ICPR 2020 Competition measure: Text Block Segmentation on a NewsEye Dataset

Bastian Laasch Max Weidemann Johannes Michael Roger Labahn

CITlab, Institute for Mathematics
University of Rostock
{first name}.{surname}@uni-rostock.de

March 17, 2020

Our final goal is the extraction of text from images for further analysis. Therefore, in terms of segmentation it is sufficient to merge the lines into blocks and not to detect regions on pixel level. Thus, leading to the introduction of a new evaluation measure (see our GitHub repository[1]).

1 Notation

Since we have to get more technical in this subsection, we want to define some notations. A baseline is described by a polygonal chain, i.e., a list of a finite number of ordered two dimensional points (= vertices of the chain), so it is to be understood as a vector. Furthermore, to evaluate the quality of our system, we have to compare the results of the Text Block Segmentation (TBS) algorithm (= hypotheses (HY)) with the ground-truth (GT) data. Per page we define:

- $\mathbf{g}_k$ is the GT baseline with index k given by human annotators, $k \in \{1, \ldots, K\}$, where K is the number of all GT baselines of the page.
- $\mathbf{h}_l$ is the HY baseline with index l computed by a Baseline Detection (BD) system, $l \in \{1, \ldots, L\}$, where L is the number of all HY baselines of the page.
- $G_i = \left\{\mathbf{g}_{i_1}, \ldots, \mathbf{g}_{i_{m_i}}\right\}$ is the GT text block with index i as a set of m_i GT baselines, $i \in \{1, \ldots, M\}$, where M is the number of all GT blocks of the page.

[1] https://github.com/CITlabRostock/citlab-article-separation-measure

- $H_j = \left\{\mathbf{h}_{j_1}, \ldots, \mathbf{h}_{j_{n_j}}\right\}$ is the HY text block with index j as a set of n_j HY baselines, $j \in \{1, \ldots, N\}$, where N is the number of all HY blocks of the page.

Remark 1. The measure is developed for an end-to-end system detecting the lines in a first step and afterwards the text blocks. In connection with this competition the BD is not needed since the baselines are given, i.e., the set of HY baselines is equal to the set of GT baselines. For the sake of completeness, however, we would like to present the entire measure.

2 R and P matrices

In the following we want to compare the given M GT text blocks with the generated N HY blocks. To this end we compute two different types of evaluation scores between every GT block and every HY block. Hence, we get two matrices of dimension $M \times N$.

At this point the BD measure presented in [1] is used, which is composed of the so called R and P value helping us to study the equality between two sets of baselines. This measure was employed successfully in the recent ICDAR 2017[2] / 2019[3] competitions on BD.

- The **R value** $\in [0, 1]$ (see [1], section 3) indicates, loosely spoken, how well a set of GT baselines is covered by a set of HY baselines. Hence, this score has similar properties to the well-known recall value.
  $\Rightarrow$ Segmentation errors, e.g., a baseline is split into two lines or two lines are merged into one, are not penalized, because we measure how reliable the text is detected (ignoring layout issues).

- The **P value** $\in [0, 1]$ (see [1], section 3) indicates, loosely spoken, how well a set of HY baselines is covered by a set of GT baselines. Hence, this score has similar properties to the well-known precision value.
  $\Rightarrow$ Segmentation errors are penalized, because we measure how reliable the structure of the baselines (layout) of the page is detected. So, this score gives us information about the over- and under-segmentation of lines.

Definition 1 (R matrix)**.** The R matrix $\boldsymbol{\mathcal{R}}\left(\{G_1, \ldots, G_M\}, \{H_1, \ldots, H_N\}\right)$ between the GT blocks $G_i, i \in \{1, \ldots, M\}$, and the HY blocks $H_j, j \in \{1, \ldots, N\}$, is defined as

[2] https://scriptnet.iit.demokritos.gr/competitions/5/
[3] https://scriptnet.iit.demokritos.gr/competitions/11/

$$
\begin{array}{c c}
 & \begin{array}{ccccc} H_1 & \cdots & H_j & \cdots & H_N \end{array} \\
\begin{array}{c} G_1 \\ \vdots \\ G_i \\ \vdots \\ G_M \end{array} & \left[\begin{array}{ccccc} & & & & \\ & & & & \\ & & \widetilde{R}\left(G_i, H_j, \mathcal{T}_i\right) & & \\ & & & & \\ & & & & \end{array} \right]
\end{array}
$$

in which $\widetilde{R}\left(G_i, H_j, \mathcal{T}_i\right)$ is the in [1] defined R value. The set $\mathcal{T}_i$ contains tolerance values for each GT baseline included by G_i effecting that minor deviations in the BD are not penalized.

Definition 2 (P matrix)**.** The P matrix $\boldsymbol{\mathcal{P}}\left(\{G_1, \ldots, G_M\}, \{H_1, \ldots, H_N\}\right)$ between the GT blocks $G_i, i \in \{1, \ldots, M\}$, and the HY blocks $H_j, j \in \{1, \ldots, N\}$, is defined as

$$
\begin{array}{c c}
 & \begin{array}{ccccc} H_1 & \cdots & H_j & \cdots & H_N \end{array} \\
\begin{array}{c} G_1 \\ \vdots \\ G_i \\ \vdots \\ G_M \end{array} & \left[\begin{array}{ccccc} & & & & \\ & & & & \\ & & \widetilde{P}\left(G_i, H_j, \mathcal{T}_i\right) & & \\ & & & & \\ & & & & \end{array} \right]
\end{array}
$$

in which $\widetilde{P}\left(G_i, H_j, \mathcal{T}_i\right)$ is the in [1] defined P value. The set $\mathcal{T}_i$ contains tolerance values for each GT baseline included by G_i effecting that minor deviations in the BD are not penalized.

3 R, P and F value for Text Block Segmentation

After the calculation of the R and P matrix based on an evaluation scheme for BD, we determine the maximum entries in these matrices in a greedy manner.

Remark 2. "Greedy manner" means, in this context, that one by one the maximal values of a given matrix are chosen with following deletion of the corresponding rows and columns. Afterwards, the resulting values are summed up (see Algorithm 1).

Furthermore, it seems to make sense to create a monotony between the BD and the proposed TBS measure that holds

$$
\text{proposed TBS measure} \leq \text{BD measure} . \tag{1}
$$

Algorithm 1 Greedy Function

procedure GREEDY($\boldsymbol{A}$) with $\boldsymbol{A} \in \mathbb{R}^{M \times N}$
  Sum $\leftarrow 0$
  $\boldsymbol{A}' \leftarrow \boldsymbol{A}$
  while $\boldsymbol{A}'$ is not empty **do**
    $a \leftarrow$ one of the maximal elements of $\boldsymbol{A}'$
    Sum $\leftarrow$ Sum $+ a$
    $\boldsymbol{A}' \leftarrow$ take $\boldsymbol{A}'$ and delete corresponding row / column of the element a
  end while
  return Sum
end procedure

The equality holds if and only if the blocks have been found perfectly. Later, this relation is useful to realize in which step in an end-to-end scenario the mistakes happened (in the BD or in the TBS step). Please note again that in the proposed competition the baselines are given and hence the BD measure is always 1.

To ensure (1), each row of the R matrix is weighted by the percentage of the GT block (compared to all GT blocks) corresponding to this row (analogous, each column of the P matrix is weighted by the percentage of the HY block (compared to all HY blocks) corresponding to this column).

For example, we consider a page with three given GT blocks G_1, G_2, G_3 and two detected HY blocks H_1, H_2. Hence, the R matrix has the dimension 3×2. The set G_1 has 10 baselines and the sets G_2 and G_3 contain together 30 baselines. Therefore, the first row in the R matrix (this row corresponds to the GT block G_1) is multiplied by $1/4$, since the block G_1 includes 25% of all GT baselines assigned to blocks.

After the explained multiplication / weighting step, the above described Greedy Function is applied on the resulting matrices. We want to express this process with

$$\mathrm{GREEDY}_{\mathrm{weighted}}\left(\boldsymbol{\mathcal{R}}\right) \qquad \text{or} \qquad \mathrm{GREEDY}_{\mathrm{weighted}}\left(\boldsymbol{\mathcal{P}}\right) \ .$$

Definition 3 (R, P and F value for Text Block Segmentation)**.** The R and P value $\in [0, 1]$ for the generated HY blocks are defined as

$$R\left(\{G_1, \ldots, G_M\}, \{H_1, \ldots, H_N\}\right) := \mathrm{GREEDY}_{\mathrm{weighted}}\left(\boldsymbol{\mathcal{R}}\right),$$

$$P\left(\{G_1, \ldots, G_M\}, \{H_1, \ldots, H_N\}\right) := \mathrm{GREEDY}_{\mathrm{weighted}}\left(\boldsymbol{\mathcal{P}}\right).$$

Thus, we obtain the F value $\in [0, 1]$ for TBS, i.e., the harmonic mean of the R and P value,

$$F\left(\{G_1, \ldots, G_M\}, \{H_1, \ldots, H_N\}\right) := \frac{2 \cdot R \cdot P}{R + P} \ .$$

The target value is 1 in all three cases.

These three values give us an appropriate tool to evaluate the result of an algorithm merging a given set of detected baselines into text blocks. The R and P values ensure that HY blocks with too many or too few baselines in comparison with the corresponding GT blocks are penalized with a lower evaluation score.

To identify the winner of the proposed tasks of the competition, the F values will be averaged over all test samples.

References

[1] 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". In: CoRR abs/1705.03311 (2017). arXiv: 1705.03311. URL:http://arxiv.org/abs/1705.03311.