

# Baseline Detection in Historical Documents using Convolutional U-Nets

Michael Fink, Thomas Layer, Georg Mackenbrock, and Michael Sprinzl

*Deutsches Medizinrechenzentrum GmbH & Co KG*

*Vienna, Austria*

*Email: {fink,layer,mackenb,sprinzl}@dmrz.de*

**Abstract**—Baseline detection is still a challenging task for heterogeneous collections of historical documents. We present a novel approach to baseline extraction in such settings, turning out the winning entry to the ICDAR 2017 Competition on Baseline detection (cBAD). It utilizes deep convolutional nets (CNNs) for both, the actual extraction of baselines, as well as for a simple form of layout analysis in a pre-processing step. To the best of our knowledge it is the first CNN-based system for baseline extraction applying a U-net architecture and sliding window detection, profiting from a high local accuracy of the candidate lines extracted. Final baseline post-processing complements our approach, compensating for inaccuracies mainly due to missing context information during sliding window detection. We experimentally evaluate the components of our system individually on the cBAD dataset. Moreover, we investigate how it generalizes to different data by means of the dataset used for the baseline extraction task of the ICDAR 2017 Competition on Layout Analysis for Challenging Medieval Manuscripts (HisDoc). A comparison with the results reported for HisDoc shows that it also outperforms the contestants of the latter.

**Keywords**—Historical Document Analysis; Baseline Extraction; Deep Neural Networks

## I. INTRODUCTION

Textline segmentation is an essential preprocessing step for handwritten text recognition (HTR) [1], and still a challenging task for heterogeneous collections of historical documents. In particular, if they are of an irregular and complex structure or of low quality such as faint typing. Moreover, ancient documents often exhibit further degradations such as, e.g., bleed-through, holes, ornamentation, annotations, etc. (cf. [2], for instance). Recent competitions, either on textline extraction itself [3], [4] or including it as a subtask towards HTR [5], [6], aim at covering a wide range of these difficulties in order to challenge and assess state-of-the-art methods and algorithms.

We present a novel approach to baseline extraction in historical documents that utilizes deep convolutional nets (CNNs) and, to the best of our knowledge, is the first to apply a U-net architecture [7], [8] with sliding window detection. While sliding windows over high resolution images has been a highly successful technique to perform object detection in images with CNNs, it has largely been considered unsuited for the detection of handwritten textlines, where a relatively large number of possibly overlapping objects of the same class need to be extracted.

We show that to the contrary, as far as the extraction of (poly-)baselines is concerned, a CNN-based sliding window approach can be very effective, witnessed for instance

by winning the ICDAR 2017 Competition on Baseline detection (cBAD) [3]. To profit from a high local accuracy of candidate baselines extracted by a CNN (in our case a residual U-net denoted as BL-net), however, our approach incorporates two further components: (i) a second U-net (DA-net), performing simple layout analysis wrt. relevant text regions and auxiliary document properties, is applied in a pre-processing step and utilized, e.g., for individual document pre-scaling; (ii) final baseline post-processing aims at pruning spurious candidate lines and assembling baseline fragments into polygons, thus compensating for inaccuracies mainly due to missing context information during sliding window detection (cf. Fig. 1 for an overview of the overall workflow). Quantifying the effect of the DA-net and post-processing on the overall task is one of the aspects covered by our experimental evaluation. Moreover, we investigate how it generalizes to different data by means of the dataset used for the baseline extraction task of the ICDAR 2017 Competition on Layout Analysis for Challenging Medieval Manuscripts (HisDoc).

Compared to most systems submitted to recent competitions on baseline extraction our approach likewise builds on neural net based machine learning techniques. However, those mainly apply recursive architectures (RNNs) (often using LSTM nodes, e.g., multi-dimensional LSTMs [9]) due to their capabilities to learn keeping track of context information (see, however, [10] for a mixed CNN and RNN architecture for baseline segmentation applied by a contestant of [6]). A notable exception from scene text recognition is [11], where a successful approach towards end-to-end printed text recognition in natural images has been developed by means of a CNN using sliding windows to detect a set of candidate lines of text. Textline segmentation was not the final aim but provided input to a second stage to obtain end-to-end results. Despite its success on printed text recognition, it was deemed unsuited for handwritten text.

More recently, [12] used CNNs for textline extraction also aiming at a robust method explicitly focusing on historical documents. However, textlines are segmented at a different level of detail, returning a surrounding polygon (Main Body Area) at pixel level. Compared to poly-baselines, this usually requires considerable additional human labeling, resp. correction, effort yielding only slightly higher HTR precision [1].

In [13] the challenge of handling high resolution images with a large amount of small objects is addressed using

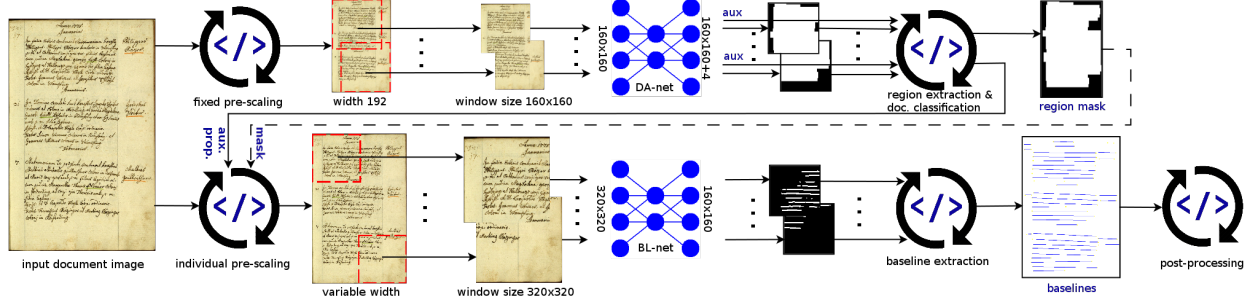


Figure 1. Workflow of our approach including central components (DA-net, BL-net, procedural units) and corresponding data flow.

multiple local prediction models with shared parameters. Textline bounding boxes are obtained detecting lower-left and upper-right corner points in a first stage and matching point pairs in a second stage through local processing. The latter uses small spatial support requiring less resources (parameters). Besides requiring bounding boxes also for training, compared to our approach the maximum number of lines per image is fixed by the architecture and, moreover, 2-dimensional LSTM neurons are added to the local network to include context.

The remainder of this paper is organized as follows. Section II is devoted to simple document analysis, introducing the common overall U-net architecture of the CNNs used, as well as the concrete specification of the DA-net. Actual baseline detection by means of the BL-net is subject to Section III, while baseline post-processing is dealt with in Section IV. In Section V we evaluate the individual components of our approach on the cBAD dataset and compare with the results reported on the HisDoc competition, before we conclude in Section VI.

## II. SIMPLE DOCUMENT ANALYSIS

The CNNs we apply follow a common U-net architecture [7] (see Fig. 2), consisting of a contracting path (left path downward) and an expanding path (right path upward). Downscaling via the contraction path is realized using max-pooling layers after each of several blocks of layers. Corresponding upscaling is achieved using up-sampling via nearest-neighbor interpolation. Horizontal red arrows indicate the usage of equal resolution feature maps from the contracting path as additional input for the expanding path, intended to enhance feature localization. In the following, we provide respective details for the DA-net architecture and describe its application to classify relevant text regions and auxiliary document properties.

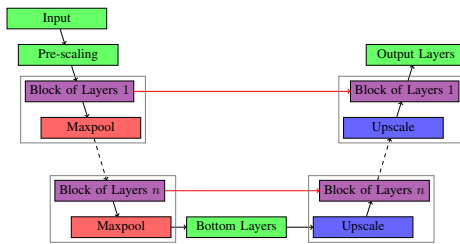


Figure 2. General U-net architecture of DA-net and BL-net.

### A. DA-net Architecture

The DA-net takes as its input images of size  $160 \times 160$  pixels (clippings from document images). No further pre-scaling layers are applied. The corresponding target output (groundtruth) is a binary mask of the same size, representing whether the corresponding pixel belongs to a relevant text-region or not. The contracting path as well as the expanding path consist of four blocks of layers. The first two blocks of layers are composed of four convolution layers with a filter size of  $3 \times 3$ , rectified linear units (ReLU) for activation, and batch normalization. Blocks 3 and 4 follow the same architecture, however, using two convolution layers. Max-pooling layers are applied with  $2 \times 2$  filters and stride 2, thus yielding a downscale by a factor of two. Upscaling doubles the size through bilinear up-sampling, effectively reversing this effect ( $160 \times 160$  output nodes). Conversely, the number of channels is doubled after each block on the contracting path and halved after each block on the expanding path, starting with 32 channels for the first block. Hence, for the bottom layers we obtain  $10 \times 10$  feature maps for 512 channels.

The bottom layers of the DA-net on the one hand connect contracting and expanding path as usual, in our case by means of two convolution layers with batchnorm as above. On the other hand we extended the U-net architecture at this point, adding a fully connected layer with 512 nodes and logistic activation after the last max-pooling layer. This layer is in turn connected to four 2-way softmax layers with corresponding auxiliary error layers (log probability). Intuitively, these layers are used for the classification of auxiliary document properties on the features of the contraction path (four additional outputs).

The block of output layers connected to the expansion path again comprises two convolution layers as above, where the first one reduces to a single channel followed by a batchnorm layer, while the second uses logistic activation with an associated dice-coefficient error layer.

### B. Region Extraction and Document Classification

Inference is realized through a sliding window approach. The input document image is scaled to a fixed width of 192 pixels and a window of size  $160 \times 160$  is applied with stride 40. Note that in this case window size is large wrt. document image size with the intention to learn/predict more global features. Every pixel of a

window yields a prediction. Since windows overlap, we thus get up to 16 predictions for every pixel of the input image (four in every direction). If one is greater than 0.5 the pixel is considered foreground (white), and background (black) otherwise. A region mask is extracted restricting to connected components of at least ten foreground pixels (4-connected). Each region (connected component) in the mask is approximated and represented through a polygon by vertically sampling the coordinates of leftmost and rightmost foreground pixels at a rate of  $1/30$  image height.

Moreover, the auxiliary two-way softmax bottom layers provide four additional classification results, i.e., values from  $[0, 1]$  corresponding to class probabilities for the following document properties:

- *leading* ( $p_{ld}$ ): the distance between successive baselines is large,
- *double-page* ( $p_{dp}$ ): the document spans two pages,
- *landscape* ( $p_{ls}$ ): the orientation is landscape,
- *noText* ( $p_{nt}$ ): the document does not contain text.

These predictions of document properties are used during baseline detection and post-processing, e.g., for scaling.

### III. BASELINE DETECTION

Candidate baselines are detected by means of a different, so-called residual U-net [8], the BL-net. It takes  $320 \times 320$  pixel input images. A  $5 \times 5$  convolution layer with stride 2 and ReLU activation is applied for prescaling. While still predicting the central  $160 \times 160$  pixels (number of output nodes), the intuition is that the net may make use of surrounding context upon downscaling via this layer. Groundtruth thus is again a binary mask, this time representing whether corresponding pixels (central  $160 \times 160$ ) belong to a baseline. Contracting path as well as expanding path consist of five blocks of layers, each composed of three residual blocks: two  $3 \times 3$  convolutional layers with ReLU activation and batch normalization, followed by an add layer that adds the input of the first convolution layer and the output of the last one and applies a logistic. In all other aspects, the architecture coincides with that of the DA-net, however, without auxiliary layers at the bottom.

Inference is again realized by sliding a window with stride 40, this time of size  $320 \times 320$ . Outputs correspond to predictions for the central  $160 \times 160$  pixels of a window. Thereof, only the central  $80 \times 80$  pixels (inner mask, cf. next subsection) are taken into consideration. Hence, we obtain up to four predictions for every pixel of the input document image. Like for the DA-net, these are turned into a classification of every pixel into foreground or background. Further on, candidate baselines are extracted from connected components with at least 50 foreground pixels by applying the linear least square method. Note that thus only straight line segments are extracted.

#### A. Dice Error Modifications

Since we employ a pixel-wise classifier trained wrt. binary masks, we opted for an overlap metric for the error layer. In particular the dice coefficient is used which is a frequent choice to specify loss and assess the quality

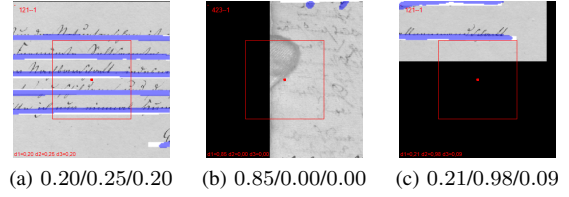


Figure 3. Variants of the dice error coefficient on three samples (a-c): values for each sample correspond to the standard dice error/the dice error with inner mask (central square, equivalent to standard dice error on that square)/the dice error with inner mask and surrounding mask (3 pixels around inner mask, not drawn); blue regions represent prediction foreground, white regions represent groundtruth foreground.

of segmentation (e.g., in medical image segmentation). However, we noticed some downsides with baselines near the border of a window: it is possible, that no text is visible but a groundtruth baseline reaches into the window (see Fig. 3b) yielding a disproportional high error. For this reason we apply a mask such that just an inner region is evaluated. Of course, when the end of baseline just reaches into that inner region, again a relatively high error is obtained for masks that are close to the groundtruth (cf. Fig. 3c). To further mitigate this problem we use another region surrounding the inner region that is treated as correct, i.e., where we use groundtruth for evaluation.

More formally, consider  $n \times n$  matrices  $Y$  and  $H$  denoting groundtruth and net output, respectively. Given integers  $a$  and  $b$ , such that  $a \leq b \leq n$ , we define a filter mask  $M(a, b)$  as the  $n \times n$  matrix given by

$$m_{ij} = \begin{cases} 1, & a \leq i, j \leq b \\ 0, & \text{else.} \end{cases}$$

Let  $M^I = M(a^I, b^I)$  be a filter mask (the *inner mask*), let  $M^Y = M(a^Y, b^Y)$  be another filter mask (*surrounding mask*), such that  $a^Y \leq a^I$  and  $b^Y \geq b^I$ , and let  $\gamma > 0$  be a constant. Then, the modified dice coefficient is given by

$$D(H, Y) = \frac{2 * |\bar{Y}\bar{H}| + \gamma}{|\bar{Y}| + |\bar{H}| + \gamma},$$

where  $\bar{Y} = Y M^Y$ ,  $\bar{H} = H M^I + Y(M^Y - M^I)$ . Note that  $a^Y = a^I$  and  $b^Y = b^I$  yields the dice coefficient on the inner mask only, if moreover  $a^Y = n$ , then we obtain the dice coefficient on the full image.

#### B. Document Scaling

Another characteristic of our approach is the application of individual document scaling on the basis of document properties predicted by the DA-net. To this end, we consider  $n_s$  different scales and an upper bound  $w_{max}$  on document width (in our case  $w_{max} = 8000$  and  $n_s = 7$ , namely 512, 640, 768, 896, 1024, 1152, and 1280 pixels). Given a document of width  $w$  and height  $h$ , a scale index value  $i_s$  is then calculated as follows:

$$i_s = \frac{n_s * w}{w_{max}} + \frac{n_s}{4} * (2 + p_{ls} + p_{dp} - 4 * p_{ld}).$$

Intuitively, the first term represents an initial offset depending on the width of the document, while the second term

may increase (landscape, double-page) or decrease this value (large leading). If the document is (almost) empty, i.e., if  $p_{nt} > 0.7$ , then this value is further decreased by  $\frac{n_s}{4} * p_{nt}$ . Also extra wide documents, i.e., if  $w/h > 2$ , yield a further decrease by  $\frac{n_s}{4} * (\frac{w}{h} - 1)$ . Eventually, the document is scaled to the width (keeping the aspect ratio constant) given by the  $i$ -th scale, where  $i = \min(\max(\lfloor i_s \rfloor, 0), n_s)$ .

#### IV. BASELINE POST-PROCESSING

The purpose of post-processing baseline candidates is twofold. First it aims at removing erroneous candidate line segments, and second, it is geared towards combining baseline candidates horizontally to a single baseline, e.g., a polygon representing an entire line in a paragraph.

*Error Pruning:* We consider short candidate line segments that either are disoriented, or that are covered by a longer line segment as unintended. More specifically, let  $l_{max}$  be an upper bound for line segments to be considered short, let  $\alpha_{max}$  be an orientation threshold, and  $d_{max}$  be a distance threshold. Then, we consider a candidate line segment  $l$  to be *short* iff  $\|l\|_2 \leq \min(0.2 * \bar{l}, l_{max})$ , where  $\bar{l}$  is the average length for a set of candidate line segments. In a first step we remove short candidate line segments that deviate more than  $\alpha_{max}$  degrees from the horizontal. In a second step, we remove short candidate lines that are covered by a longer candidate line segment, where coverage of a line segment  $l_1 = (s_1, e_1)$  by another line segment  $l_2 = (s_2, e_2)$  requires two conditions to be satisfied: The projections of  $s_1$  and  $e_1$  onto the line corresponding to  $l_2$  result in points on the segment  $l_2$ , and the minimal distance of any point on  $l_1$  from the line corresponding to  $l_2$  is smaller than  $d_{max}$ .

*Joining Baseline Segments:* is parametrized by a horizontal ( $d_x$ ), a vertical ( $d_y$ ), and an angular ( $d_\alpha$ ) threshold and proceeds recursively as follows. In a first run, we only allow joins to the right, i.e., the joined baseline segments must not overlap horizontally, whereas in a second pass we also allow for overlapping (leftward) joins. In particular, for a baseline candidate  $l_1$ , another baseline candidate  $l_2$  is a potential join line candidate wrt.  $l_1$  iff (i) the line containing  $s_2$  and  $e_2$  does not deviate more than  $d_\alpha$  degrees from the horizontal, (ii)  $|x_{s_2} - x_{e_1}| \leq d_x$ , and (iii)  $|y_{s_2} - y_{e_1}| \leq d_y$ . In case of non-overlapping joins, we additionally require that  $|x_{s_2} - x_{e_1}| \geq 0$ , for leftward joins  $x_{e_2} > x_{e_1}$  and  $\|s_2 - e_1\|_2 > d_x/3$  must hold in addition. Regarding preference, non-overlapping joins are always preferred over leftward joins, among non-overlapping joins minimal vertical distance prevails, among leftward joins it is minimal distance from the end point of  $l_1$ . Having determined preferred join line candidates (if any) for every baseline candidate, we assemble new candidate baselines (polygons) applying the respective join(s) recursively, starting from any base line candidate that itself is not a preferred join line candidate for another baseline.

#### V. EXPERIMENTS

In this section we first evaluate components and properties of our approach for baseline detection individually by

means of the cBAD competition<sup>1</sup> data and then compare to results on the HisDoc competition<sup>2</sup>.

*The cBAD Dataset and Competition:* was organized into two tracks using a dataset that contains various page layouts and degradations. The data [14] is composed of historical documents from 9 different archives, the documents span different time periods and were split into two sets. The first set (TRACK A) contains 755 samples that are simpler than the 1280 samples in the second set (TRACK B), since the former just contain text in paragraph form and, moreover, a segmentation into relevant text regions is provided. In contrast, TRACK B comprises more complex samples with various image distortions (e.g., noise, text-lines rotated up to 180°), tables, empty pages, marginalia, and without text region information.

A subset of the data can be used for training (215, resp. 269, samples) with publicly available annotated baselines serving as groundtruth. In addition, an evaluation tool for baselines is provided that computes precision and recall wrt. groundtruth lines (cf. [14] for details), as well as their harmonic mean (F-score). The remaining data served as test data with non-public groundtruth used for the competition with F-score as the decisive objective. Our approach yielded a final F-score of 97.13% on TRACK A, and without final joining of baselines 85.86% on TRACK B, winning both tracks [3].

*Training and Experimental Setup:* We reserved about 8% of the training data for cross validation (CVS set: 17, resp. 19, samples). Groundtruth binary masks were obtained from the text regions supplied with TRACK A for the DA-net, respectively from the baselines provided with both tracks for the BL-net (turning them into areas using 5 pixels height for every line segment). We additionally labelled the training and CVS data wrt. auxiliary document properties, i.e., whether or not the document is double-page, in landscape format or empty, and regarding leading (quantified into small, normal or large).

Our nets were trained by Adadelta (decay rate  $\rho = 0.95$ ) with a batch size of  $2 \times 32$  samples. Each sample is a random clipping ( $160 \times 160$  pixels, resp.  $320 \times 320$  pixels) from a prescaled training document image, where the scale is fixed to 192 pixel width for the DA-net, and computed individually (cf. Section III-B) for the BL-net. We also applied augmentation to 90% of the samples consisting of random rotation up to 180° and random scaling by  $\pm 0.5$ . Moreover, for the BL-net a dice coefficient with inner mask  $M^I = M(40, 120)$  and  $M^Y = M(37, 123)$  has been used. The post-processing step is also parameterized individually taking into account detected document properties. In particular, given an input image of width  $w$  pixel, we used  $l_{max} = w * 0.1$  (resp.  $l_{max} = w * 0.05$  if  $p_{dp} > 0.7$ ),  $\alpha_{max} = 30$ , and  $d_{max} = 20 + 50 * p_{ld}$  for error pruning. For joining we set  $d_x = w * 0.2$  (resp.  $d_x = w * 0.1$  if  $p_{dp} > 0.7$ ),  $d_y = 20 + 50 * p_{ld}$ , and  $d_\alpha = 50$ .

<sup>1</sup><https://scriptnet.iit.demokritos.gr/competitions/5/>

<sup>2</sup><http://diuf.unifr.ch/main/hisdoc/icdar2017-hisdoc-layout-comp>



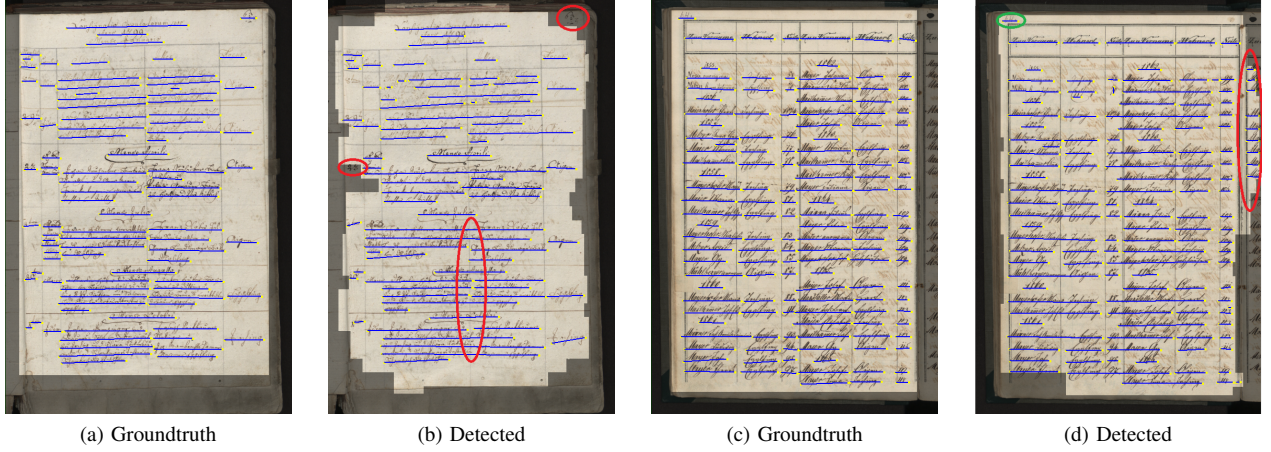


Figure 4. Samples of detected regions of interest (including detected baselines) for (the more complex) TRACK B.

#### A. Component Assessment and cBAD Results

**Simple Document Analysis:** To evaluate the extraction of relevant text regions by the DA-net, the dice coefficient of the detected mask wrt. the groundtruth region has been calculated. Table I lists the results obtained on the CVS set, as well as for the test data of TRACK A, since relevant regions are provided for this track. In both cases roughly 90% of the relevant areas are correctly detected. Visual inspection reveals that in many cases non-overlapping areas do not contain relevant text and result from a more fine-grained sampling of the mask compared to the rectangular groundtruth (see Fig. 4). However, sometimes also relevant text is missed, e.g., in Fig. 4b: a table entry and the page number (marked by red ellipses to the left and the upper right), or the mask partially contains marginal text as in Fig. 4d: the page number is contained (green) but also spurious text from the next page (red).

	region	landscape	dbl.-page	no text	leading
CVS	90.89	100	94.74	97.37	81.58
Test	90.21	99.29	95.93	94.19	73.85

Table I  
EVALUATION OF SIMPLE DOCUMENT ANALYSIS (% GT).

Concerning auxiliary document properties, outputs for landscape, double-page, and no text are binarized thresholding at 0.5, and leading has been quantified into small, medium, or large using thresholds 0.3 and 0.6. For post-competition evaluation we also labeled the test data wrt. these properties. While stable results are obtained for the binary properties, correctness drops for leading. A more rigorous and consistent labeling than subjective visual classification may help to improve in this regard.

**Baseline Detection:** Table II reports precision and recall for baselines detected in various settings on the cBAD dataset. Results are listed for both tracks on the CVS set as well as on the test set. Entries in the first column correspond to detections obtained by applying the BL-net standalone, i.e., with fixed pre-scaling to 1024 pixel width and without post-processing (for the test data with a precursor of the final BL-net). Values in the second column are obtained applying individual document scaling

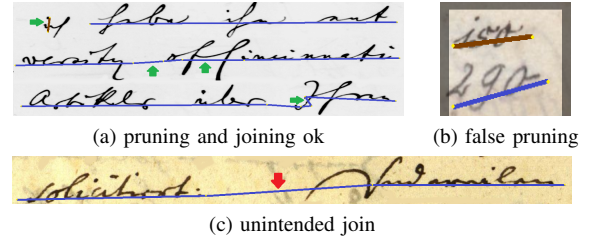


Figure 5. Samples of final baselines after post-processing.

on behalf of the document properties returned by the DA-net, but still without post-processing. Document scaling thus clearly increases precision for both tracks. Moreover, note that the test set for TRACK B does not specify relevant text regions, corresponding results are obtained applying the region masks extracted by the DA-net.

	BL-net fixed scale	no post-p. ind. scale	add. pruning ind. scale	add. joining ind. scale
CVS A	90.88/95.95	92.73/95.47	93.28/94.93	96.92/95.07
CVS B	85.44/85.63	86.22/88.02	86.79/87.78	n.a.
Test A	76.59/96.76	81.41/96.27	91.08/94.78	<b>97.27/96.99</b>
Test B	76.87/90.08	78.82/89.04	<b>85.42/86.30</b>	n.a.

Table II  
EVALUATION OF BASELINE DETECTION (% PREC./% REC.).

**Baseline Post-Processing:** Precision increases further when error pruning is applied (cf. third column in Table II). However, besides pruning erroneous lines like the short and almost vertical line in Fig. 5a (top left green arrow), false positives also eliminate intended lines like in Fig. 5b (brown line) thus decreasing recall. In summary, it still has a slightly positive effect (increases F-score).

Baseline joining only applies to TRACK A which is restricted to text in paragraph form. There, however, it achieves a significant improvement (forth column in Table II) with an overall increase wrt. both, precision and recall. Unintended joins as in Fig. 5c are rare (see Fig. 5a for several positive joins indicated by green arrows, including recursive joins and a leftward join).

#### B. HisDoc Baseline Detection Results

For the HisDoc competition a collection of pages from three medieval manuscripts has been selected with re-

gard to the complexity of their layout [5]. Compared to cBAD the most distinctive characteristic is that they also contain interlinear glosses (explanatory notes) in addition to marginal annotations and additions. The competition comprised three different tasks, one (Task 2) addressing baseline extraction. It was laid out to make the competition results comparable to other competitions such as cBAD, e.g., using the same evaluation tool (with parameters:  $-\text{tTF } 0.75 -\text{minT } 20 -\text{maxT } 20$ ).

We thus compare our approach using the same parametrization as before except for three minor adaptations: the upper bound on document width is changed to reflect the actual scale of the data ( $w_{\text{max}} = 5000$ ); the distance threshold for error pruning is enlarged to account for larger leadings due to interline glosses ( $d_{\text{max}} = 20 + 500 * p_{\text{ld}}$ ), while the corresponding orientation threshold is decreased ( $\alpha_{\text{max}} = 3$ ) for stricter short line removal (no tabular data).

	CB55	CSG18	CSG863	Total
HisDoc best	<b>98.96</b>	<b>98.79</b>	<b>98.30</b>	<b>98.22</b>
unmodified	95.73	80.79	80.46	86.01
BL-net train ctd.	99.91	97.43	97.96	98.44
BL-net train new	<b>99.91</b>	<b>99.25</b>	<b>98.52</b>	<b>99.23</b>

Table III  
RESULTS ON HISDOC PRIVATE TEST SET (% F-SCORE).

Table III lists F-score percentages obtained for the three manuscripts individually and in total. The first row repeats the best results reported in [5] for each class (rather than the winning system), while the second line gives the results of our system without any additional training. The scores reflect that the BL-net had not been trained with interline glosses and is penalized for returning baselines for them. Training the model (just the BL-net) for another 30k iterations (as long as for cBAD) on the HisDoc training samples yields a system scoring slightly better than the winning entry. Training the BL-net from scratch (also 30k iterations) using color images rather than greyscale images (which we used for cBAD) our approach clearly outperforms the HisDoc contestants in all three categories.

## VI. DISCUSSION AND CONCLUSION

We presented a novel approach to baseline extraction for historical documents that utilizes U-nets and a sliding window approach for simple layout analysis in a pre-processing step, as well as for actual baseline detection. It is complemented by procedural baseline post-processing for pruning spurious candidate lines and potential assembly of candidate line segments into poly-baselines. It yields convincing results for heterogeneous collections of historical documents exhibiting various degradations (e.g., bleed-through, ornamentations, interlinear glosses), as long as they are of a simple structure (text in paragraph form). To wit, besides winning the 2017 cBAD competition, we showed that it also outreaches the results reported for a corresponding task of the 2017 HisDoc competition.

For more complex structured documents however, there is still potential for further improvement and several issues remain for further research. For instance, increasing the accuracy of relevant text region detection (DA-net)

would affect recall considerably avoiding the inclusion of marginal text (as in Fig. 4d). To this end, dilated residual networks [15] are promising architectures.

Moreover, extending our simple document analysis towards multi-dimensional classification in order to segment various document regions (table boundaries, marginalia, ornamentations, etc.) is an obvious next step. E.g., if we had had a more detailed segmentation, we would not have had to retrain the BL-net in the presence of interlinear glosses (but simply skip baselines from such regions). This illustrates another benefit of pursuing a modular approach: adaption to different application cases (including glosses or not) not necessarily requires re-training.

Eventually, it seems necessary to consider context for baseline post-processing in complex scenarios. While errors (such as in Figures 5b and 5c) were rare in simple settings, preliminary experiments with procedural baseline splitting along vertical lines (in cases as in Fig. 4b) indicate a need for more sensitive decisions. Resorting to recursive networks, with local spatial context information, would be an interesting route towards learning corrective post-processing actions (joining, splitting, pruning).

## REFERENCES

- [1] V. Romero, J. Sánchez, V. Bosch, K. Depuydt, and J. de Does, "Influence of text line segmentation in handwritten text recognition," in *ICDAR 2015*, pp. 536–540.
- [2] L. Likforman-Sulem, A. Zahour, and B. Taconet, "Text line segmentation of historical documents: a survey," *IJDAR*, vol. 9, no. 2-4, pp. 123–138, 2007.
- [3] M. Diem, F. Kleber, S. Fiel, B. Gatos, and T. Grüning, "c-BAD: ICDAR2017 competition on baseline detection," in *ICDAR 2017*, pp. 1355–1360.
- [4] M. Murdock, S. Reid, B. Hamilton, and J. Reese, "ICDAR 2015 competition on text line detection in historical documents," in *ICDAR 2015*, pp. 1171–1175.
- [5] F. Simistira, M. Bouillon, M. Seuret, M. Würsch, M. Alberti, R. Ingold, and M. Liwicki, "ICDAR2017 competition on layout analysis for challenging medieval manuscripts," in *ICDAR 2017*, pp. 1361–1370.
- [6] J. Sánchez, V. Romero, A. H. Toselli, and E. Vidal, "ICFHR2016 competition on handwritten text recognition on the READ dataset," in *ICFHR 2016*, pp. 630–635.
- [7] E. Shelhamer, J. Long, and T. Darrell, "Fully convolutional networks for semantic segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 640–651, 2017.
- [8] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR 2016*, pp. 770–778.
- [9] A. Graves, S. Fernández, and J. Schmidhuber, "Multi-dimensional recurrent neural networks," in *ICANN 2007*, LNCS vol. 4668, pp. 549–558.
- [10] B. Shi, X. Bai, and C. Yao, "An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition," *TPAMI*, vol. 39, no. 11, pp. 2298–2304, 2017.
- [11] T. Wang, D. J. Wu, A. Coates, and A. Y. Ng, "End-to-end text recognition with convolutional neural networks," in *ICPR 2012*, pp. 3304–3308.
- [12] J. Pastor-Pellicer, M. Z. Afzal, M. Liwicki, and M. J. C. Bleda, "Complete system for text line extraction using convolutional neural networks and watershed transform," in *DAS 2016*, pp. 30–35.
- [13] B. Moysset, J. Louradour, C. Kermorvant, and C. Wolf, "Learning text-line localization with shared and local regression neural networks," in *ICFHR 2016*, pp. 1–6.
- [14] T. Grüning, R. Labahn, M. Diem, F. Kleber, and S. Fiel, "READ-BAD: A new dataset and evaluation scheme for baseline detection in archival documents," *CoRR*, vol. abs/1705.03311, 2017.
- [15] F. Yu, V. Koltun, and T. A. Funkhouser, "Dilated residual networks," in *CVPR 2017*, pp. 636–644.