ICDAR2017 Competition on Recognition of Documents with Complex Layouts – RDCL2017

C. Clausner, A. Antonacopoulos, and S. Pletschacher
Pattern Recognition and Image Analysis (PRImA) Research Lab
School of Computing, Science and Engineering, University of Salford
Greater Manchester, M5 4WT, United Kingdom
www.primaresearch.org

***Abstract*—This paper presents an objective comparative evaluation of page segmentation and region classification methods for documents with complex layouts. It describes the competition (modus operandi, dataset and evaluation methodology) held in the context of ICDAR2017, presenting the results of the evaluation of seven methods – five submitted, two state-of-the-art systems (commercial and open-source). Three scenarios are reported in this paper, one evaluating the ability of methods to accurately segment regions and two evaluating both segmentation and region classification (one focusing only on text regions). For the first time, nested region content (table cells, chart labels etc.) are evaluated in addition to the top-level page content. Text recognition was a bonus challenge and was not taken up by all participants. The results indicate that an innovative approach has a clear advantage but there is still a considerable need to develop robust methods that deal with layout challenges, especially with the non-textual content.**

Keywords - performance evaluation; page segmentation; region classification; layout analysis; OCR; recognition; datasets;

I. INTRODUCTION

Layout Analysis (Page Segmentation and Region Classification) is a critical 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). Frequently, methods are devised with a specific application in mind and are fine-tuned to the image dataset used by their authors. However, the variety of documents encountered in real-life situations (and the issues they raise) is far wider than the target document types of most methods.

In addition, OCR, largely abandoned by academic researchers, faces challenges in large-scale digitisation and is still not performing well enough to not require costly manual post-correction. Systematic evaluation is crucial to study the issues involved and attempt to make progress.

The aim of the ICDAR Page Segmentation competitions (the oldest running ICDAR competition since 2001) has been to provide an objective evaluation of methods, on a realistic contemporary dataset, 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 [1]) and/or the limited scope of the dataset (e.g. the structured documents used in [2]. In addition, a characteristic of other competition reports has been the use of rather basic evaluation metrics. Since the 2009 edition of the ICDAR Page Segmentation competition a more extensive evaluation scheme has been used [3], allowing for higher level goal-oriented evaluation and much more detailed region comparison, going far beyond simple precision/recall metrics. In addition, the used datasets have been selected from curated repositories [4][5] containing realistic and representative documents. This edition (RDCL2017) is based on the same principles established and refined by the 2011, 2013, and 2015 competitions on historical and contemporary document layout analysis [6] but its focus is on documents with complex layouts. The evaluation scenarios selected for this competition reflect the 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 3, the evaluation dataset used and its general context are described. The performance evaluation methodology is described in Section 4, while each participating method is summarised in Section 5. Finally, different comparative views of the results of the competition are presented and the paper is concluded in Sections 6 and 7.

II. THE COMPETITION

RDCL2017 had the following 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). Delving deeper, the second objective was a detailed analysis of the performance of each method in different scenarios from the simple ability to correctly identify and label regions to a text recognition scenario where the reading order needs to be preserved. This analysis facilitates a better understanding of the behaviour of methods in different digitisation scenarios across the variety of documents in the dataset. Finally, the third objective was a placement of the participating methods into context by comparing them to leading commercial and open-source 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. Three weeks before the competition closing date, registered authors of candidate methods were able to 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.

Figure 1. Page images in the example set.

III. The Dataset

The importance of the availability of realistic datasets for meaningful performance evaluation has been repeatedly discussed and the authors have addressed the issue for contemporary documents by creating the PRImA Layout Analysis dataset with ground truth [4] and making it available to all researchers. The overall dataset contains a wide selection of contemporary documents (with complex as well as simple layouts) together with comprehensive ground truth and extensive metadata. Emphasis is placed on magazines (mostly) and technical articles, which are likely to be the focus of digitisation efforts.

For this competition, the evaluation set consisted of 75 images selected from the PRImA Layout Analysis dataset as a representative sample ensuring a balanced presence of different issues affecting layout analysis and OCR. Such issues include the presence non-rectangular shaped regions, varying text column widths, varying font sizes, presence of separators and regions of “reverse video” text (light-coloured text on a dark background). The presence of running headers and captions of illustrations/photos in addition to the main body of text, pose difficulties in the identification of the correct reading order of the page.

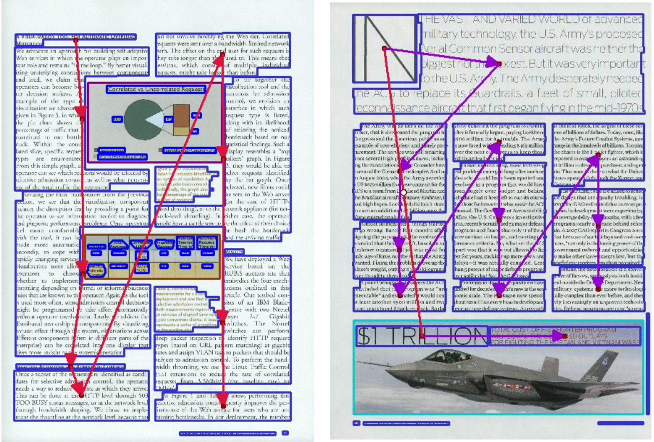


Figure 2. Sample images showing the region outlines (blue: text, purple: chart, brown: table, cyan: image) and reding order.

In addition to the evaluation set, six representative images were selected as the example set that was provided to the authors with ground truth. The pages from the latter can be seen in Fig. 1.

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 the 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 about *language*, *font*, *reading direction*, *text colour*, *background colour*, *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. Structured content can be modelled with regions nesting (regions within regions). For this competition only nested text was taken into account (table cells, text on images, chart labels etc.). Sample images with ground truth description can be seen in Fig. 2.

IV. Performance Evaluation

A. Layout Analysis

The performance analysis method used for this competition [9] can be divided into three parts. First, all regions (polygonal representations of ground truth and method results for a given image) are transformed into an interval representation, which allows efficient comparison and calculation of overlapping/missed parts. Second, correspondences between ground truth and segmentation result regions are determined. Finally, errors are identified, quantified and qualified in the context of one or more 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:

- *Merger*: A segmentation result region overlaps more than one ground truth region.
- *Split*: A ground truth region is overlapped by more than one segmentation result region.
- *Miss (or partial miss)*: A ground truth region is not (not fully) overlapped by a result region.
- *False detection*: A segmentation result region does not overlap any ground truth region.

In terms of Region Classification, considering also the *type* of a region, an additional situation can be determined:

- *Misclassification*: A ground truth region is overlapped by a result region of another type.

Based on the above, the segmentation and classification errors are *quantified*. The amount (based on overlap area) of each single error is recorded (raw evaluation data).

This raw data (errors) are then *qualified* by their significance using two levels of error significance. 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 instance, to build the table of contents for a print-on demand facsimile edition of a book, the correct segmentation and classification of page numbers and headings is very important (e.g. a merger between those regions and other text should be penalised more heavily).

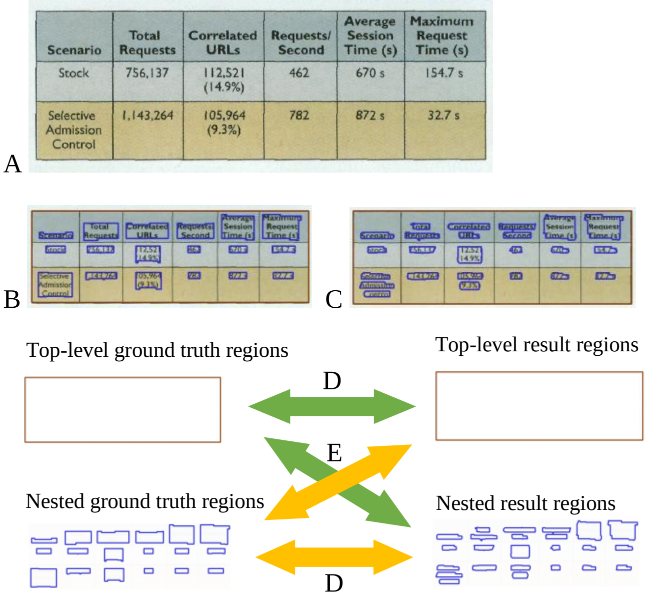
Both levels of error significance are expressed by a set of weights, referred to as an *evaluation profile* [9]. Each evaluation scenario has a corresponding evaluation profile.

Appropriately, the errors are also weighted by the size of the area affected (excluding background pixels). 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. To this end, bitonal images are produced using the Sauvola method (window size 20, weight 0.4).

For comparative evaluation, the weighted errors are combined to calculate overall error and success rates. A non-linear function is used in this calculation to better highlight contrast between methods and to allow an open scale (due to the nature of the errors and weighting).

Nested regions (regions within regions) require special treatment. The evaluation method was extended accordingly. Top-level regions (parent regions) and nested regions (child regions) are thereby treated as being in different layers. For each ground truth region, two error values are calculated: one in the same layer (top-level to top-level) and one across layers (top-level to nested or nested to top-level). The lower of the two is then used as final value for the region. This helps to avoid over-penalisation when the region structure of the page analysis result is different than the ground truth, but most of the content was recognised. For example, in Figure 3. , if the table region would have been missed, but the table cells were recognised (as top-level regions), then the across-layer error would be lower. A penalty is already applied from the same-layer evaluation of the ground truth top-level table region.

Errors where nested regions are involved are globally weighted at 50% in combination with the normal weights defined by the evaluation profile.



Scenario	Total Requests	Correlated URLs	Requests/ Second	Average Session Time (s)	Maximum Request Time (s)
Stock	756,137	112,521 (14.9%)	462	670 s	154.7 s
Selective Admission Control	1,143,264	105,964 (9.3%)	782	872 s	32.7 s

Figure 3. Evaluation of nested regions. A: example snippet; B: ground truth; C: analysis result; D: same-layer, E: across-layer error calculation.

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, it will get a very low performance score. Special care was therefore taken when selecting the evaluation measures.

The Character Accuracy [11] 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.

The word-based measure called *Bag of Words* (see [10]) 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.

All evaluation methods (and datasets) are available at the authors' website [12].

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. The LIPADE Straight-Line-based Method

This method was submitted by Héloïse Alheritiere (supervised by F. Cloppet, C. Kurtz, and N. Vincent) of the LIPADE Computer Science Department of the Paris Descartes University.

The employed method for complex layout extraction is based on features of higher level than pixels obtained from a document straight line based covering catching some information on the dimensions and directions of the objects contained in the binarised document image.

It is proposed to capture straight line segments based on a new transform called the Local Diameter Transform (LDT), which integrates the local spatial organization of the segments contained in the document content. Such segments may be used to approximate filled forms, lines and drawings that constitute another level of document primitives from a topological point of view. Furthermore, according to the length of the segments, some document parts can be discriminated.

The proposed transform is applied simultaneously on the foreground (related to the document content) and the background pixels, in order to take advantage of the duality of information present in both parts of the document to extract its layout.

The segmented document regions are then labelled with respect to different classes of the document layout (text areas, images, separators and tables) thanks to a decision-tree procedure involving high-level rules also derived from the straight line segments properties. As far as text is concerned, the method extracts text lines as regions to be labelled. Then some post processing based on Gestalt theory is used to build paragraphs but no semantic rules are considered to extract reading order and then correct the line merging.

B. The CVML Layout Analysis Method

This system was submitted by Sangyu Han, Soyeon Kim and Hyung Il Koo from the Ajou University, Suwon, Korea.

The method first extracts text lines in images, and estimates paragraph structures using the detected text lines. Then, other regions (e.g. separators, pictures) are extracted in non-text regions.

Text line extraction is performed by extracting connected components (CCs) and grouping the CCs into text lines [13]. The text/non-text classifier in [14] allows the method to work in the presence of noisy components.

After the text region processing, the LSD algorithm [15] is applied to other regions (i.e. non-text regions) in order to detect vertical/horizontal separators. Then, text regions and separators are removed with an inpainting method [16] and picture regions are detected by localizing salient objects in the inpainting results.

C. The MHS 2017 System

The *Page Segmentation Using Multilevel Homogeneous Structure (MHS)* method was submitted by Tuan Anh Tran from the HoChiMinh City University of Technology (Ho Chi Minh City, Viet Nam) and Hai Duong Nguyen, Hong Trai Tran, In Seop Na, and Soo Hyung Kim from Chonnam National University (Gwangju, Republic of Korea).

The method uses the following steps:

1. Binarisation - A combination of Sauvola technique and Otsu's method is used.

2. Text and Non-text classification - The main stage of text and non-text classification in the MHS-2017 system is the Minimum Homogeneity Algorithm (MHA) which was first introduced in 2016 [17]. This algorithm based on the connected component analysis [18] in a statistical approach. In 2017, an essential update in the core of this algorithm, the MLL classification [19] which uses the combination of Multilevel and Multilayer homogeneity structure, is presented.

3. Text segmentation and image classification - In this step, text documents are segmented to get text regions, and non-text elements are classified into different types. A combination of text line extraction, paragraph segmentation, and adaptive mathematic morphology is applied to get text regions [19]. Based on properties of non-text elements, they are classified into negative-text region, line, table, separator, and image. The system also contains a robust table detection method which was introduced by Tran et al. [20] in 2016.

4. Region refinement and labelling - Based on the boundary of each region, the rectangular shapes of text and non-text regions are extracted. All of the identified regions are labelled (heading, page number, etc.) based on their text size and position.

5. Optical Character Recognition - All text regions are then recognized via Tesseract OCR in Computer Vision System Toolbox™ (Matlab).

D. The AOSM Method

This method was submitted by Ha Dai-Ton from Ha Long High School for Gifted Student, Ha Long City, Vietnam.

AOSM (Adaptive Over-Split and Merge algorithm) [21] is a hybrid page segmentation method combined top-down and bottom-up approaches. It firstly over-segments page image using a set of white-spaces covering the whole document background. It then groups over-segmented text regions using adaptive parameters. Finally, local context analysis sub-divides (under-segmented) text regions into paragraphs.

For the white-spaces detection, after connected components are detected and filled, a set of white-spaces covering the whole document background is determined using WhiteSpace algorithm [22].

E. *The JU_Aegean Method*

This method was submitted by S. Bhowmik, S. Kundu, B. Kumar De, R. Sarkar, and M. Nasipuri from Jadavpur University (India) and N. Vasilopoulos and E. Kavallieratou from the University of the Aegean (Greece).

Starting point is a combination of pre-processing steps, including conversion to greyscale, contrast stretching, region filling, and binarisation before connected components are analysed to eliminate separators and margins. The resulting image is segmented into horizontal segments based on sufficiently thick horizontal white space. A further step produces an image I_{large} without small components by applying a morphological closing operation with a dynamically chosen structuring element to each horizontal segment. Based on this and the binarised image, a corresponding image I_{small} containing only small components is obtained.

The actual region segmentation process is now started on I_{small} using iterative morphological dilation. The dimension of the structuring element is changed in each epoch based on the size of the connected components present in the image generated during the previous iteration. In addition, during each epoch dilation is performed twice (with 0° and 90° rotation of the structuring element). This process is continued until the number of components present in the currently generated image is reduced to a certain threshold.

Text/non-text separation is then performed on I_{large}. The components are classified as text or non-text on the basis of *solidity*, *perimeter*, *area*, *aspect ratio*, *number of neighbours* etc. of the component under consideration. To arrive at valid regions, the image containing all text is segmented in the same way as I_{small} before.

The final result is produced by combining all text segments obtained from I_{large} with the segmentation result originating from I_{small} to represent text regions as well as labelling all non-text regions from I_{large} accordingly.

VI. Results

Evaluation results for the above methods are presented in this section in the form of graphs. For comparison purposes, the layout analysis and recognition components of a leading product, ABBYY FineReader® Engine 11, and that of the popular open-source system, Tesseract 3.04 are also included. It must be noted that FineReader and Tesseract have been evaluated with no prior training or knowledge of the dataset.

Three scenarios have been defined for the competition, each with a corresponding evaluation profile. The first profile is used to measure the pure segmentation performance. Therefore, misclassification errors are ignored completely. Miss and partial miss errors are considered worst and have the highest weights. The weights for merge and split errors are set to 50%, whereas false detection, as the least important error type, has a weight of only 10%. Results for this profile are shown in Figure 4.

The second profile ("Segmentation + Classification") also evaluates region classification, in the context of a typical OCR system, focusing on text but not ignoring the non-text regions. Accordingly, this profile is similar the first but misclassification of text is weighted highest and all other misclassification weights are set to 10%. Results for this profile are shown in Figure 4.

The third profile ("Text regions only") is based on the previous profile but focuses solely on text, ignoring non-text regions. Results for this profile are shown in Figure 4. A breakdown of the layout analysis errors made by each method (Segmentation + Classification scenario) is given in Figure 5.

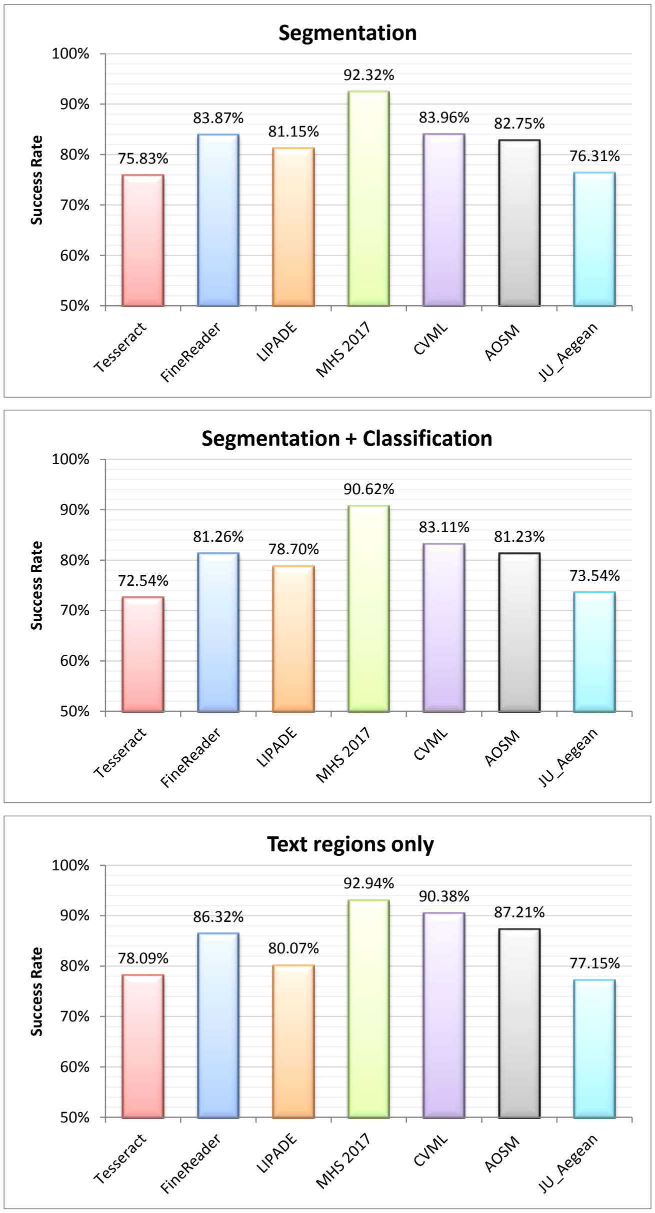


Figure 4. Results using three different evaluation profiles.

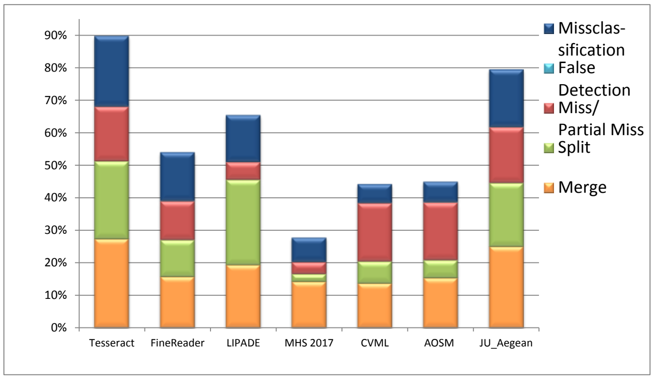


Figure 5. Breakdown of errors made by each method.

Text recognition was a bonus challenge and only submitted by the MHS team. Figure 6. shows the results using the bag of words evaluation method and Figure 7. shows the character accuracy. The better layout analysis method of MHS, in comparison to Tesseract's native approach, leads to better OCR performance. FineReader performs best in this setup.

One interesting observation can be made from the differences of traditional character accuracy and flex character accuracy. The closer the two values, the closer the reading order of the OCR result to the ground truth. However, this does not necessarily reflect the exact reading order detection performance because of the complexity of the material (ambiguous order).

VII. Concluding Remarks

Despite the increased difficulty (e.g. table regions), good page segmentation results were achieved by the submitted methods. For general segmentation and region classification, the MHS 2017 method is a clear winner. The competitors are much closer together when looking at the text-regions-only scenario. All methods outperform the open source state-of-the-art method and the top three also outperform ABBYY FineReader with regard to layout analysis.

The breakdown by error type is particularly useful to identify shortcomings of an algorithm. Mergers represent the largest problem across all methods. False detection, on the other hand, is negligible. By addressing miss error, CVML and AOSM could catch up with the winner.

The OCR evaluation clearly shows that the page layout analysis has an impact on the text recognition performance. The MHS team use Tesseract but achieve better results than the standalone Tesseract.

All submitted methods use a multi-step approach including rule-based decisions and, in many cases, connected components or a variation thereof. Neural nets have not (yet?) taken centre stage for page analysis.

Although progress has been made, page analysis for complex layouts is an unsolved problem and further research is needed to reach success rates close to 100%, as is common for text recognition for contemporary documents.

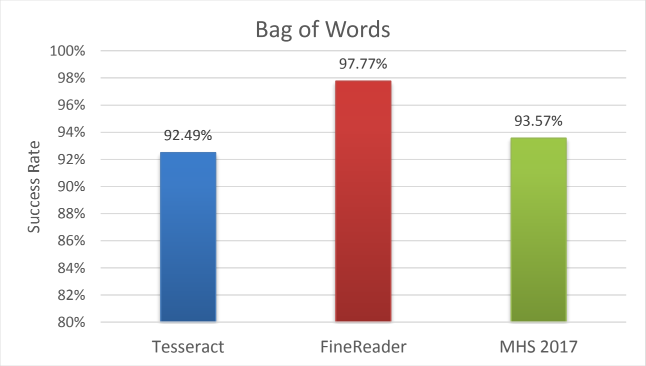


Figure 6. OCR evaluation result using bag of words method.

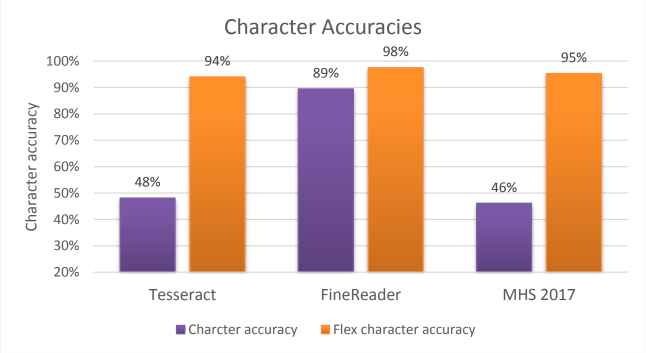


Figure 7. OCR evaluation result using character accuracy methods.

References

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

[2] 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.

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

[4] A. Antonacopoulos, D. Bridson, C. Papadopoulos and S. Pletschacher, "A Realistic Dataset for Performance Evaluation of Document Layout Analysis", *Proc. ICDAR2009*, Barcelona, Spain, July 2009, pp. 296-300.

[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, "ICDAR2015 Competition on Recognition of Documents with Complex Layouts – RDCL2015", *Proc. ICDAR2015*, Nancy, France, Aug 2015, pp. 1151-1155.

[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, S. Pletschacher and A. Antonacopoulos, "Scenario Driven In-Depth Performance Evaluation of Document Layout Analysis Methods", *Proc. ICDAR2011*, Beijing, China, Sept 2011.

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

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

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

[13] H.I.Koo and N.I.Cho “State estimation in a document image and its application in text block identification and text line extraction” *in ECCV*, volume 6312, pages 421–434, 2010.

[14] H. I. Koo, "Text-Line Detection in Camera-Captured Document Images Using the State Estimation of Connected Components," in *IEEE Trans Image Proc*, vol. 25, no. 11, pp. 5358-5368, Nov. 2016.

[15] R. von Gioi, J. Jakubowicz, J.-M. Morel and G. Randall “LSD: A fast line segment detector with a false detection control”, *IEEE Trans PAMI*, 32(4):722–732, April 2010.

[16] H.I. Koo, Y.K. Baik, and B.S. Kim, *Efficient blending methods for ar applications*, Jan. 3 2013. US Patent App. 13/283,462.

[17] T. A. Tran, I. S. Na, S. H. Kim, “Page Segmentation using Minimum Homogeneity Algorithm and Adaptive Mathematical Morphology,” *Int J on Doc Anal and Rec*, vol. 19, pp. 191-209, 2016.

[18] T. A. Tran, I. S. Na, S. H. Kim, “Separation of Text and Non-text in Document Layout Analysis using a Recursive Filter,” *KSII Transactions on Internet and Information Systems*, vol. 9, pp. 4072-4091, 2015.

[19] T. A. Tran, I. S. Na, S. H. Kim, “A Robust System for Document Layout Analysis using Multilevel Homogeneity Structure,” *Expert Systems With Applications*, vol. 85, pp. 99-113, 2017.

[20] T. A. Tran, H. T. Tran, I. S. Na, S. H. Kim, G. S. Lee, H. J. Yang, “A Mixture Model using Random Rotation Bounding Box to Detect Table Region in Document Image,” *Int J Vis Comm and Image Representation*, vol. 39, pp. 196-208, 2016.

[21] Dai-Ton, Ha, Nguyen Duc-Dung, and Le Duc-Hieu. "An adaptive over-split and merge algorithm for page segmentation." *Pattern Recognition Letters* 80 (2016): 137-143.

[22] Breuel, Thomas M. "Two geometric algorithms for layout analysis." *Int workshop on Doc Anal Systems*. Springer Berlin Heidelberg, 2002.