ICFHR 2018 Competition on Recognition of Historical Arabic Scientific Manuscripts – RASM2018

Christian Clausner
School of Computing, Science & Engineering
University of Salford
Salford, United Kingdom
c.clausner@primaresearch.org

Apostolos Antonacopoulos
School of Computing, Science & Engineering
University of Salford
Salford, United Kingdom
a.antonacopoulos@primaresearch.org

Nora McGregor
Digital Research Team
British Library
London, United Kingdom
Nora.McGregor@bl.uk

Daniel Wilson-Nunn
Alan Turing Institute
British Library
London, United Kingdom
dwilson-nunn@turing.ac.uk

***Abstract*— This paper presents an objective comparative evaluation of page analysis and recognition methods for historical scientific manuscripts with text in Arabic language and script. It describes the competition (modus operandi, dataset and evaluation methodology) held in the context of ICFHR2018, presenting the results of the evaluation of six methods – three submitted and three baseline systems. The challenges for the participants included page segmentation, text line detection, and optical character recognition (OCR). Different evaluation metrics were used to gain an insight into the algorithms, including new character accuracy metrics to better reflect the difficult circumstances presented by the documents. The results indicate that, despite the challenging nature of the material, useful digitisation outputs can be produced.**

Keywords*— *performance evaluation, page analysis, optical character recognition, OCR, layout analysis, recognition, datasets

I. INTRODUCTION

The British Library's collection of Arabic manuscripts [1] is internationally recognised as one of the largest and finest in Europe and North America, comprising almost 15,000 works in some 14,000 volumes. Since 2012, the Library, in partnership with The Qatar Foundation and Qatar National Library, has digitised and made freely available over 950,000 images and counting, featuring the cultural and historical heritage of the Gulf and wider region, on Qatar Digital Library (QDL).

Ranging from the early eighth century CE to the nineteenth century, the manuscripts are drawn from both Arab countries and other countries with Arab or Muslim communities including India, China, Indonesia, Malaysia, and West Africa, and they display fascinating variations in style and script.

As part of this project, this competition was organised to pose a challenge focussing on finding an optimal solution for accurately and automatically transcribing our vast and growing digital archive of historical Arabic scientific handwritten manuscripts within the QDL. The aim is to improve accessibility of this rich content by enabling full-text search and discovery, as well as enabling large-scale text analysis.

Page Analysis (here page segmentation, region classification, and text recognition) is a central step in the recognition workflow. Its performance significantly influences the overall success of a digitisation system, not only in terms of OCR accuracy but also in terms of the usefulness of the extracted information (in different use scenarios).

This competition was organised in collaboration with the British Library and is a spin-off from a long-standing series of ICDAR page segmentation competitions (the oldest running ICDAR competition since 2001). The aim has been to provide an objective evaluation of methods, on realistic datasets, enabling the creation of a baseline for understanding the behaviour of different approaches in different circumstances. Other evaluations of page segmentation methods have been constrained by their use of indirect evaluation (e.g. the OCR-based approach of UNLV [2]) and/or the limited scope of the dataset (e.g. the structured documents used in [3]. In addition, a characteristic of most competition reports has been the use of rather basic evaluation metrics. While the latter point is also true to some extent of early editions of this competition series, which used precision/recall type of metrics, the 5th edition of the ICDAR Page Segmentation competition (ICDAR2009) [4] made significant additions and enhancements.

This edition (RASM2018) is based on the same principles established and refined by the 2011, 2013, 2015, and 2017 competitions on historical document layout analysis [5] but its focus is on documents with Arabic text. The evaluation metrics selected for this competition reflect the significant need to identify robust and accurate methods for large-scale digitisation initiatives.

An overview of the competition and its modus operandi is given next. In Section III, the evaluation dataset used and its general context are described. The performance evaluation methodology is described in Section IV, while each participating method is summarised in Section V. Finally, different comparative views of the results of the competition are presented and the paper is concluded in Sections VI and VII.

II. The Competition

RASM2018 had three objectives. The first was a comparative evaluation of the participating methods on a representative dataset (i.e. one that reflects the issues and their distribution across library collections that are likely to be scanned / processed). The second objective was a detailed analysis of the performance of each method from different angles. Finally, the third objective was a placement of the participating methods into context by comparing them to established systems currently used in industry and academia.

The competition proceeded as follows. The authors of candidate methods registered their interest in the competition and downloaded the example dataset (document images and associated ground truth). The Aletheia [7] ground-truthing system (which can also be used as a viewer for results) and code for outputting results in the required PAGE format [8] (see below) were also available for download. Two weeks before the competition closing date, registered authors of candidate methods could download the document images of the evaluation dataset. At the closing date, the organisers received both the executables and the results of the candidate methods on the evaluation dataset, submitted by their authors in the PAGE format. The organisers then verified the submitted results and evaluated them.

Although the goal is to discover a good end-to-end digitisation method for the material at hand, three distinct challenges were proposed: Page segmentation (into regions, also known as blocks or zones), text line detection (with polygonal outlines), and text recognition. Participants could decide for which challenges to take part in.

Fig. 1. Example page images

III. The Dataset

The importance of the availability of realistic datasets for meaningful performance evaluation has been repeatedly discussed (e.g. [9]) and the British Library selected a subset of current digitisation endeavours. For the most part, the scanned images contain single column lines of handwritten text (very regular), with a small amount containing illustrations as well as text. Some pages also contain marginal data such as numbers, handwritten notes, and stamps.

For this competition, the evaluation set consisted of 85 page images as a representative sample ensuring a balanced presence of different issues affecting layout analysis and OCR. Such issues include non-straight text lines, show-through or bleed-through, faded ink, decorations, the presence non-rectangular shaped regions, varying text column widths, varying font sizes and various aging- and scanning-related issues.

In addition to the evaluation set, 15 representative images were selected as the example set that was provided to the authors with ground truth. Examples can be seen in Fig. 1.

Fig. 2. Sample images showing region outlines (blue outline: text, green outline: graphic, green highlight: text lines) and text line outlines with transcribed text of a selected text line.

The ground truth is stored in the XML format which is part of the PAGE (Page Analysis and Ground truth Elements) representation framework [8]. For each region on a page there is a description of its outline in the form of a closely fitting polygon. A range of metadata is recorded for each different type of region. For example, text regions hold information their logical label (e.g. heading, paragraph, caption, footer, etc.) among others. Moreover, the format offers sophisticated means for expressing reading order and more complex relations between regions. Sample images with ground truth description can be seen in Fig. 2. The text transcription was initially collected via crowdsourcing and finalised by authors.

IV. Performance Evaluation

A. Layout Analysis

The page layout and text line segmentation performance analysis method used for this competition [10] can be divided into two main parts. First, correspondences between ground truth and segmentation result regions (or text lines) are determined based on overlapping and missed parts. Secondly, errors are identified, quantified and qualified in the context of different use scenarios.

The region correspondence determination step identifies geometric overlaps between ground truth and segmentation result regions. In terms of Segmentation, the following situations can be determined: merge, split, miss / partial miss, and false detection. In terms of Region Classification, considering also the type of a region, an additional situation can be determined: misclassification (not applicable for text lines).

Based on the above, the segmentation and classification errors are quantified, recoding the amount of each single error. This data (errors) is then qualified by the significance, using two levels. The first is the implicit context-dependent significance. It represents the logical and geometric relation between regions. Examples are allowable and non-allowable mergers. A merger of two vertically adjacent paragraphs in a given column of text can be regarded as allowable, as the result will not violate the reading order. On the contrary, a merger between two paragraphs across two different columns of text is regarded as non-allowable, because the reading order will be violated. To determine the allowable/non-allowable situations accurately, the reading order, the relative position of regions, and the reading direction and orientation are taken into account.

The second level of error significance reflects the additional importance of particular errors according to the use scenario for which the evaluation is intended. For the evaluation of text line segmentation, for example, False Detection errors were disregarded entirely because the ground truth contains only the text lines of the main text blocks. Detected lines within marginalia were not to be penalised.

Both levels of error significance are expressed by a set of weights, referred to as an evaluation profile [10]. Appropriately, the errors are also weighted by the size of the area affected (excluding background pixels). In this way, a missed region corresponding to a few characters will have less influence on the overall result than a miss of a whole paragraph, for instance.

For comparative evaluation, the weighted errors are combined to calculate overall error and success rates.

B. *Text Recognition*

For the evaluation of OCR results, character-based and word-based measures were used. The former gives a detailed insight into the recognition accuracy of a method while the word-based approach is more realistic in terms of use scenarios such as keyword-based search.

A significant problem for the evaluation is the influence of the reading order of text regions. For simple page layouts, the order is obvious, but for more complex layouts, the reading order can be ambiguous. In such cases, measures that are affected by the reading order are less meaningful. An OCR method might recognise all characters perfectly, but if it does not return the regions in the same order as in the ground truth (or with merge/split errors), it will get a very low performance score. Special care was therefore taken when selecting the evaluation measures.

The Character Accuracy [12] is based on the edit distance (insertions, deletions and substitutions) between ground truth and OCR result. The method was extended by the authors to reduce the influence of the reading order. The edit distance is thereby calculated for parts of the texts, starting with good matches and marking matched parts as "visited" until the whole text was processed (unmatched parts count as deletion or insertion errors). The extended measure is called Flex Character Accuracy. Fig. 3 illustrates a simple example. Assuming all characters were recognised perfectly by two OCR systems X and Y, the Character Accuracy scores can differ significantly merely due to differences in the detected reading order of paragraphs. If the order in an OCR result is different from the ground truth order, the edit distance is high and the resulting score low. For the example, the Flex Character Accuracy measure would return high scores for both method X and method Y.

The word-based measure called Bag of Words (see [11]) disregards reading order since it only looks at the occurrence of words and their counts, not at the context or location of a word.

Due to the historic and, in some cases, unusual spelling and use/lack of diacritics, a text normalisation was performed for both ground truth and OCR results. The evaluation was performed twice, once with the original texts and once with the normalised texts. The normalised version is less strict.

All evaluation methods and the datasets are available at the PRImA website [13].

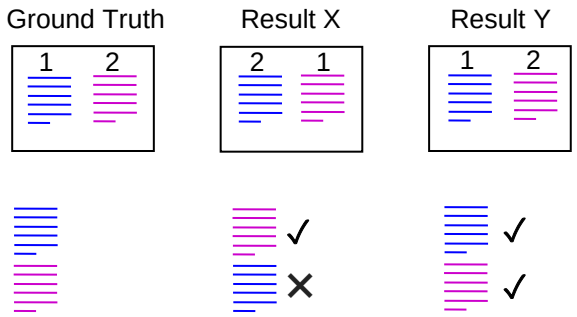

Fig. 3. Impact of reading order on Character Accuracy (top: paragraph order on page; bottom: serialised text). Given two paragraphs, the order in which the text is serialised is significant. OCR result X has the opposite paragraph order than the ground truth and will get a low Character Accuracy score. OCR result Y has the same paragraph order as the ground truth and will get a high Character Accuracy score.

V. Participating Methods

Brief descriptions of the methods submitted to the competition are given next. Each account has been provided by the method's authors and summarised by the organisers.

A. *Google Cloud Vision API*

The Google entry for RASM2018 is a small client program that communicates with the publicly accessible Google Cloud Vision API: https://cloud.google.com/vision/. The DOCUMENT_TEXT_DETECTION feature is selected, which instructs the service to expect dense, book-like page images, as opposed to material such as natural scene images, which is supported under the TEXT_DETECTION feature. The client program is a thin wrapper around the cloud service call without any customization; hence the results are what any other user of the cloud service would have obtained at the time of this submission (May 2018). Since Cloud Vision models get updated periodically, re-running at a later date may produce different results.

Behind the API, the OCR process is split into five stages: text detection, direction identification, script identification, line decoding, and layout analysis. Details of each stage can be found in "A Web-Based OCR Service for Documents" [14]. If the language of the source images is known, the API accepts a language hint provided as a BCP-47 language code, which bypasses the script identification stage. For the

RASM2018 entry, an Arabic language hint ("ar") was provided.

The API returns a response in JSON format containing information from each detected text line, including text results, bounding box coordinates, and confidence scores. The lines are grouped into higher-order structures such as paragraphs and blocks.

B. KFCN

This method was submitted by Berat Kurar from the Ben-Gurion University of the Negev.

Page segmentation method is based on Fully Convolutional Network (FCN) [16]. FCN was trained on 2372 patches generated from the 10 pages of example set using a sliding window of 800×800 pixels with a step size of 400 pixels. Trained FCN was used to predict page layouts of 85 evaluation set pages. From the predicted page layouts, paragraph parts were cropped for text line segmentation. Text line segmentation method is based on anisotropic Gaussian smoothing [17]. This method's output is pixel level labels. We converted pixel level labels into non-intersecting polygons using concave hull of the pixels in a given text line.

C. RDI

This method was submitted by Hany Ahmed from the RDI Company, Cairo University.

RDI-Corporation has its own Historical Arabic Handwritten/Typewritten Optical Character Recognition (OCR) system which has been built from different historical manuscripts. This system has been used before for converting many Arabic typewritten historical manuscripts to its corresponding text to be searchable and editable, for this reason our main concern is to get the corresponding text only without any diacritics or dotless letters. The proposed system can deal with different layout and different fonts. In this version, the system is able to extract lines from the input images without main text block detection using instance segmentation algorithm. The algorithm is expected to precise locations and classes of all lines in the image. Furthermore, the algorithm is able to distinguish different instances of the same class in the image. After the process of lines segmentation, we start recognizing line by line. For the competition, we adapted our system on the given samples and different images which are similar to the given samples. We participate in this competition targeting two main challenges, Text line detection / segmentation (Challenge 2), and Text Recognition (Challenge 3).

D. Baseline methods

Tesseract OCR 3.04, 4.0 (beta version) [15][18] and ABBYY FineReader Engine 11 were used for comparison. Tesseract 4.0 is based on a long short-term memory (LSTM) approach. No training was required as a language model for Arabic is available. The PRImA Tesseract-to-PAGE wrapper tool was used to create PAGE XML for Tesseract 3.04. Tesseract 4.0 was executed using the native command line tool and the output was converted from hOCR format to PAGE XML format using the PRImA PageConverter tool. FineReader results were produced by using the provided API and exporting results in PAGE XML.

It should be noted that Tesseract and FinerReader are optimised for printed text. Nevertheless, they represent a baseline of readily available methods that can process manuscripts, as long as the font is not too variable.

VI. Results

Evaluation results for the above methods are presented in this section in the form of a table and, in part, with corresponding graphs.

0shows the results of all methods for the three challenges. Cases where a method did not return results for a particular challenge are marked as "N/A".

Fig. 4 shows the success rates for the page segmentation challenge and Fig. 5 shows the respective error analysis. The KFCN method returned very good results, given the quality of the images. The Google method focuses on text blocks and its results should be interpreted respectively. The baseline methods score low success rates with Tesseract 3 and FineReader often misclassifying text areas as illustrations.

The page segmentation results (challenge 1) are likely to have an impact on the text line segmentation (and the OCR). For text lines, the main differentiating factor between KFCN and RDI is miss error (KFCN has higher miss error rate). FineReader's region misclassification leads to a large proportion of missed text lines. Fig. 6 shows the measured errors. As mentioned earlier, false detection errors were excluded for text lines.

Fig. 7 shows the text recognition evaluation results (Flex Character Accuracy). The RDI method reaches very high accuracies. Google's results are good as well, considering they did not specifically train or optimise for this competition.

The difference between Character Accuracy and Flex Character Accuracy is most noticeable for Tesseract (up to about 10% difference). Most of the pages in the competition dataset have a simple layout and the reading order is not ambiguous. Tesseract has an overall low page segmentation score, most likely including issues with the reading order.

Comparing the three different text-based measures for the original texts and the normalised texts, it clear that the normalisation has a big influence on the accuracies. RDI's word success rate, for example, jumps from 42% to 61%.

RDI outperforms all methods for all text-based measures.

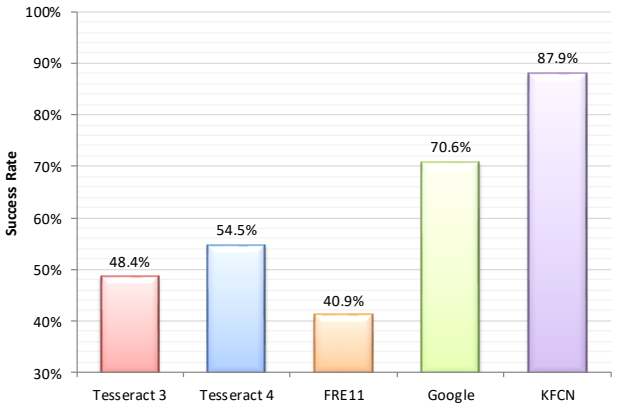

Fig. 4. Page segmentation results

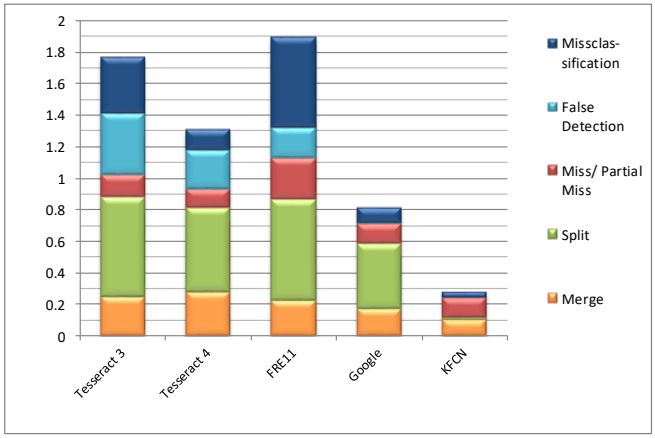

Fig. 5. Page segmentation errors

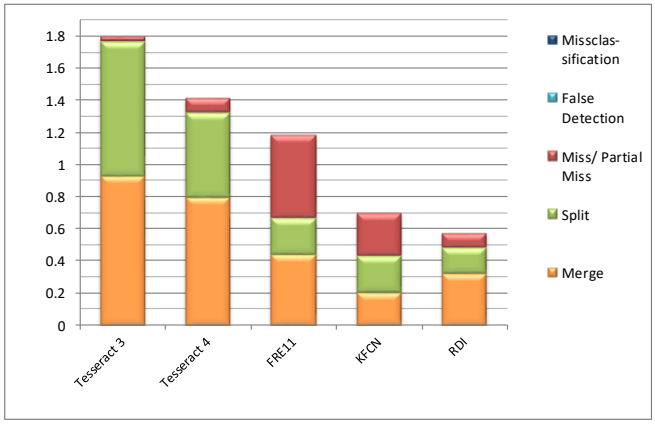

Fig. 6. Text line segmentation errors

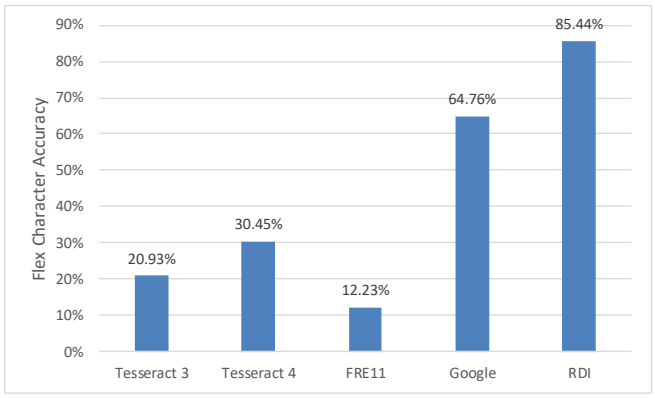

Fig. 7. OCR accuracy (normalised text)

TABLE I. EVALUATION RESULTS (VALUES IN PERCENT SUCCESS)

	Tesseract 3	**Tesseract 4**	**FRE11**	**Google**	**KFCN**	**RDI**
Challenge 1 (Page segmentation)	48.4	54.4	40.9	70.6	**87.9**	N/A
Challenge 2 (Text lines)	28.8	44.2	43.2	N/A	67.7	**81.6**
Challenge 3 (OCR)						
	Original text					
Character accuracy	13.0	18.3	11.0	60.4	N/A	**78.1**
Flex character accuracy	20.9	27.8	10.8	60.6	N/A	**78.1**
Bag of words success rate	2.2	4.7	0.4	20.9	N/A	**42.3**
	Normalised text					
Character accuracy	13.3	19.2	12.3	64.4	N/A	**85.4**
Flex character accuracy	20.9	30.5	12.2	64.8	N/A	**85.4**
Bag of words success rate	2.5	5.5	0.4	26.7	N/A	**60.6**

VII. CONCLUDING REMARKS

To the best of the authors' knowledge, this competition constitutes the first objective comparative evaluation of page analysis and recognition approaches for historical scientific manuscripts in Arabic. It has highlighted the technical difficulties faced by the most advanced methods currently available from academia and industry.

The KFCN method delivered the best results for page segmentation whereas the RDI method is the winner for challenges two and three (text lines and OCR). Google only submitted for challenge three, their results for challenge one are only for comparison.

Good results were achieved in all three challenges. Areas for improvement include region separation. Marginalia close to the main text often disrupted the digitisation methods. Dedicated historical dictionaries could improve the OCR results further.

ACKNOWLEDGEMENTS

We thank all curators and ground truthers (from the British Library or elsewhere) who helped creating the competition dataset.

REFERENCES

[1] https://www.bl.uk/collection-guides/arabic-manuscripts, The British Library, accessed 29/06/2018

[2] J. Kanai, S.V. Rice, T.A. Nartker and G. Nagy, "Automated Evaluation of OCR Zoning", IEEE PAMI, 17(1), 1995, pp. 86-90.

[3] F. Shafait, D. Keysers and T.M. Breuel, “Performance Evaluation and Benchmarking of Six Page Segmentation Algorithms” IEEE PAMI, 30(6), 2008, pp. 941–954.

[4] A. Antonacopoulos, S. Pletschacher, D. Bridson, C. Papadopoulos, “ICDAR2009 Page Segmentation Competition”, Proc. ICDAR2009, Barcelona, Spain, July 2009, pp. 1370-1374.

[5] C. Papadopoulos, S. Pletschacher, C. Clausner, A. Antonacopoulos, “The IMPACT dataset of Historical Document Images”, Proc. HIP2013, Washington DC, USA, August 2013, pp. 123-130.

[6] A. Antonacopoulos, C. Clausner, C. Papadopoulos, S. Pletschacher, “ICDAR2013 Competition on Historical Newspaper Layout Analysis – HNLA2013”, Proc. ICDAR2013, Washington DC, USA, Aug 2013.

[7] C. Clausner, S. Pletschacher and A. Antonacopoulos, “Aletheia - An Advanced Document Layout and Text Ground-Truthing System for Production Environments”, Proc. ICDAR2011, Beijing, China, 2011.

[8] S. Pletschacher and A. Antonacopoulos, "The PAGE (Page Analysis and Ground-Truth Elements) Format Framework", Proc. ICPR2008, Istanbul, Turkey, August 23-26, 2010, IEEE-CS Press, pp. 257-260.

[9] C. Clausner, C. Papadopoulos, S. Pletschacher, A. Antonacopoulos “The ENP Image and Ground Truth Dataset of Historical Newspapers”, Proc. ICDAR2015, Nancy, France, Aug. 2015, pp. 931-935.

[10] C. Clausner, S. Pletschacher and A. Antonacopoulos, “Scenario Driven In-Depth Performance Evaluation of Document Layout Analysis Methods”, Proc. ICDAR2011, Beijing, China, Sept 2011.

[11] A. Antonacopoulos, C. Clausner, C. Papadopoulos, S. Pletschacher, “ICDAR2013 Competition on Historical Book Recognition – HBR2013”, Proc. ICDAR2013, Washington DC, USA, Aug 2013.

[12] S.V. Rice, “Measuring the Accuracy of Page-Reading Systems”, PhD thesis, University of Nevada, Las Vegas December 1996.

[13] PRImA Performance Evaluation Tools http://www.primaresearch.org/tools/PerformanceEvaluation

[14] J. Walker, Y. Fujii, A. C. Popat “A Web-Based OCR Service for Documents” in Proceedings of the 13th IAPR International Workshop on Document Analysis Systems (DAS), Vienna, Austria, Apr. 2018.

[15] R. Smith, "An Overview of the Tesseract OCR Engine," Ninth International Conference on Document Analysis and Recognition (ICDAR 2007), Parana, 2007, pp. 629-633. doi: 10.1109/ICDAR.2007.4376991

[16] B. Kurar and J. El-sana, “Binarization free layout analysis for arabic historical documents using fully convolutional networks” in Arabic Script Analysis and Recognition (ASAR), 2018 2nd International Workshop on. IEEE, 2018.

[17] M. Kassis, B. Kurar, R. Cohen, J. El-Sana, and K. Kedem, “Using scale-space anisotropic smoothing for text line extraction in historical documents,” Submitted to International Journal on Document Analysis and Recognition (IJDAR), 2018.

[18] Tesseract OCR: https://github.com/tesseract-ocr, accessed 11/07/2017