a high contrast between the groups of elements in the scene. Because the image document may be slightly skewed due to its positioning in the scanning device, it is vital to follow the step of skew detection and correction ("deskew"). Moreover, skew correction not only facilitates the process of recognition but also improves the proposed solution's results.

To calculate the angle of inclination of the page, the lines of text from the document have to be detected. Detection is made using the probabilistic Hough transformation because it is an optimum method for recognizing collinear points in an image. Moreover, this method returns extreme collinear coordinates of the lines (the coordinates of the start or end) which helps to calculate more precisely the angle of inclination.

The algorithm is simple and it must follow the steps bellow:

- the thresholding method using Otsu;
- applying the probabilistic Hough transformation, which detects extreme coordinates of the lines of text;
- determining the inclination angle of the entire image, calculating a weighted average of the length for each line;
- the last step consists of the image rotation to compensate for the calculated angle.

Figure 7 illustrates an example of the skew detection algorithm's steps. It shows how the document in the original image has a slight inclination to the right (this is possibly due to incorrect positioning inside the scanning device). The second image is representing the step in the skew detection algorithm named color inversion (step necessary to use Hough transformation [26]). The third image is the result of the application of probabilistic Hough transform (note how a sufficient number of lines is detected to calculate the inclination angle of the page with a good enough accuracy). The last picture shows the skew correction determined by the angle and how the text lines [27] are, in the end, horizontally aligned.


ADVERTISEMENT.

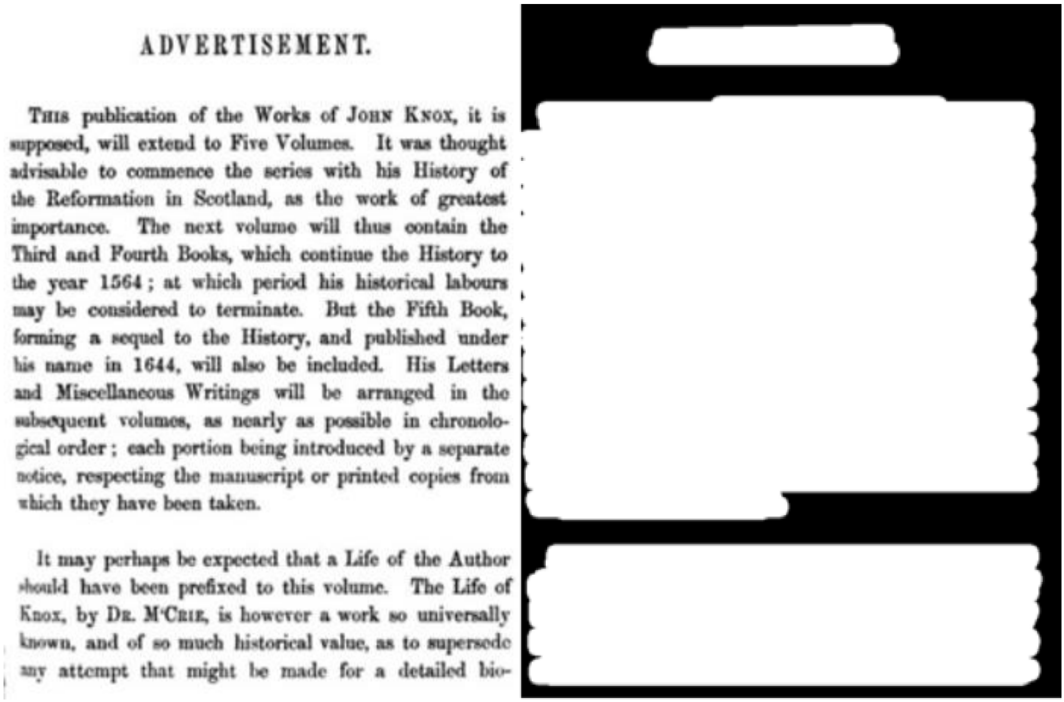
This publication of the Works of John Knox, it is supposed, will extend to Five Volumes. It was thought advisable to commence the series with his History of the Reformation in Scotland, as the work of greatest importance. The next volume will thus contain the Third and Fourth Books, which continue the History to the year 1564; at which period his historical labours may be considered to terminate. But the Fifth Book, forming a sequel to the History, and published under his name in 1644, will also be included. His Letters and Miscellaneous Writings will be arranged in the subsequent volumes, as nearly as possible in chronological order; each portion being introduced by a separate notice, respecting the manuscript or printed copies from which they have been taken.

It may perhaps be expected that a Life of the Author should have been prefixed to this volume. The Life of Knox, by Dr. M'Crie, is however a work so universally known, and of so much historical value, as to supersede any attempt that might be made for a detailed bio-

Figure 7. An example of the skew detection algorithm's steps

Because the image resolution does not change, the varying entry or the coordinates of the blocks do not change. Thus, it creates the common base necessary for the voting system allowing for both the opportunity to vote in areas that relate to the same text and the granularity required to create a viable voting system.

The template building process is also based on the probabilistic Hough transform. The first stage detects text lines based on extreme collinear points. After the detection process finishes, the coordinate lines are drawn between extreme points in order to obtain a model of the original, followed by morphological transformations to thicken previously drawn lines until they unite. The last step consists in extracting the coordinates of regions containing blocks of text.



ADVERTISEMENT.

THIS publication of the Works of JOHN KNOX, it is supposed, will extend to Five Volumes. It was thought advisable to commence the series with his History of the Reformation in Scotland, as the work of greatest importance. The next volume will thus contain the Third and Fourth Books, which continue the History to the year 1564; at which period his historical labours may be considered to terminate. But the Fifth Book, forming a sequel to the History, and published under his name in 1644, will also be included. His Letters and Miscellaneous Writings will be arranged in the subsequent volumes, as nearly as possible in chronological order; each portion being introduced by a separate notice, respecting the manuscript or printed copies from which they have been taken.

It may perhaps be expected that a Life of the Author should have been prefixed to this volume. The Life of Knox, by DR. M'CRIE, is however a work so universally known, and of so much historical value, as to supersede any attempt that might be made for a detailed bio-

Figure 8. An example of a black and white (region-separation) mask: original image (left) and the resulted mask (right)

Figure 8 illustrates a black and white (region-separation) mask obtained by using the aforementioned processing method. We can observe how the document was divided into three regions. The number of regions varies from image to image and depends on the spacing between rows. This number of regions can be controlled by a number of parameters applied to the probabilistic Hough transform and/or dilation operator. For example, we can consider each text-line of the image as a block of text, while the voting will be performed between those lines. However, the results of the OCR engine are better if we consider larger blocks of text due to the mechanism of feedback that is implemented in the two-step recognition process.

The second type of preprocessing filters is represented by morphological operators of dilation or erosion as well as a combination of them (an operator of erosion followed by a dilation, or the reverse process). These morphological changes will achieve varying input data, so the process will involve passing the input image through the thresholding step, correcting the slope and building a template followed by the application of a set of distinct preprocessing.

The structural element may acquire various forms including the most popular ones: star, ellipse and rectangle, all having various sizes. The size of the structural element will not exceed a predefined threshold as the application of such a filter will change too much the text features thus becoming unusable for the OCR process.

The proposed solution is independent of any particular OCR engine that is treated as a black box system. The voting system is considered an extra layer of abstraction that tries to improve the final result, regardless of the OCR engine used. To enhance the abstraction of the communication with the OCR engine, we consider a common interface (wrapper) that connects the preprocessing module and the voting module. The common interface may include implementations for multi-engine OCR usage in which case the interface for interconnection must be carefully designed to ensure all the required basic functionalities are met. Thus, irrespective of the OCR engine used the input interface receives a number of preprocessed image data (based on morphological transformations and combinations previously established) and returns the same number of results multiplied by the number of text blocks determined in the black and white mask.

Also, the voting system is based on the percentage of accuracy of output text, so it is necessary that the OCR engine is able to calculate such a rate for each region in the template. Typically, OCR engines calculate this percentage at character level and by applying an arithmetic average can precisely calculate the accuracy detection percentage of these areas.

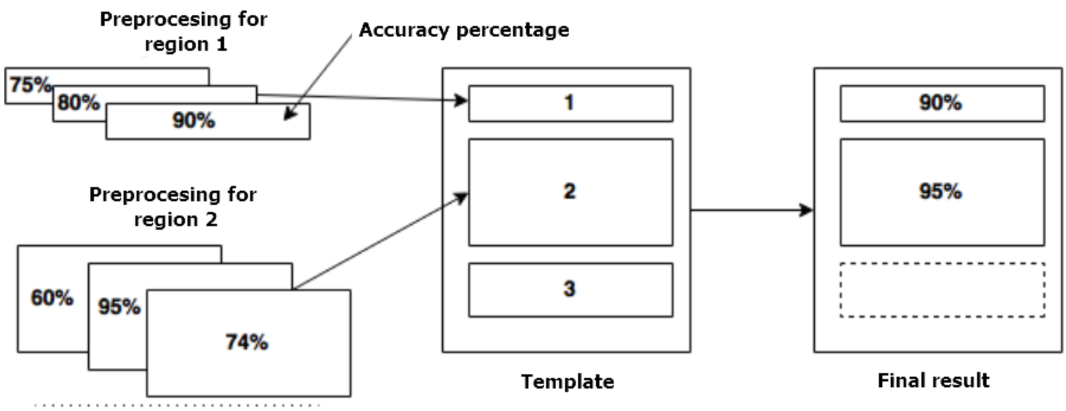


Figure 9. The voting process

The operation performed by this module is very simple. It receives multiple sets of data from OCR engine and combines them to obtain the final result (Figure 9). The received data is stored in an associative structure; each entry is represented by a block of text index previously determined in the stage of building the template preprocessing module. Also, at the same time, for each preprocessing result the detected percentage of accuracy is stored.

The voting process means the processing of each such aforementioned region (thus completing every index) and the election of the most suitable partial results that will be included in the final result.

4. CONCLUSIONS

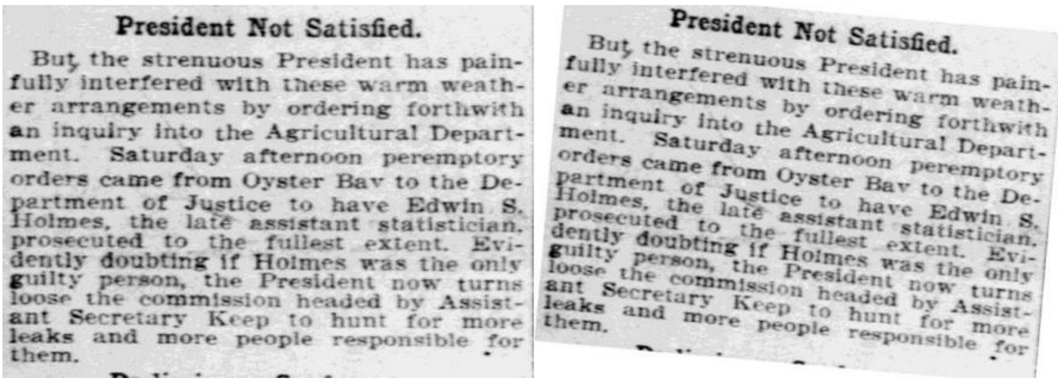
The Validator Application and Test Scenarios

Some libraries used in the implementation of the image processing modules and our proposal validator application are Tesseract 3.02 and Leptonica 1.68. Tesseract offers an API for the C++ programming language to develop applications and due to the fact that the voting system requires constant variations applied on the input, by switching between different image processing filters and morphological operators, the OpenCV 3.0 library was chosen to complete the set of libraries and technologies needed in our implementation.

To illustrate the validity of the voting system, a set of test scenarios were carried. The tests include, among others, the sample pages represented in Figure 10 and Figure 13 that contain problems like noise, uneven background and illumination, various artifacts and/or an angle of inclination, allowing to exemplify a comparison between a simple execution of Tesseract and the proposed solution.

The Proposed Approach Results

Scenario 1 (skew-free and skewed input image document)



President Not Satisfied.

But, the strenuous President has painfully interfered with these warm weather arrangements by ordering forthwith an inquiry into the Agricultural Department. Saturday afternoon peremptory orders came from Oyster Bay to the Department of Justice to have Edwin S. Holmes, the late assistant statistician, prosecuted to the fullest extent. Evidently doubting if Holmes was the only guilty person, the President now turns loose the commission headed by Assistant Secretary Keep to hunt for more leaks and more people responsible for them.

President Not Satisfied.

But, the strenuous President has painfully interfered with these warm weather arrangements by ordering forthwith an inquiry into the Agricultural Department. Saturday afternoon peremptory orders came from Oyster Bay to the Department of Justice to have Edwin S. Holmes, the late assistant statistician, prosecuted to the fullest extent. Evidently doubting if Holmes was the only guilty person, the President now turns loose the commission headed by Assistant Secretary Keep to hunt for more leaks and more people responsible for them.

Figure 10. Test "Test_A" and “Test_B” input images

In Figure 10 both images have a resolution of 300 dpi, the latter being obtained from the first by using a rotation of 5 degrees to the right. Picture "Test_A" was extracted from the first page of "The Washington Times", 17 July 1905. The sample image quality is not suitable for direct OCR processing, since bold characters are often wrongfully joined together. Also, the background is having a gray tone that approaches the text color, making it difficult or almost impossible to recognize certain characters printed with faded colors.

The normal Tesseract execution (Figure 11) was compared with the proposed solution (Figure 12) and the DiffChecker tool [25] was used for the purpose of accurate identification of the differences between the two approaches. The text lines were vertically-aligned in both figures for the left and right text blocks, so that the correspondence between the same information is as easy as possible to be followed visually.

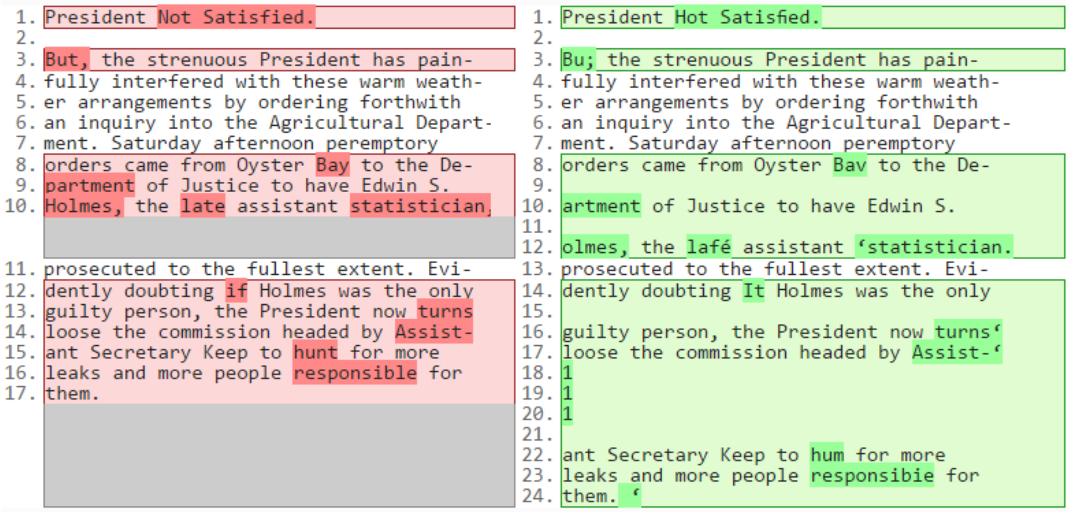


Figure 11. Comparison between the input "Test_A" reference text (left) and the basic Tesseract execution result (right)

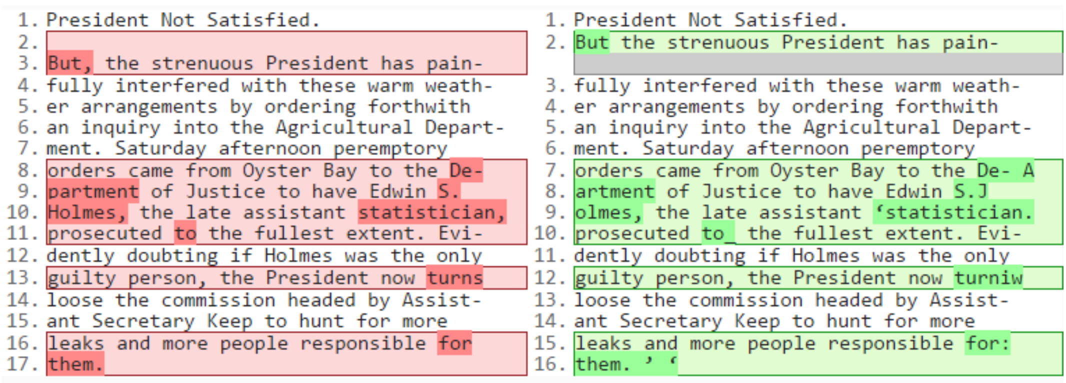


Figure 12. Comparison between the input "Test_A" reference text (left) and the proposed voting-method result (right)

The set of preprocessing steps used in the voting process contains a complex series of morphological operators applied as a succession of two transforms, each such transformation being applied with different sizes and shapes of structural elements. The two-transform sequence results are also measured independently on the two detected regions of the test document. Table 2 shows in a detailed manner, the filters used with the percentage of accuracy for each of the two regions.

Table 2. The filters used in the proposed solution and the obtained accuracy

#	Transform 1			Transform 2			Accuracy %	
	Applied operator	**Structural element size (pixels)**	**Structural element shape**	**Applied operator**	**Structural element size (pixels)**	**Structural element shape**	**Region 1**	**Region 2**
1	dilation	1 x 1	plus	erosion	1 x 1	square	75.65	75.44
2	erosion	1 x 1	square	dilation	1 x 1	ellipse	81.09	82.83
3	dilation	1 x 1	plus	erosion	1 x 1	ellipse	81.09	**84.27**
4	erosion	1 x 1	ellipse	dilation	1 x 1	plus	81.09	73.13
5	dilation	3 x 3	plus	erosion	3 x 3	square	75.78	76.88
6	erosion	3 x 3	square	dilation	3 x 3	ellipse	75.78	76.36
7	dilation	3 x 3	plus	erosion	3 x 3	ellipse	81.54	79.43
8	erosion	3 x 3	ellipse	dilation	3 x 3	plus	**85.03**	82.68

Table 3. Comparison between the proposed voting-based method and Tesseract 3.02 in test scenario 1

Test image	Accuracy of characters (%)	Accuracy of words (%)	Technology used
Test_A	97.01	87.64	Tesseract 3.02
Test_B	90.26	75.28	
Test_A	**98.13**	**91.01**	**Proposed method**
Test_B	**97.51**	**88.74**	

Table 3 shows a comparison between the standard Tesseract 3.02 results and the voting-based method proposed in this paper for the test scenario 1. The accuracy percentages are calculated by manually counting the wrong characters/words and computing the percentage based on the total number of characters (483) or the total number of words (89) found in the reference text.

Scenario 2 (noise-free and variable text quality and contrast)

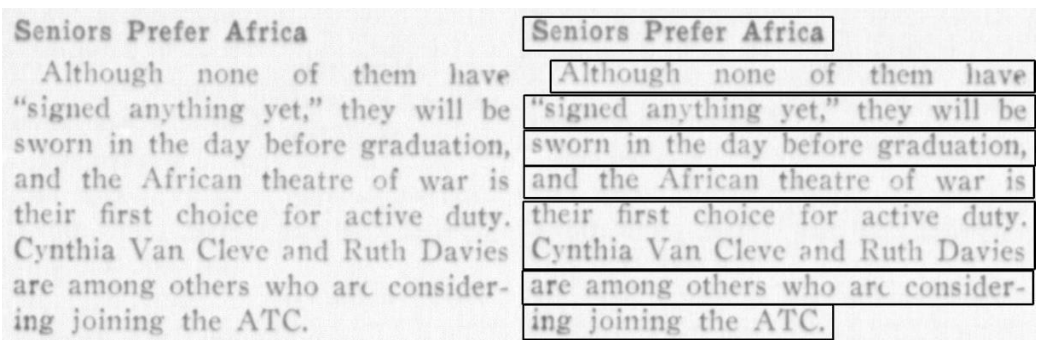

Seniors Prefer Africa
Although none of them have "signed anything yet," they will be sworn in the day before graduation, and the African theatre of war is their first choice for active duty. Cynthia Van Cleve and Ruth Davies are among others who arc consider-ing joining the ATC.

Figure 13. The input "Test_C" original image (left) and the current proposal detected template regions (right)

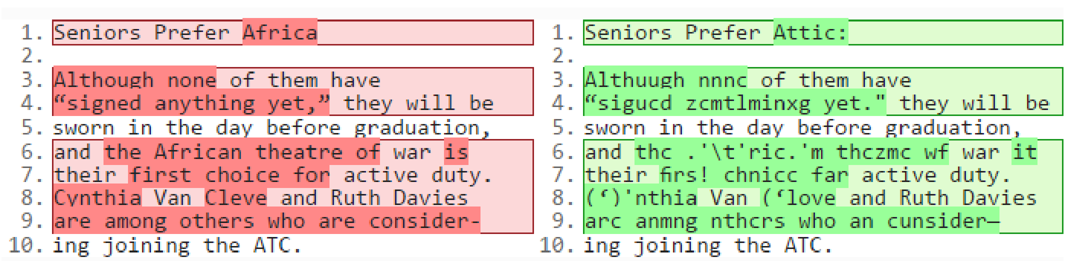
In Figure 13 the original image has also a resolution of 300 dpi being extracted from the page 4 of „The Vassar Chronicle", 24 March 1944.

The "most successful" morphological operator succession and kernel type and size for every region, is represented in table 4 and the result comparison for standard Tesseract and the proposed voting-based approach is presented in Figures 14 and 15.

Finally, the comparison for the second test scenario in terms of efficiency of the proposed voting-based approach in comparison to the standard Tesseract OCR engine is presented in Table 4.

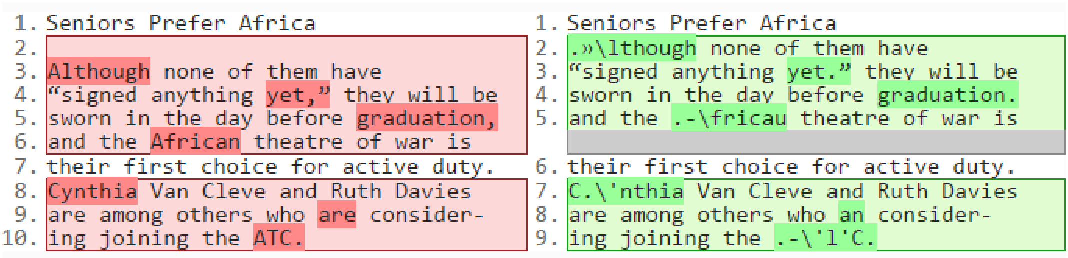
Table 4. The proposed solution's operators and kernels used, and the resulted accuracies

#	**Applied operator (in order)**	**Region affected by the applied operator**	**Accuracy %**
1	Dilation (plus, 1 x 1) Erosion (ellipse, 5 x 5)	Seniors Prefer Africa	86.63
2	Dilation (plus, 5 x 5) Erosion (square, 1 x 1)	Although none of them have	83.24
3	Dilation (plus, 1 x 1) Erosion (ellipse, 5 x 5)	"signed anything yet," they will be	90.22
4	Erosion (square, 1 x 1) Dilation (ellipse, 5 x 5)	sworn in the day before graduation,	89.88
5	Erosion (ellipse, 1 x 1) Dilation (plus, 1 x 1)	and the African theatre of war is	85.68
6	Dilation (plus, 1 x 1) Erosion (ellipse, 5 x 5)	their first choice for active duty. Cynthia Van Cleve and Ruth Davies	84.53
7	Erosion (square, 1 x 1) Dilation (ellipse, 5 x 5)	are among others who arc consider-	87.27
8	Dilation (plus, 1 x 1) Erosion (ellipse, 5 x 5)	are among others who arc consider-	80.72


1. Seniors Prefer Africa
2.
3. Although none of them have
4. "signed anything yet," they will be
5. sworn in the day before graduation,
6. and the African theatre of war is
7. their first choice for active duty.
8. Cynthia Van Cleve and Ruth Davies
9. are among others who are consider-
10. ing joining the ATC.

1. Seniors Prefer Attic:
2.
3. Althuugh nnnc of them have
4. "sigucd zcmtlminxg yet." they will be
5. sworn in the day before graduation,
6. and thc .'\t'ric.'m thczmc wf war it
7. their firs! chnicc far active duty.
8. (')'nthia Van ('love and Ruth Davies
9. arc anmng nthcrs who an cunsider—
10. ing joining the ATC.

Figure 14. Comparison between the input "Test_C" reference text (left) and the basic Tesseract execution result (right)



```
1. Seniors Prefer Africa
2.
3. Although none of them have
4. “signed anything yet,” they will be
5. sworn in the day before graduation,
6. and the African theatre of war is
7. their first choice for active duty.
8. Cynthia Van Cleve and Ruth Davies
9. are among others who are consider-
10. ing joining the ATC.
```

```
1. Seniors Prefer Africa
2. .»\lthough none of them have
3. “signed anything yet.” they will be
4. sworn in the dav before graduation.
5. and the .-\fricau theatre of war is
6. their first choice for active duty.
7. C.\'nthia Van Cleve and Ruth Davies
8. are among others who an consider-
9. ing joining the .-\'1'C.
```

Figure 15. Comparison between the input “Test_A” reference text (left) and the proposed voting-method result (right)

Table 4. Comparison between the proposed voting-based method and Tesseract 3.02 in test scenario 2

Test image	Accuracy of characters (%)	Accuracy of words (%)	Technology used
Test_C	82.32	63.26	Tesseract 3.02
Test_C	**94.82**	**89.79**	**Proposed method**

The purpose of this paper was to propose and demonstrate the viability of voting-based methods to improve the accuracy of an entire document image processing system, including the final OCR results, by designing and implementing a voting system based on varying OCR input data using a set of image processing filters. The filters used are only based on morphological transformations along with a global thresholding method. Various combinations were tried in terms of size, shape and operator applied (erosion, dilation) which in the end provided an average of 4-5% better text accuracy.

5. FUTURE WORK

The implementation of the proposed solution may be further refined to obtain better results in the OCR processing. Moreover, optimization and efficient implementation of a greater number of preprocessing methods should be considered for further improvements while also designing and implementing a parallel processing architecture considering the significant time requirements for preprocessing (obtaining voting candidates) as well as for the required intermediate steps (skew correction, template construction) followed by the execution of the OCR engine processing.

In conclusion, the proposed voting system can be successfully used to improve the accuracy of OCR in case of significantly deteriorated document images, when the quality of character detection is more important than processing speed.

Another approach that can be considered is implementing a system of voting between two or more existing OCR technologies. Each such technology candidate (OCR engine) would have an empirically grade associated that would be a representation of how good it is compared with the other alternatives. The proposed approach may be further improved by combining this pre-evaluated grade with the accuracy percentage obtained by the OCR engine returned for a specific input.

6. REFERENCES

[1] R. Smith, „Tesseract OCR," Google, [Interactive]. Available: https:// code.google.com /p/ tesseract-ocr/. [Accessed 5 June 2015].

[2] R. Smith, „An Overview of the Tesseract OCR Engine," in International Conference on Document Analysis and Recognition, 2007.

[3] R. Smith, „Tesseract OCR Engine," in OSCON, 2007.

[4] S. V. Rice, F. R. Jenkins, T. A. Nartker, „The Fourth Annual Test of OCR Accuracy," Information Science Research Institute, July 1995.

[5] N. Otsu, „A threshold selection method from gray-level histograms," IEEE Trans. Syst. Man Cybern. 9, p. 62-66, 1979.

[6] R. Gupta, P. Jacobson, E. K. Gracia, „OCR binarization and image pre-processing for searching historical documents," Pattern Recognition 40, p. 389 - 397, 2007.

[7] C. A. Boiangiu, M. Simion, V. Lionte, „Voting-Based Image Binarization," Journal of Information Systems & Operations Management (JISOM), Vol. 8, Issue 2, pp. 127-136, 2014.

[8] P. Daniel Ratna Raju and G. Neelima, "Image Segmentation by using Histogram Thresholding," IJCSET, vol. 2, no. 1, pp. 776-779, 2012.

[9] J. Sauvola, M. Pietikainen, „Adaptive document image binarization," Pattern Recognition 33 (2), p. 225¬236, 2000.

[10] Naveed Bin Rais, M. Shehzad Hanif, A. T. Imtiaz, „Adaptive Thresholding Technique for Document Image Analysis," in Multitopic Conference, 2004. Proceedings of INMIC 2004. 8th International, Lahore, Pakista, Dec. 2004.

[11] G. Stockman, L. G. Shapiro, Computer Vision, Prentice-Hall, 2001, pp. 305-315.

[12] R. C. Gonzalez, R. E. Woods, Digital Image Processing Second Edition, Pretice Hall, 2002, pp. 612-626.

[13] C. A. Boiangiu, R. Ioanitescu, „Voting-Based Image Segmentation," Journal of Information Systems & Operations Management (JISOM), Vol. 7, No. 2, pp. 211-220, 2013.

[14] J. Roerdink, A. Meijster, „The watershed transform: definitions, algorithms and parallelization strategies," Fundamenta Informaticae, vol. 41, pp. 187-288, 2000.

[15] H. Jaekyu, M. Haralick, T. Phillips, „Recursive X-Y Cut using Bounding Boxes of Connected Components," in Third International Conference on Document Analysis and Recognition (ICDAR '95), 1995.

[16] C. A. Boiangiu, P. Boglis, G. Simion, R. Ioanitescu, „Voting-Based Layout Analysis," Journal of Information Systems & Operations Management (JISOM), Vol. 8, No. 1, pp. 39-47, 2014.

[17] T. M. Breuel, „Two Geometric Algorithms for Layout Analysis," Lecture Notes in Computer Science, vol. vol. 2423/2002, pp. 687-692, 2002.

[18] G. Bradsky, A. Kaehler, Learning OpenCV, O'Reilly, 2008, pp. 115-124.

[19] M. Helinski, M. Kmieciak, T. Parkota, „Report on the comparison of Tesseract and ABBYY FineReader OCR engines," 2012.

[20] Y. Bassil, M. Alwani, „OCR Context-Sensitive Error Correction Based on Google Web 1T 5-Gram Data Set," American Journal of Scientific Research, No. 50, 2012.

[21] R. Hart, O. Duda, „Use of the Hough Transformation to Detect Lines," Comm. ACM, vol. 15, pp. 11-15, 1972.

[22] J. Knox, „Project Gutenberg's the Works of John Knox Vol. 1," 2007. [Interactive]. Available: http:// www.gutenberg.org/ files/ 21938/ 21938-h/ 21938-h.htm #advertisement. [Accessed 23 July 2015].

[23] Gonzalez, Rafael C. & Woods, Richard E. (2002). Thresholding. In Digital Image Processing, pp. 595–611. Pearson Education.

[24] Razvan-Costin Dragomir, "OCR prin votare", License Thesis, Unpublished Work, Bucharest, Romania, 2015.

[25] DiffChecker, Available: https://www.diffchecker.com/, [Accessed 17 April 2016].

[26] Manolache Florentina Cristina, Sofron Angela, Stanciu Adelina, Costin-Anton Boiangiu – "Study on the Optimizations of the Hough Transform for Image Line Detection", Journal of Information Systems & Operations Management (JISOM), Vol. 7 No. 1, pp. 141-155, 2013.

[27] Costin-Anton Boiangiu, Mihai Cristian Tanase, Radu Ioanitescu, "Text Line Segmentation in Handwritten Documents Based On Dynamic Weights", Journal of Information Systems & Operations Management (JISOM), Vol. 7 No. 2, pp. 247-254, 2013.