ICDAR2019 Competition on Recognition of Early Indian Printed Documents – REID2019

C. Clausner[1], A. Antonacopoulos[1], T. Derrick[2] and S. Pletschacher[1]

1: Pattern Recognition and Image Analysis Research Lab
School of Computing, Science and Engineering
University of Salford
Greater Manchester, M5 4WT, United Kingdom
www.primaresearch.org

2: Digital Scholarship
British Library
London, NW1 2DB, United Kingdom
www.bl.uk/subjects/digital-scholarship

***Abstract*—This paper presents an objective comparative evaluation of page analysis and recognition methods for historical documents with text mainly in Bengali language and script. It describes the competition rules, dataset, and evaluation methodology. Results are presented for five methods – three submitted, one re-run, and one open source state-of-the-art system. The focus is on optical character recognition (OCR) performance. 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 deep learning approaches are promising, but there are still significant challenges for historic material of this nature.**

Keywords - performance evaluation; page analysis; optical character recognition; OCR; layout analysis; recognition; datasets;

I. INTRODUCTION

Since 2016 the British Library (BL) has been digitising unique and rare early Indian printed books drawn from the Library's South Asian printed books and periodicals collection. More than 3,600 books (1713-1914) have been digitised and are being made available openly online through the *Two Centuries of Indian Print* project. The printed books are supplemented by catalogues known as Quarterly Lists, containing tables of data recording of all books published in India between 1867 and 1947. These too have been made available as open access.

The books encompass an extensive range of academic disciplines and topics, yet up until now much of the material has only been accessible in physical form by visiting the Library. Providing accurate transcriptions will therefore be of great benefit to the research community, enabling full-text analysis of the material which may yield new insights into areas of South Asian studies.

The quality of information extraction from printed material heavily depends on the performance of individual processing steps such as page segmentation, region classification and OCR. The usefulness of the extracted information is subject to the use scenario the data is intended for. The evaluation of digitisation methods should therefore be flexible to be able to reflect different scenarios.

Recent deep learning technologies promise to advance OCR beyond traditional approaches. Previous competitions and reviews show, however, that historical material poses additional challenges (little training data, low image quality, spelling variations etc.) which have not yet been overcome in a satisfactory way.

This competition was organised in collaboration with the British Library and is the second edition of a spin-off from a long-standing series of ICDAR page segmentation competitions. 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 (REID2019) is based on the same principles established and refined by the 2011 to 2017 competitions on historical document layout analysis [5] but its focus is on text recognition performance. The evaluation metrics selected for REID 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

REID2019 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).

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 open-source systems currently used in industry and academia.

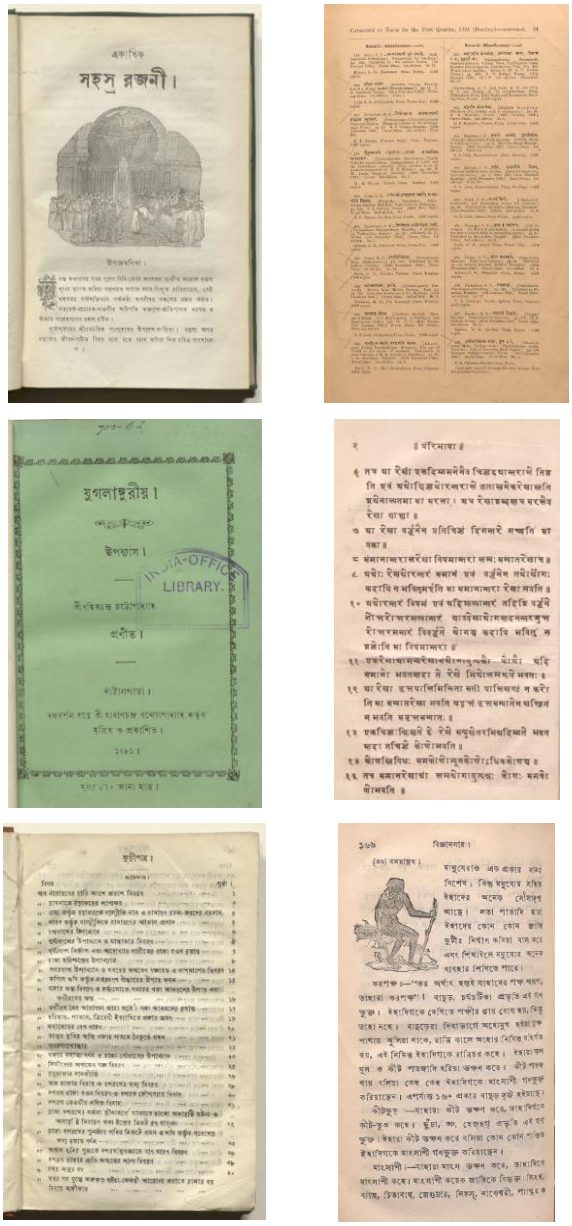

Figure 1. Example page images.

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.

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. The competition was originally composed of two challenges, but no submissions were made for the Quarterly Lists bonus challenge (recognition of tabular material in both English and Bengali), leaving only the Bengali texts. The corresponding digitisation project at the BL will be digitising 2,700 printed books written in Bengali (1713-1914), amounting to about 1,000,000 pages in TIFF format. For the most part, the scanned images contain single column lines of text, with a small amount containing illustrations as well as text. Some pages contain marginal data such as numbers, handwritten notes, and decorative frames.

For this competition, the evaluation set consisted of 56 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, presence of separators and various aging- and scanning-related issues.

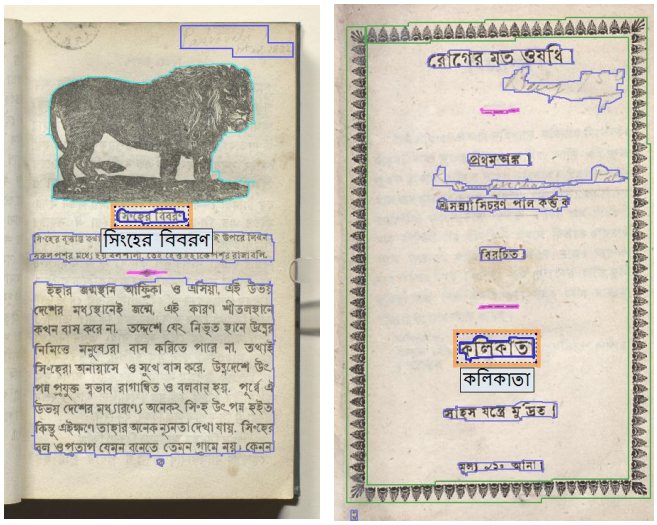

Figure 2. Sample images showing the region outlines (blue: text, magenta: separator, green: graphic, cyan: image) and text content of a selected region.

In addition to the evaluation set, 25 representative images were selected as the example set that was provided to the authors with ground truth. Examples from both sets can be seen in Fig. 1.

The ground truth is stored in PAGE XML [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 provided by the School of Cultural Texts and Records at Jadavpur University.

The dataset, including all ground truth, is available for download at primaresearch.org/datasets.

IV. Performance Evaluation

A. Layout Analysis

The page layout performance analysis method used for this competition [10] can be divided into two main parts. First, correspondences between ground truth and segmentation result regions 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 Page 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.

Based on the above, the segmentation and classification errors are *quantified*, recording 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.

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 major 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 (see [13]).

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

Because some of the document pages contain padding characters such as "….." or "- - - -", a pre-processing step is performed to remove special characters from all ground truth and OCR result texts. These include: hyphen, dash, full stop, tilde, asterisk, equal sign, bullet, and double quotes. In addition, unnecessary white spaces are removed (e.g. multiple spaces and trailing line breaks). This helps to focus the evaluation on the more interesting parts of the documents.

All evaluation methods and the datasets are available at the PRImA website [14].

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. ABCD

This method was submitted by Showmik Bhowmik, Soumyadeep Kundu, and Ram Sarkar from the Department of Computer Science and Engineering, Jadavpur University, India.

In this method an input color image I is initially converted into its corresponding grayscale image and then into its binary version I_b. After that the components of I_b are examined on the basis of their height, width, density and area to separate the non-text components from the text components. In this stage the text-only image I_t and non-text-only image I_{nt} are generated. Next, a region segmentation process is performed

on I_t. For that purpose, morphological dilation is applied iteratively on I_t. In this process, the dimension of the structuring element gets changed at each epoch based on the size of the connected components present in the image generated at previous iteration. That means the dimension of the structuring element for dilation in i^{th} epoch is decided on the basis of the size of the connected components present in the image generated at $(i-1)^{th}$ epoch. In addition, at each epoch, dilation is performed twice, one with the 0^0 rotation of the structuring element and another is with 90^0 rotation of the same structuring element. This dilation is continued until the number of components present in the currently generated image is reduced to an experimentally chosen threshold value. At the end, the segmented image I_t^{seg} is generated. Finally, I_t^{seg} represents the segmented text regions and I_{nt} represents the segmented non-text regions.

B. Bangla OCR

This OCR system has been submitted by Tanmoy Nandi and Sumit Kumar Saha, Gnosis Lab, Kolkata, India, Chandranath Adak, School of Software, University of Technology Sydney, Australia, and Bidyut B. Chaudhuri, Techno India University, Kolkata, India.

This end-to-end system works only with printed Bengali (endonym, Bangla) script. Since the "REID2019: Main Challenge - Recognition of Bengali Books" database contains old printed documents, some rigorous pre-processing is necessary.

In the pre-processing stage, at first, median filtering is performed on the entire document image to remove noise. Then, an erosion-dilation-based morphological operation is adopted to join "broken" components, if exists any. Finally, the document image is cleaned using an improved version of [15].

For the next stage (text recognition) Tesseract OCR's [22] open-source modules were used, pre-trained on Bengali document images. Here, the segmented words, text-lines, and blocks are classified/recognized separately and combined with a hierarchical combinational logic.

Finally, the system produces the Unicode character text and generates a PAGE XML for each document image.

Since this REID2019 competition is only focused on recognition of Bengali pages, the submitted "BanglaOCR" system has not taken care of any other scripts. Therefore, if there exist any other scripts (e.g. Devanagari or English), this system as its current form produces erroneous results.

C. DS

This method was submitted by Soumyadeep Dey and Rohit Srivastava of Microsoft India.

In this work, a technique was developed to detect various regions from a scanned image of early printed documents. The method is not dependent on any script of the document. The provided document images contain various types of noise, such as ink bleed, margin noise, etc. In presence of these issues, identification of text regions is especially difficult. The proposed technique initially cleans the document. After that, a region identification approach is applied to efficiently detect the text regions.

The overall methodology of the proposed work is described below.

1. An adaptive binarization technique is applied to separate foreground pixels from the background using the method proposed in [16].
2. Foreground pixels are grouped together using morphological gap filling operations in horizontal and vertical directions [17] at various stages of the algorithm. At each of these stages, various noise elements are removed.
3. Margin noise is removed using the method proposed in [18].
4. The input document is segmented into components/blocks with the help of the technique presented in [19].
5. EAST [20], is used to detect text from both uncleaned and cleaned images. This information is further used to identify text regions from the initial blocks generated with the method [19]
6. Lastly, text regions image snippets are passed to the Tesseract OCR engine [22] to obtain Unicode text content.

D. Google Multi-Lingual OCR (2017)

The Google entry for REID2017 was rerun (by the competition organisers) on the 2019 dataset. Although not a new submission, the method uses a current Google backend (web service).

The method has 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. No pre-processing or post-processing is performed by the client program; it relies entirely on the publicly available cloud service for the entire operation. Because the 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 three phases: text detection, line decoding, and layout analysis.

Text detection locates individual lines of text in the image; these regions are then extracted and provided to the line decoding phase, described below. Text detection follows the approach described by Bissacco et al. [21].

More information can be found in [5].

E. State-of-the-art Method (Tesseract)

Tesseract 4.0 [22][23] was used for comparison. This version of Tesseract is based on a long short-term memory (LSTM) approach. No training was performed – the available language models for Bengali and English were applied. The PRImA Tesseract-to-PAGE wrapper tool was used to create PAGE XML.

VI. Results

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

Although the primary focus of this competition is text recognition, the performance analysis of page segmentation and region classification also give useful insights to pinpoint problems and improve the OCR methods. Fig. 3 and Fig. 4

show the layout evaluation results using general page analysis profile and a text-focused profile (i.e. errors on non-textual regions are weighted less significantly). Fig. 5 shows the breakdown of the different error types of the evaluation measure.

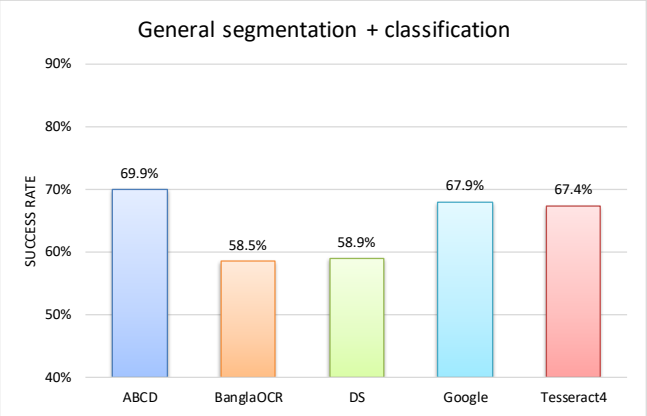

Figure 3. Results using the evaluation profile for general page recogniton (segmetnation + region classification).

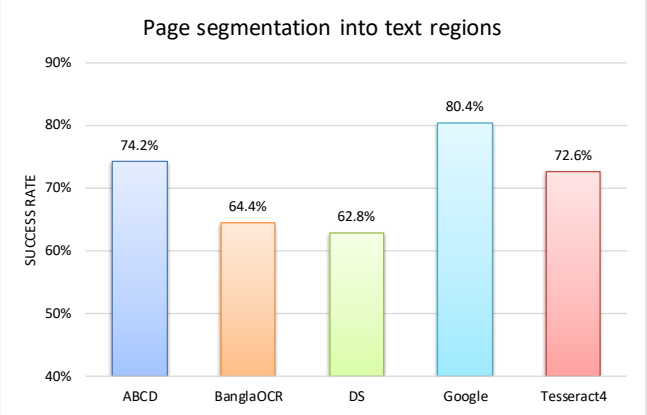

Figure 4. Results using the text region-focused evaluation profile.

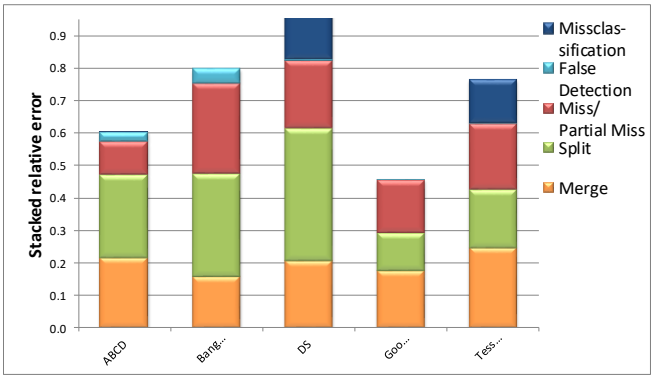

Figure 5. Breakdown of errors made by each method (text-focused profile).

Page segmentation results in REID are far behind of what can be achieved for contemporary documents of good quality. For general page segmentation and region classification (i.e. finding text, illustrations, decorations, etc.), success rates of about 70% can be achieved (ABCD method performs best). However, when focussing on textual regions, Google's strong OCR engine tips the balance in favour of their method, achieving over 80% success rate compared to 74% of the runner-up.

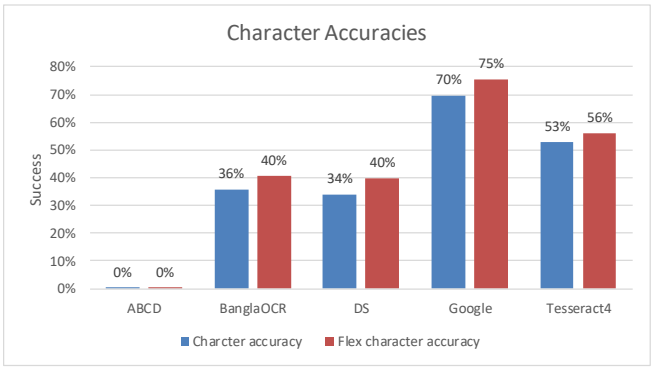

Figure 6. Character accuracy and flex character accuracy.

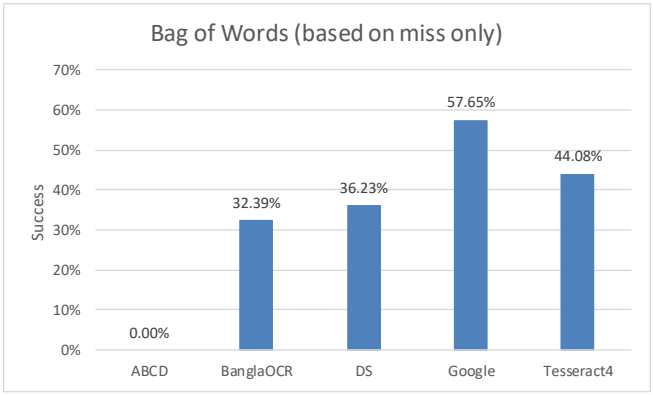

Figure 7. Bag of words success rate (based on miss error).

Fig. 6 shows the traditional and the modified (Flex) character accuracy results. As explained in Section IV.B, there is a clear difference between the two measures, originating from reading order and segmentation variations. The Flex character accuracy is more meaningful with respect to the actual character recognition. TABLE I. shows the accuracies for different subsets. The "2017" set contains only the images that were used in REID2017. The "2019" set contains only the images that were added for this year's competition. Comparing the results of the 2017 subset with the results from REID2017 [5], there is little movement. All methods represented in both competitions show virtually no change.

Although the Google Cloud Vision outperforms the other methods in the given scenario, but an accuracy of less than 78% is far from satisfactory.

Considering real-world use cases such as page retrieval via keyword search, a word-based measure is more meaningful. Fig. 7 shows the results for the Bag of Words measure. As can be expected, the success values are lower than the character-based values (one character can cause a whole word to be wrong). The success rate is only based on "miss"

errors (words that are in the ground truth but missing or misspelled in the OCR result). False detection (insertion of non-existent words) is disregarded, reflecting the use scenario of page retrieval. The Google OCR method is still ahead, but the margins are slightly narrower.

TABLE I. FLEX CHARACTER ACCURACY PER SUBSET (IN %)

METHOD	FLEX CHARACTER ACCURACY PER SUBSET		
	2017 (26 pages)	**2019 (30 pages)**	**All (56 pages)**
ABCD	N/A	N/A	N/A
BANGLAOCR	40.5	65.1	53.66
DS	39.6	57.1	49.01
GOOGLE	**75.4**	**79.7**	**77.68**
TESSERACT4	55.8	67.7	62.18

VII. CONCLUDING REMARKS

To the best of the authors' knowledge, this series of competitions constitutes the first objective comparative evaluation of page analysis and recognition approaches for historical Bengali documents. It has highlighted the technical difficulties faced by the most advanced methods currently available from academia and industry. The method from Google outperforms the other methods in this instance but there is much room for improvement for all methods. In fact, in certain situations, other methods outperform the Google's method, especially for pages containing a table of content.

In general, the evaluation shows that even a basic task such as indexing pages based on OCR results will be of limited success. Word-based error rates are 42% and higher.

A clear first candidate for improvement is the pre-processing stage – especially since the material is of historical nature. This could include a robust binarisation to clearly isolate textual characters and developing a classifier that can handle a variety of historical fonts. A sophisticated approach to recognise both text and decorative elements would also be beneficial. In addition, historical spelling and script variations posed a problem which could be overcome by training and/or dictionary creation in a dedicated project.

REFERENCES

[1] www.bl.uk/projects/two-centuries-of-indian-print, The British Library, accessed 11/07/2017

[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. Clausner, A. Antonacopoulos, T. Derrick, S. Pletschacher, "ICDAR2017 Competition on Recognition of Early Indian Printed Documents – REID2017", *Proceedings of the 14th International Conference on Document Analysis and Recognition (ICDAR2017)*, Kyoto, Japan, November 2017, pp. 1411-1416.

[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 text library on GitHub, https://github.com/PRImA-Research-Lab/prima-text

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

[15] K. Ntirogiannis, B. Gatos, I. Pratikakis, "A Combined Approach for the Binarization of Handwritten Document Images", *Pattern Recognition Letters*, vol. 35, pp. 3-15, 2014.

[16] J. Sauvola and M. Pietik¨ainen. Adaptive document image binarization. Pattern Recognition, 33:225–236, 2000.

[17] Soumyadeep Dey, Jayanta Mukhopadhyay, Shamik Sural, and Partha Bhowmick. Margin noise removal from printed document images. In Proceeding of the Workshop on Document Analysis and Recognition, DAR'12, pages 86–93, New York, NY, USA, 2012. ACM.

[18] Soumyadeep Dey, Barsha Mitra, Jayanta Mukhopadhyay, and Shamik Sural. A comparative study of margin noise removal algorithms on marnr: A margin noise dataset of document images. In 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), volume 04, pages 35–39, Nov 2017.

[19] Soumyadeep Dey, Jayanta Mukherjee, and Shamik Sural. Consensus-based clustering for document image segmentation. IJDAR, 19(4):351–368, 2016.

[20] Xinyu Zhou, Cong Yao, He Wen, Yuzhi Wang, Shuchang Zhou, Weiran He, and Jiajun Liang. EAST: an efficient and accurate scene text detector. CoRR, abs/1704.03155, 2017.

[21] A. Bissacco, M. Cummins, Y. Netzer, H. Neven "PhotoOCR: Reading Text in Uncontrolled Conditions" in *IEEE Int. Conference on Computer Vision (ICCV)*, Sydney, Australia, Dec. 2013.

[22] 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

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