cBAD: ICDAR2019 Competition on Baseline Detection

Markus Diem, Florian Kleber, and Robert Sablatnig
Computer Vision Lab
TU Wien
Vienna, Austria
{*diem, kleber, sab*}*@cvl.tuwien.ac.at*

Basilis Gatos
Computational Intelligence Laboratory
National Center for Scientific Research Demokritos
Athens, Greece
bgat@iit.demokritos.gr

***Abstract*—Baseline detection is a simplified text-line extraction that typically serves as pre-processing for Automated Text Recognition. The cBAD competition benchmarks state-of-the-art baseline detection algorithms. It is the successor of cBAD 2017 with a larger dataset that contains more diverse document pages. The images together with the manually annotated groundtruth are made publicly available which allows other teams to benchmark and compare their methods. We could also evaluate the winning method of cBAD 2017 on the newly introduced dataset which now serves as baseline. This competition shows that the performance of automated baseline detection increased substantially since 2017.**

***Keywords*-cBAD, baseline detection; text-line extraction; document analysis; competition;**

I. Introduction

The performance of layout analysis systems is crucial because they are typically used as pre-processing steps for other document analysis tasks such as Automated Text Recognition (ATR). Hence, errors from this stage are propagated to all subsequent stages. There were three competitions dedicated to layout analysis [1], [2], [11], [6] and two competitions dedicated to baseline detection [3], [11] organized in conjunction with ICDAR 2017 which indicates an active interest of the document analysis community in benchmarking layout analysis systems. The performance of benchmarked methods shows on the one hand that we could substantially improve layout analysis systems by utilizing Deep Neural Networks (DNN) for this task. On the other hand, there is still a need for improving Document Image Analysis (DIA) systems - especially when it comes to generalizing the models for different tasks.

The cBAD: ICDAR2019 competition on BAseline Detection benchmarks automated baseline extraction. It thus continues the successful cBAD: ICDAR2017 competition on BAseline Detection [3] that targets text extraction as pre-processing step for Automated Text Recognition (ATR). We therefore keep the tradition of text line extraction competitions but the challenging dataset with more than 3021 page images sets new benchmarking standards.

Despite of this challenging dataset, the competition attracted four teams from across Europe and China. We present the dataset, the evaluation scheme, and the competition protocol in the next section. The teams present their method in Section III and the respective results are presented in Section IV. Section V concludes this paper.

II. The Competition

The competition is organized using ScriptNet[1]. The training and evaluation sets together with the groundtruth are published at Zenodo[2] which serves as sneak preview and for training supervised approaches. We again use the PAGE XML Schema [9] which is well-established in the document analysis community and best serves our needs. A minimal PAGE XML sample is shown in Listing 1. The baseline evaluation scheme that was introduced in conjunction with the last cBAD [3] is deployed to assess the performance of methods submitted. The full dataset including ground truth annotation of the test set was publicly released after the submission deadline using Zenodo. Zenodo is chosen because it promises long term preservation, it allows for versioning (which improves comparability of results), and it creates a DOI that can be used for citing independent to the databases actual location.

A. Dataset

We sampled 3028 document images from 175567 images using a freely available python script[3] that guarantees the data to be sampled uniformly. The 3028 document images thus sampled were annotated by DigiTexx[4]. Aftwards, the GT was inspected by two independent operators who removed 7 images because of wrong baseline annotations resulting in a final dataset size of 3021. The dataset is split into a train set with 755 $(= 25\%)$ images, an evaluation set with 755 $(= 25\%)$ images and a test set with 1511 $(= 50\%)$ images. Participants were provided with the groundtruth of the train and evaluation sets. They had to run their method on all images of the test set whose groundtruth was not published until after the submission deadline.

[1]https://scriptnet.iit.demokritos.gr/competitions/
[2]https://doi.org/10.5281/zenodo.2567397
[3]https://github.com/TUWien/Benchmarking
[4]https://digi-texx.vn/en

2379-2140/19/$31.00 ©2019 IEEE
DOI 10.1109/ICDAR.2019.00240

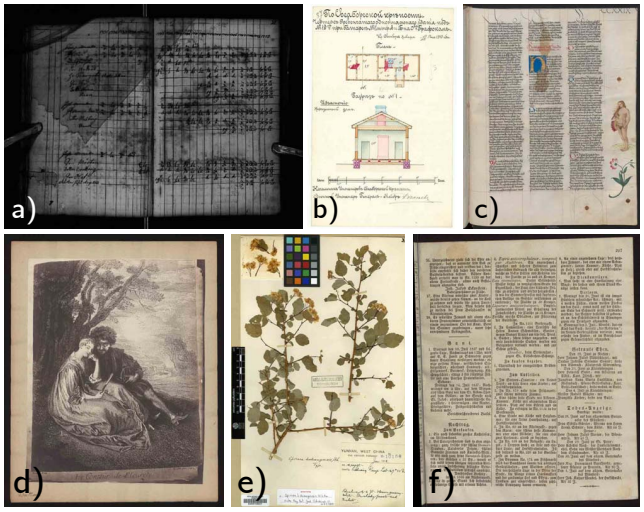

Figure 1. Sample images from different collections. The database consists of document pages with different layouts and origins. There are heavily structured pages (a); sparsely inscribed pages (b), (d), and (e); drawings (b) and engravings (d); and printed documents (f).

The dataset contains mainly archival documents with latin text written in different languages. Figure 1 shows six exemplary images. The collection contains tables (a), drawings (b), medieval handwriting (c), engravings (d), pages of the George Forrest Herbarium (e) and historical prints (f).

B. Groundtruth

For benchmarking baseline detection approaches, a baseline is defined in the typographical sense as the virtual line where all characters rest upon and descenders extend below [3]. The objective of this benchmark is to asses the quality of baseline detection methods in the wild. In contrast to the previous cBAD, we do not limit the data to handwritten archival documents. Hence, narrow columns in printed documents (see Figure 1 (f)) now pose a particular challenge. Additionally, sparsely inscribed pages such as (b), (d), or (e) test the methods' capabilities of locating text in images.

All documents were manually annotated and saved as PAGE XML. A pixel accurate location of baselines is not necessary since the evaluation scheme permits small deviations. Figure 2 shows an annotated image. Baselines, which are drawn pink are used for evaluation. They can have an arbitrary orientation and local skew. In other words, baselines are not necessarily straight lines.

C. Evaluation Scheme

The baseline evaluation scheme that was used at the last cBAD [3] is used again. This scheme utilizes a coverage function that aligns hypothesis baselines (HY) with groundtruth baselines (GT). The R-value then indicates

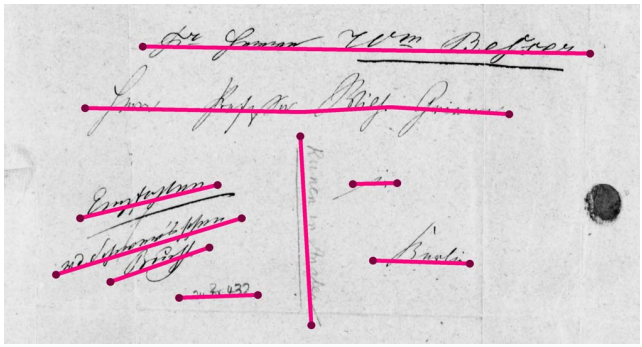

Figure 2. Groundtruth example. Baslines (pink) are annotated and evaluated.

(similar to the well-known recall) the amount of GT lines that have corresponding HY baselines while the P-value penalizes segmentation errors (similar to precision). An F-value is computed per page as the harmonic mean of the P and R. Participating methods are ranked with respect to the average F-value of the page-wise results.

The baseline evaluation is implemented in Java and publicly available[5] as a standalone command line tool licensed under LGPLv3. A detailed explanation of the evaluation scheme can be found in [7].

III. Participants

As previously mentioned, the competition was carried out using ScriptNet. Teams could download the training and evaluation sets along with GT and all images of the test set. For evaluation, teams uploaded the resulting PAGE XMLs (one per image) which were evaluated in ScriptNet. Registered teams were able to see the results of their submissions (but not those of other teams). The number of submissions was not limited and results presented in this paper represent the best submission per team.

Methods of four different teams were submitted. A short method description provided by the participating teams is given below. They are listed in alphabetical order.

A. DMRZ

Georg Mackenbrock, Michael Fink, Thomas Layer, Michael Sprinzl
Deutsches Medizinrechenzentrum GmbH & Co KG, Vienna, Austria
mackenb@dmrz.de

Our submission to the ICDAR 2019 cBAD Competition utilizes deep convolutional nets and is a follow-up on the method presented in [5]. Compared to the latter, we (i) scale the input image to a fixed width, (ii) directly apply a residual U-net (BL-net) to detect baseline candidates, which (iii) uses larger windows and (iv) is trained with auxiliary error layers aiming at the detection of starting point and

[5]https://github.com/Transkribus/TranskribusBaseLineEvaluationScheme

Listing 1. Minimal sample of a PAGE XML.

```
<?xml version="1.0" encoding="UTF-8" standalone="yes"?>
<PcGts
  xmlns="http://schema.primaresearch.org/PAGE/gts/pagecontent/2013-07-15"
  xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  xsi:schemaLocation="http://schema.primaresearch.org/PAGE/gts/pagecontent/2013-07-15
  http://schema.primaresearch.org/PAGE/gts/pagecontent/2013-07-15/pagecontent.xsd">
    <Metadata>
      <Creator>CVL</Creator>
      <Created>2018-11-29T08:46:03Z</Created>
      <LastChange>2018-11-30T10:18:12Z</LastChange>
    </Metadata>
<Page imageFilename="document.tif" imageWidth="2959" imageHeight="4332">
  <TextRegion id="R0" type="Handwritten">
    <Coords points="2401,228 2647,228 2647,399 2401,399"/>
       <TextLine id="L0">
         <Coords points="2439,306 2574,310 2573,360 2438,356"/>
         <Baseline points="2438,351 2573,355"/>
       </TextLine>
  </TextRegion>
</Page>
</PcGts>
```

end point for baselines in addition the baselines themselves. Moreover, a further convolutional net has been trained on images extracted homogeneously around detected candidate baselines in order to classify them as either to be kept or to be pruned from the candidate set. Deleting those baseline candidates whose corresponding images are classified as to be pruned yields a final set of detected baselines.

B. Planet

Tobias Grüning (1) and Max Weidemann (2)
Planet AI GmbH (1), CITlab University of Rostock (2)
tobias.gruening@planet.de

The proposed baseline detection system is composed of a CNN-based orientation estimation, an ARU-Net to detect baselines/separators and an image processing based methodology to extract the baselines in a parametrized form. The orientation CNN is trained to estimate the predominant text orientation modulo $90°$ ($0°, 90°, 180°$ or $270°$) in the current document page. After correcting the detected orientation, the trained ARU-Net generates maps which encode the probability of the presence of the classes baseline and separator. These maps are utilized in an image processing based super pixel clustering process to estimate the baselines in form of polygonal chains. The details of this approach are described in [8]. The proposed system differs slightly from [8], e.g., a different training scheme for the ARU-Net is used. Finally, the hypothesis baselines are evaluated with respect to their confidence, quantity as well as positions to handle false positives. Our different submissions basically differ in the orientation as well as ARU-Nets.

C. TJNU

Rubo Bai, and Yuanping Zhu
Tianjin Normal University
zhuyuanping@tjnu.edu.cn

Our baseline detection method is based on a FCN model which consists of 7 convolution layers. Reffering the first 6 layers of VGG16, our network uses the same size and numbers of filters. However, our network uses the dilated convolutions instead of pooling. Dilated convolution is used to enlarge receptive fields and detect text line effectively without large context. The first two layers are standard convolutions with a dilation of 1, then two layers with a dilation of 2 and two layers with a dilation of 4. Finally, an output layer is added to get predictions with a dilation of 1 and a filter with size of 1. The network can predict the binary mask of pixels which are in a small 3-pixel radius of the training baselines. Moreover, on-the-fly data augmentation strategy is adopt in the training.

D. UPVLC

Lorenzo Quirós, Moisés Pastor-i-Gadea and Jose R. Prieto
Universitat Polotècnica de València, Pattern Recognition and Human Language Technology group (PRHLT), Spain.
lorenzoqd@gmail.com

The method submitted to this competition is a two stage method based on our previous works presented in [10] and [4]. The first stage is an Conditional Generative Adversarial Neural Network (CGANN) trained to estimate the probability of each pixel in the input image to belong to a baseline, as explained in [10]. The second stage is a modified version of [4] where probability map computed on the first stage is used

to estimate a set of interest points, then DBScan algorithm is used for clustering those points into a set of baselines. The images are first resized to 1024x768 pixels in order to constrain the computational resources required. CGANN architecture and main hyper-parameters are the same used for experiments presented in [10], and trained during 300 epochs with a batch-size of 6 images.

E. Baseline (Winner of cBAD 2017)

Georg Mackenbrock, Michael Fink, Thomas Layer, Michael Sprinzl
Deutsches Medizinrechenzentrum GmbH & Co KG, Vienna, Austria
mackenb@dmrz.de

The winning method of cBAD 2017 was submitted by Fink et al [5]. Their approach utilized a CNN for classifying basic document properties (i.e. text regions) and a second CNN that detects and extracts baselines (see also [3]). The authors tested their 2017 version on the newly introduced dataset which allows us to compare both datasets and to determine the performance increase of automated baseline detection methods since 2017.

IV. Results

The evaluation is carried out with the aforementioned evaluation scheme on all 1511 document images of the test set. The median F-value of all 29 submissions is 0.90. We see an increased F-value of 14% when compared to cBAD 2017 Track B [3] which was 0.76. This indicates an impressive improvement of state-of-the-art baseline detection methods considering that the dataset of cBAD 2019 is more challenging which we will demonstrate later in this section.

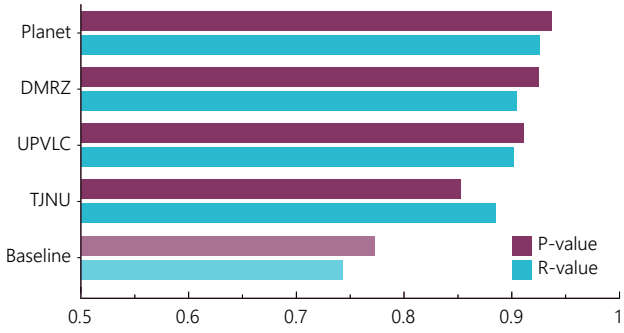

Figure 3. P-value (upper bar) and R-value (lower bar) of all participating methods when evaluated on the test set. The winning method of cBAD 2017 (baseline) is also evaluted on the newly introduced dataset to show the overall performance increase since the last two years.

Figure 3 shows the P-value and R-value of each team's best submission. The results are sorted with respect to the F-value which is the harmonic mean of P-value and R-value and used for ranking the approaches. The best performance with an F-value of 0.931 is achieved by *Planet* (see Table I). The methods submitted by *DMRZ* and *UPVLC* achieve comparably good results (F-value = 0.915 and F-value = 0.907 respectively). *DMRZ* benchmarked their winning method of 2017 on the cBAD 2019 dataset. These results serve as baseline to show the progress since 2017. Table I shows that the winning method of cBAD 2019 achieves a performance increase of 17.3% when compared to the winner of cBAD 2017 using the same dataset.

Method	P-value	R-value	F-value	Rank
Planet	0.937	0.926	0.931	1
DMRZ	0.925	0.905	0.915	2
UPVLC	0.911	0.902	0.907	3
TJNU	0.852	0.885	0.868	4
Baseline	0.773	0.743	0.758	-

Table I
RESULTS: P-VALUE (PSEUDO PRECISION), R-VALUE (PSEUDO RECALL), AND F-VALUE (PSEUDO F-SCORE) OF ALL PARTICIPATING METHODS. THE METHODS ARE RANKED WITH RESPECT TO THE F-VALUE.

In order to analyze the newly introduced dataset, the results achieved by the best performing method of cBAD 2017 is compared with the previous datasets in Figure 4. The performance drops significantly if the new dataset is used for benchmarking which indicates that the cBAD 2019 dataset is more diverse and challenging than its predecessor.

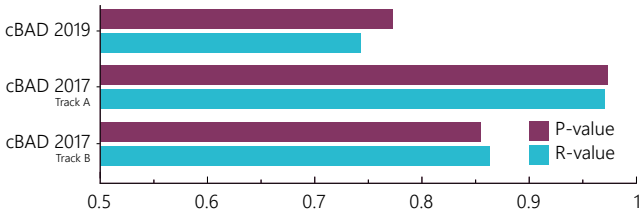

Figure 4. This figure shows a dataset comparison. P-value and R-value of the baseline method (winner of cBAD 2017) when evaluated using the different cBAD datasets.

V. Conclusion

The ICDAR 2019 competition on Baseline Detection was organized as successor of cBAD 2017 with more challenging data. While three out of five teams utilized deep learning in 2017, all participating methods of this years competition make use of deep learning. Two of which propose using a hybrid architecture with U-Nets and traditional post-processing. *UPVLC* are the first who utilize GANs for the task of baseline detection with quite promising results. Tobias Grüning (*Planet*) and Max Weidemann (*CITlab University of Rostock*) submitted the best performing method which achieves an F-value of 0.931. The results also show that significant progress was made in the field of baseline detection. New methods are capable of correctly detecting baselines even if documents with different modalities are presented.

We keep the submission system open on ScriptNet which allows for comparing new methods with the presented ones. Moreover, the dataset along with the groundtruth is publicly available which should stimulate future development in the context of baseline detection.

REFERENCES

[1] Christian Clausner, Apostolos Antonacopoulos, Tom Derrick, and Stefan Pletschacher. ICDAR2017 competition on recognition of early indian printed documents - REID2017. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*. IEEE, nov 2017.

[2] Christian Clausner, Apostolos Antonacopoulos, and Stefan Pletschacher. ICDAR2017 competition on recognition of documents with complex layouts - RDCL2017. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*. IEEE, nov 2017.

[3] Markus Diem, Florian Kleber, Stefan Fiel, Tobias Gruning, and Basilis Gatos. cBAD: ICDAR2017 competition on baseline detection. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*. IEEE, nov 2017.

[4] Ahmed Fawzi, Moisés Pastor, and Carlos D. Martínez-Hinarejos. Baseline detection on arabic handwritten documents. In Kenneth P. Camilleri and Alexandra Bonnici, editors, *Proceedings of the 2017 ACM Symposium on Document Engineering, DocEng 2017, Valletta, Malta, September 4-7, 2017*, pages 193–196. ACM, 2017.

[5] Michael Fink, Thomas Layer, Georg Mackenbrock, and Michael Sprinzl. Baseline detection in historical documents using convolutional u-nets. In *13th IAPR International Workshop on Document Analysis Systems, DAS 2018, Vienna, Austria, April 24-27, 2018*, pages 37–42. IEEE Computer Society, 2018.

[6] Liangcai Gao, Xiaohan Yi, Zhuoren Jiang, Leipeng Hao, and Zhi Tang. ICDAR2017 competition on page object detection. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*. IEEE, nov 2017.

[7] Tobias Grüning, Roger Labahn, Markus Diem, Florian Kleber, and Stefan Fiel. READ-BAD: A new dataset and evaluation scheme for baseline detection in archival documents. *CoRR*, abs/1705.03311, 2017.

[8] Tobias Grüning, Gundram Leifert, Tobias Strauß, and Roger Labahn. A two-stage method for text line detection in historical documents. *CoRR*, abs/1802.03345, 2018.

[9] Stefan Pletschacher and Apostolos Antonacopoulos. The PAGE (page analysis and ground-truth elements) format framework. In *2010 20th International Conference on Pattern Recognition*. IEEE, aug 2010.

[10] Lorenzo Quirós. Multi-task handwritten document layout analysis. *CoRR*, abs/1806.08852, 2018.

[11] Fotini Simistira, Manuel Bouillon, Mathias Seuret, Marcel Wursch, Michele Alberti, Rolf Ingold, and Marcus Liwicki. ICDAR2017 competition on layout analysis for challenging medieval manuscripts. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*. IEEE, nov 2017.