НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ
май–июнь 2023 Том 23 № 3 http://ntv.ifmo.ru/
SCIENTIFIC AND TECHNICAL JOURNAL OF INFORMATION TECHNOLOGIES, MECHANICS AND OPTICS
May–June 2023 Vol. 23 No 3 http://ntv.ifmo.ru/en/
ISSN 2226-1494 (print) ISSN 2500-0373 (online)

doi: 10.17586/2226-1494-2023-23-3-585-594

Joint recognition of text and layout in historical Russian documents

Samah Mohammed[1]✉, Nikolay Teslya[2]

[1] ITMO University, Saint Petersburg, 197101, Russian Federation
[2] St. Petersburg Federal Research Center of the Russian Academy of Sciences, Saint Petersburg, 199178, Russian Federation

[1] samahslimanmhmd93@gmail.com✉, https://orcid.org/0000-0002-8009-7222
[2] teslya@iias.spb.su, https://orcid.org/0000-0003-0619-8620

Abstract
In this paper, we evaluated the Document Attention Network (DAN), the first end-to-end segmentation-free architecture on Historical Russian Documents. The DAN model jointly recognizes both text and layout from whole documents, it takes whole documents from any size as an input and output the text as well as logical layout tokens. For comparison purposes, we conduct our experiments on Digital Peter dataset as it has been recognized at line-level. Dataset consists of documents of Peter the Great manuscripts; ground truths are represented according to a sophisticated XML schema which enables an accurate detailed definition of layout and text regions. We achieved good results at page-level: 18.71 % for Character Error Rate (CER), 39.7 % for Word Error Rate (WER), 14.11 % For Layout Ordering Error Rate (LOER), and 66.67 % for mean Average Precision (mAP).

Keywords
document understanding, handwritten text recognition, layout analysis, fully connected networks, transformers

Acknowledgments
The study was carried out at the expense of state funding, topic project No. FFZF-2022-0005.

For citation: Mohammed S., Teslya N. Joint recognition of text and layout in historical Russian documents. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2023, vol. 23, no. 3, pp. 585–594. doi: 10.17586/2226-1494-2023-23-3-585-594

УДК 004.932.75

Совместное распознавание текста и оформления в исторических документах на русском языке

Самах Мохаммед[1]✉, Николай Тесля[2]

[1] Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация
[2] Санкт-Петербургский Федеральный исследовательский центр Российской академии наук, Санкт-Петербург, 199178, Российская Федерация

[1] samahslimanmhmd93@gmail.com✉, https://orcid.org/0000-0002-8009-7222
[2] teslya@iias.spb.su, https://orcid.org/0000-0003-0619-8620

Аннотация
Рассмотрена сквозная, свободная от сегментации архитектура Document Attention Network (DAN), на примере распознавания исторических документов на русском языке. Архитектура DAN способна распознать текст или макет документа любого размера и вывести распознанный текст, а также логические области макета оформления. Выполнено сравнение полученных результатов экспериментов с набором данных Digital Peter, по которому обучены модели распознавания рукописного текста, имеющие высокую точность распознавания на уровне строк. Набор данных состоит из документов рукописей Петра Великого. Эталонные данные для архитектуры DAN представлены в соответствии со сложной схемой формата XML, которая обеспечила точное определение макета оформления и текстовых областей. Получены следующие результаты распознавания текста на уровне страницы: 18,71 % для коэффициента ошибок символов (Character Error Rate, CER), 39,7 % — коэффициента

ошибок в словах (Word Error Rate, WER), 14,11 % при упорядочении макета слов (Layout Ordering Error Rate, LOER) и 66,67 % для средней точности (mean Average Precision, mAP).

Ключевые слова
понимание документов, распознавание рукописного текста, анализ макета оформления, полносвязные сети, преобразователи

Благодарности
Исследование выполнено за счет средств государственного финансирования, тема FFZF-2022-0005.

Ссылка для цитирования: Мохаммед С., Тесля Н. Совместное распознавание текста и оформления в исторических документах на русском языке // Научно-технический вестник информационных технологий, механики и оптики. 2023. Т. 23, № 3. С. 585–594 (на англ. яз.). doi: 10.17586/2226-1494-2023-23-3-585-594

Introduction

Document Understanding [1] is a maturing field that includes a set of tasks whose purpose is to extract, classify, interpret, contextualize, and search information from documents. It implies, among others, document layout analysis and Handwritten Text Recognition (HTR).

The most state-of-the-art approaches for the task of recognition of handwritten documents are handling this task through a complex pipeline implying two main steps: segmentation and recognition. These current approaches require a huge amount of segmentate annotated documents which are very costly to produce. Moreover, the prediction error rate will be higher because of accumulating errors from both stages, this prevents a holistic understanding of the document. Also, the recognition at line-level and even paragraph-level can't preserve the global coherence of the content which is only feasible with an understanding of the layout. In addition, methods with segmentation stage have no notion of the ordered sequence among the different text regions in the same document, and that limits the ability to learn the reading order. The recognition of historical Russian documents is useful for historian's needs, to avoid experts to transcribe the texts themselves, which is very time-consuming. The Document Attention Network (DAN) model proposed by Coquenet et al. [2] jointly recognizes both text and layout at document level. The model is based on a Fully Connected Network (FCN) encoder whereas the transformer proposed by Vaswani et al. [3] had been chosen to be the decoder since it has proven its robustness HTR tasks.

In this paper, we describe the first historical Russian dataset at document level; manuscripts written by Peter the Great were used, ground truth for each document is represented by a sophisticated XML schema which is a part of the Page Analysis and Ground truth Elements (PAGE) format framework [4]. To generate this schema, a ground-truthing system Aletheia proposed by Clausner et al. [5] had been used. The XML schema is flexible, extensible, and it provides a detailed explanation of layout regions and thus it generates a serialized representation of the document that is further used as a ground truth to recognize this document.

Recent works to recognize the handwritten Russian data have applied segmentation at character level. Later on, segmentation was applied at line level [6, 7]. To our knowledge, this is the first attempt to recognize the Russian manuscripts at document level.

As we aim to recognize both text and layout from historical Russian documents, we briefly make the following contributions:

— An end-to-end architecture (DAN) [2] is trained to jointly recognize the text at document level and to label logical layout information without the need of any segmentation labels.
— We rework a historical Russian dataset at page level where the ground truth is represented by an XML paradigm to generate a serialized representation of the document.

Related work

Since our purpose is to recognize both text and layout from historical documents, this section is first devoted to HTR, and then it concentrates on Document layout Analysis.

Handwritten Text Recognition

Most works have focused on isolated line recognition and only a few studies have dedicated to multi-line recognition [8]. In the literature, many approaches have been proposed: combination of Convolutional Neural Networks (CNNs) to extract the features from an image and Bidirectional Long Short-Term Memory (BLSTM) to predict the sequence of the characters [9–11], or Multi-Dimensional LSTM [12], the problem with the MD-LSTM is the high computational cost. Alternative model proposed by Puigcerver et al. [13] which relies only on convolutional and one-dimensional recurrent layers, achieves better results and runs faster. These previous models are utilizing the Connectionist Temporal Classification (CTC) objective function [14] where probability distribution is computed over all possible output sequences for a given input sequence. However, CTC-based architectures are subject to internal limitations, such as strict Input/Output alignment and output sequence length correlated with input length.

On the contrary, sequence-to-sequence (Seq2Seq) models that follow the encoder-decoder structure are more flexible, suited to the temporal nature of text and are able to focus on the most relevant features of the input by incorporating attention mechanisms. Although, attention mechanism allows the networks to model the language structures. The encoder in the TrOCR model proposed by Li et al. [15] can be initialized with pre-trained ViT models [16–18], while the decoder can be initialized with pre-trained BERT models [19–21]. The model proposed by Bluche et al. [12] based on covert and overt attention with

MDLSTM network to recognize at paragraph level without any prior segmentation but this approach is limited by the prohibitive memory requirements. Similar model proposed by [22] to perform full page handwritten recognition without image segmentation but this model need to be trained with longer sequence length and reduce the encoder size (22 million parameters). Rouhou et al. [23] proposed a transformer-based model that jointly operates Handwriting and Named Entity Recognition only at paragraph level. All the approaches that proposed so far for the task of Russian historical documents recognition use explicit line segmentation and this is very costly to produce. The authors in [6] used the Convolutional Recurrent Neural Network (CRNN) model proposed by Shi et al. [11] to recognize Peter the Great manuscripts at line level. Later, the authors introduced two novel data augmentation methods [7], strikethrough text algorithm Handwritten Blots, and handwritten text generation algorithm StackMix; the two methods are applied to the Resnet-BiLSTM-CTC model.

Document Layout Analysis

Document Layout Analysis (DLA) is the process of extracting and identifying the physical regions of interest in a textual document. This process is based on physical ground truth annotations that describe the related semantic labels to each document. DLA is a maturing field and here are some recent approaches. DeepDeSRT proposed by Schreiber et al. [24] is the first end-to-end model for table understanding. Later Fully Convolutional Neural Networks (FCNNs) are proposed for pixel-wise segmentation on historical documents in order to detect tables and figures in these documents [25]. A Multimodal-FCN (MFCN) introduced by Yang et al. [26] for document semantic structure, it uses both textual and visual information. In [27], the LayoutLM model introduced for Visual Document Understanding (VDU) tasks like named entity recognition and key-value pair information extraction, the model uses the masked visual-language model and the multi-label document classifications as the training objectives and the BERT architecture used as the backbone. Later on, an improved version of LayoutMv2 [28] presented where the visual information is integrated in the pre-training stage to learn the cross-modality interaction between visual and textual information.

For instance level recognition, the first End-to-End Document Image Segmentation Transformer (DocSegTr) introduced by Biswas et al. [29], this model shows high performance with overlapped layout objects but doesn't manifest much improvements for smaller regions. The self-supervised pre-training for Document image Transformer proposed by Li et al. [30], the model is pre-trained with large scale unlabeled document images where each document image is divided into nonoverlapping patches before passing it into a stack of transformers. Table 1 provides a brief comparison between the previous approaches.

Data Preparation

In this section we describe the procedures using the ground-truthing system tool Aletheia to obtain XML paradigm for each handwritten document of Peter the Great manuscripts that is further used as a ground truth of the document image. We suppose this dataset may be beneficial for researchers to train HTR at document level.

Mark et al. [6] convert the manuscripts of Peter the Great (662 full copies) into lines, they had to split each document image manually using Computer Vison Annotation Tool (CVAT)[1], the resulting dataset consists of 9694 images and text files. Because the text for each document image, as the document images have the following format x_y_z, where x is the series number, y is the page number, and z is the line number on this page, then we could be able to reassign each line with its document image and obtain a full transcription of the document.

As a prior stage, some documents need to be cropped to remove all the items that can effect on the recognition process, such as the stamp of Federal Archival Agency, irrelevant writing in the background, etc., and we keep only Peter the Great writing. Then, we manually detect all the text regions in the document image by drawing a rectangle around each one; we also manually provide the reading orientation and the corresponding labels for all the text regions. Based on these regions, we automatically

[1] Available at: URL: https://doi.org/10.5281/ZENODO.4009388 (accessed: 01.11.2022).

Table 1. State-of-art approaches for Document Layout Analysis

DLA Model	Main architectures	Document of Artificial Intelligence tasks
dhSegment [25]	Single Res-Net 50	— Page and baseline extraction. — Document Layout Analysis. — Ornament detection. — Photo-collection extraction.
MFCN [26]	Multimodal Fully Convolutional Networks	Extract semantic structures from document images
LayoutLM [27]	BERT architecture	Information extraction from scanned documents
LayoutLMv2 [28]	Multi-Modal Transformer	Model the interaction among text, layout and image in a single multimodal framework.
DocSegTr [29]	Hybrid CNN based Transformer	— Instance-level extraction. — Document Layout analysis.
DiT [30]	Vanilla Transformer architecture	— Document image classification. — Document Layout Analysis. — Table extraction.

added a class to each text region among these five classes: Page (P), Page Number (N), Body (B), Annotation (A) and Section (S) where the section is a group of linked annotations and body. For the text lines, we also draw a rectangle or polygons surrounding each line, we assign the baseline points, the labels and reading order for each line in each text region. We also highlight that reading order is defined by hand, based on the sequence of the text lines and text regions, this could cause some errors in case of slanted lines. Then, a PAGE XML diagram is generated for each document and this diagram is further used as a ground of the document image.

The dataset consists of 350 documents at page level with the corresponding PAGE XML diagram for each document.

Model Architecture

We opted for DAN model [2] as it is the first end-to-end encoder-decoder architecture to recognize both the text and layout from a document image, each document is represented by a sequence y of tokens with length L_y as shown in Fig. 1.

The FCN had been chosen as an encoder because it is known for its ability to handle input images of different sizes. In [31], the FCN encoder takes as input a document image $X \in R^{H\times W\times C}$, where H, W, C are the height, width, and number of channels, respectively ($C = 3$ for an RGB image). The encoder extracts 2D features maps from the input document: $f_{2D} \in R^{H_f\times W_f\times C_f}$, where $H_f = \frac{H}{32}$, $W_f = \frac{W}{8}$ and $C_f = 256$. The positional encoding describes the position of an entity in a sequence so that each position is assigned a unique representation; mathematically it can be explained as sine and cosine functions with different frequencies. Then, the 2D features maps are added to a 2D positional encoding to get 1D features maps. It is used to preserve spatial information. The 1D features maps with the previously predict tokens ($\hat{y}_0$, $\hat{y}_1$, …, $\hat{y}_{t-1}$) form the input to the decoder which made up of a stack of 8 transformers based on multihead attention mechanism followed by a 1 × 1 convolutional layer to compute the next token probabilities, the multihead attention is based on previous predictions where queries Q, keys K and values V are from the same input. Also, an attention window is used to reduce to computation time, It means that given an input sequence s of length L_s, the t^{th} output frame o_t is computed over the range $[s_a, s_{t-1}]$ with $a = \max(0; t-100)$. The DAN model is trained using the cross-entropy function over the sequence of tokens.

$$\varsigma = \sum_{t=1}^{L_y+1} \varsigma_{CE}(y_t, p_t),$$

where ς_{CE} is the cross entropy loss for each token and ς is the total cross-entropy loss over the sequence of tokens.

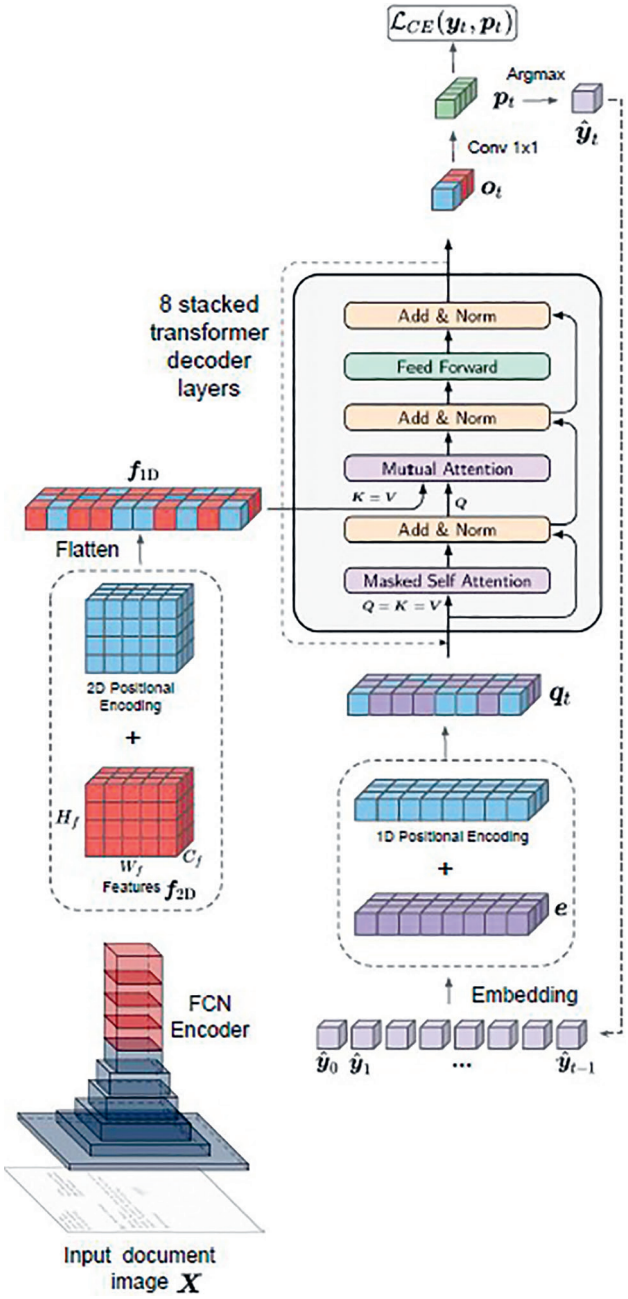


Fig. 1. Document Attention Network [1]
(Abbreviations of the variables explained later in the text)

Training strategies and Metrics

Training strategies

As a prior stage before training, synthetic printed lines and synthetic documents are generated where the synthetic printed lines are used to train the feature extraction part of the DAN, and the synthetic documents are used to learn the reading order as the reading order is the same between the real and synthetic documents. For synthetic printed lines, the real documents dataset D_{doc} is used to extract isolated text lines transcriptions and generate lines dataset D_{line} which is used to generate arbitrarily synthetic printed lines; we chose a set of more than 40 different fonts with different sizes to have more variability and make the model more robust. The generation of synthetic documents is explained by the algorithm proposed in [2] where a document image is chosen randomly from the input documents to be a template to generate the synthetic documents.

The main points in the training process can be summarized as follows:

— Synthetic printed text lines are generated to train a line-level OCR model as proposed in [31], the weights then used to initialize the weights of the FCN encoder and the last convolutional layer of the decoder for training the DAN on our dataset.
— In order to reduce the overfitting, a data augmentation strategy with a probability of 90 % is used, transformations, such as color jittering, erosion, dilation, gaussian blur, gaussian noise and resolution modification, are applied in random order with a probability 10 % for each one.
— A teacher forcing is used at training to parallelize the computations by predicting the whole sequence at once where the previously predicted tokens are replaced by the ground truth.
— A curriculum strategy is applied to improve the convergence where the training process begins with 90 % synthetic documents to learn the reading order first, then the percentage is reduced gradually to 20 % to fine tune on the real documents.
— The DAN model is trained on heterogenous documents; from historical German dataset (READ_16 dataset) and historical Russian dataset (Digital Peter dataset).
— After training, the unpaired predicted layout tokens are handled through a forward pass on the entire document by adding a missing end token or removing an isolated one. The predicted layout tokens will be used to evaluate the layout recognition.

Metrics

As the DAN model jointly recognizes text and layout, the evaluation of the model is correlated with the evaluation of the text recognition, the layout recognition, and the evaluation of joint recognition of both layout and text.

The evaluation of text recognition. The performance of text recognition is evaluated using the Character Error Rate (CER) which is the most common metric to evaluate the text recognition approaches [8, 13, 31, 32]. It is the sum of Levenshtein distance (d_{lev}) among the ground truth y^{text} and the predictions $\hat{y}^{text}$ (after removing all the layout tokens) from The XML diagram, normalized by the total length of the ground truth $y^{text}_{len_i}$:

$$CER = \frac{\sum_{i=1}^{K} d_{lev}(y^{text}, \hat{y}^{text})}{\sum_{i=1}^{K} y^{text}_{len_1}},$$

where d_{lev} is the minimum number of single-character (or word) edits (i.e., insertions, deletions, or substitutions) required to change one word (or sentence) into another.

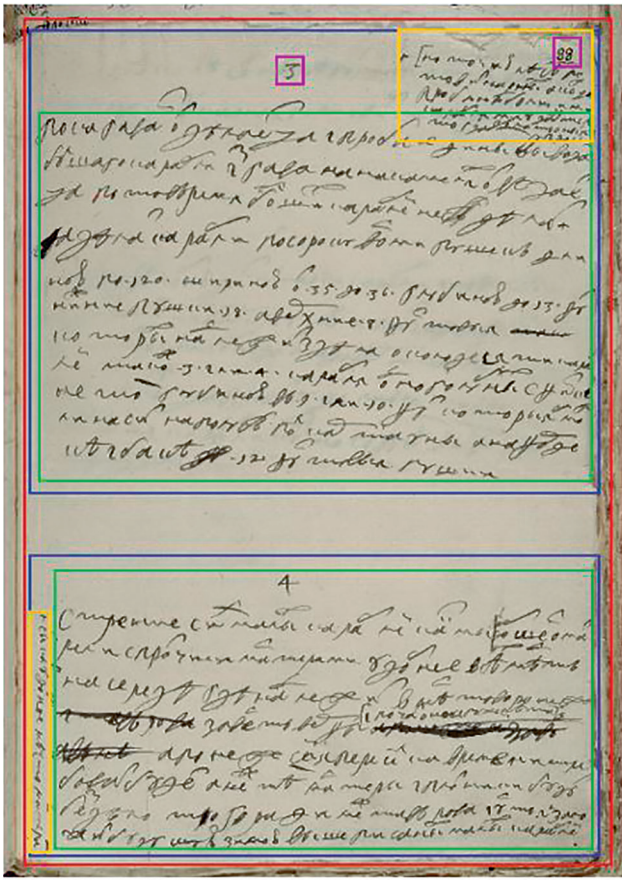
Word Error Rate (WER) is also used to evaluate text recognition and it is computed in the same way but at word level.

The evaluation of layout recognition. The evaluation of the layout recognition should be considering the evaluation of reading order and the hierarchical relations between layout entities. We have to model the layout as an oriented graph by first computing the ground truth y^{graph} and the predictions $\hat{y}^{graph}$ (after all but layout tokens are removed). e.g. y^{graph} = <D> <S> <P> <P/> <S/> <D/>, where D is a document, P is a paragraph, S is a section, N is number, A is annotation, and B is body. Then by following the hierarchical rules of dataset, we map the sequence of layout tokens into an oriented graph as shown in Fig. 2 where the nodes are the document entities (Document, Section, Annotation, Number).

The dashed arrows represent the hierarchy of the entities within the document and the solid arrows represent the reading order of these entities. The Layout Ordering Error Rate (LOER) (represented below) is used to evaluate the layout recognition, it is calculated as the Graph Edit Distance (GED) for K samples, normalized by the sum of the number of the edges n_{e_i} and nodes n_{en_i} in the ground truth:

$$LOER = \frac{\sum_{i=1}^{K} GED(y^{graph}, \hat{y}^{graph})}{\sum_{i=1}^{K} n_{e_i} + n_{en_i}}.$$

a



b

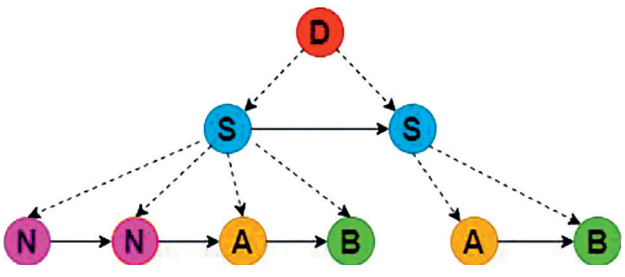


Fig. 2. Document image (*a*) with associated layout graph annotation (*b*)

The graph edit distance is computed using a unit cost of edition whether it is for addition, removing or substitution, and whether it is for edges or for nodes. However, computing CER and LOER is not sufficient to evaluate the correct recognition of document. These metrics can't evaluate the association between the layout and the text parts of the document.

The evaluation of joint recognition of both text and layout. Based on the mean average precision mAP score proposed by [33] for object detection approaches, Coquenet et al. [1] proposed a similar new metric to evaluate the joint recognition of both text and layout where the predicted sequence $\hat{y}$ and the ground truth y sequence are split into sub-sequences extracted using the begin and the end tokens of the same class c, and thus the mean average precision metric is computed as a weighted sum over the different layout classes weighted by the number of characters len_c in each class c:

$$mAP_{CER} = \frac{\sum_{c \in S} AP_{CER_c}^{\theta_{min}:\theta_{max}:\Delta\theta} len_c}{\sum_{c \in S} len_c},$$

where θ_{min}:θ_{max}:$\Delta\theta$ are different CER thresholds.

Experiments

In this section we evaluate the DAN model on the Digital Peter dataset at document level, we considered the same configurations for pre-training and training on DAN model. Also, we provide an ablation study to emphasize the key components that made it feasible for these outcomes to be achieved. The DAN model is evaluated on the test set of RIMES dataset [34] and READ 2016 dataset [32] and compared with the most state-of-the-art approaches at line-level as shown in Table 2 and 3.

One can notice that the values of CER, WER are slightly better for the page level dataset, and this can be explained by the higher necessity to understand the layout.

The Digital Peter dataset is split in training, validation, and test sets on-line level and page level (Table 3). Also, we provide the number of characters for each and the number of layout tokens.

To generate synthetic documents, we set the maximum number of lines per page $l_{max} = 35$ to match the dataset properties. For all our experiments on DAN model, we use the Adam optimizer with an initial learning rate (lr = 0.00005). Pre-training and training are performed on a single GPU RTX 3090 (24 GB).

The DAN model had been trained with mini-batch size of 16 for line-level training and mini-batch size of 1 for training at page level. The DAN model is first trained a line-level OCR model on synthetic printed lines, this pre-training step is carried out during 24 hours and then for transfer learning purposes, the weights are used to initialize the FCN encoder and the last convolutional layer of the decoder where training at page level is carried out during 5 days. Moreover, we didn't use any language model.

For the Digital Peter dataset, we evaluated the DAN model on the test set at page-level and compared the results with the state-of-the-art models at line-level as detailed in Table 4.

The DAN is based on an autoregressive prediction process. This is not a problem at training time since computation are parallelized through teacher forcing. However, this recurrence issue is significant at prediction time: it grows linearly with the number of tokens to be predicted. This way, the average prediction time for a test sample is 4.4s. We aim at reducing this prediction time in future works. An example of an input document with the output of the DAN model is shown in Fig. 3.

Table 2. Evaluation of the DAN model on the test datasets READ 2016 and RIMES with the line-level recognition approaches, %

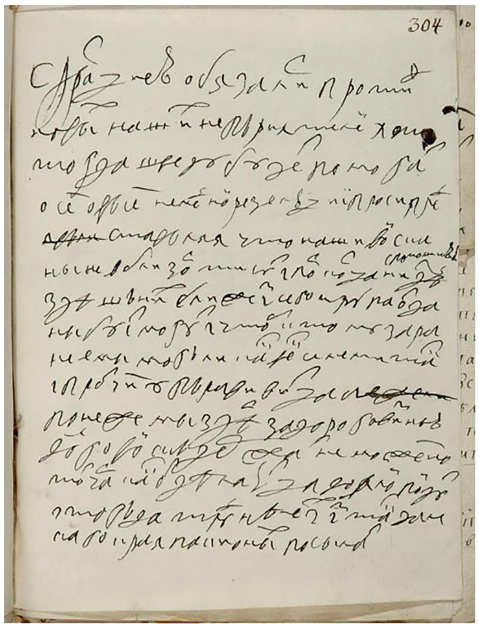
Model	CER	WER	LOER	mAP_{CER}
READ 2016 dataset				
Line level CNN + RNN [32]	5.10	21.10	—	—
FCN + BLSTM [31]	4.10	16.29	—	—
Page level DAN	3.53	13.33	5.94	92.75
RIMES dataset				
Line level FCN + BLSTM [31]	3.04	8.32	—	—
CNN + BLSTM [13]	2.30	9.60	—	—
Page level DAN	4.54	11.85	3.82	93.74

Table 3. Dataset partitions with associated number of characters and layout tokens

Dataset	level	training	validation	test	no. of char	no. of layout tokens
Digital Peter	Line	6237	1527	1930	84	—
	Page	250	50	50	84	10

Table 4. Evaluation of the DAN model on the test set of Digital Peter dataset with the line-level recognition approaches, %

Model	CER	WER	LOER	mAPCER
Line level CRNN [6]	7.10	39.7	—	—
Resnet + BLSTM [7]	2.50	14.6	—	—
Page level DAN	18.71	39.7	14.11	66.67



```
<page>
 <section>
   <body>
скфранциею обязалис протиф
но вых наших непрiятелей х т
тогда шведу будет помогат
о сем о всем немедной резелюциi
проси пред
авъля ставъляя что наши войска
ныне в близости сут i посконцани дѣл
здѣшъних ближе iз сего крупа отда
ны буiт могут i чтоб к тому зара
не емы могъ ли как векселеми так
i протчим упъравивитца а вжели
понеже мы здѣс за дорого визною
долговойскъ держат не можемно
тотчас как отдѣлаютца до алой пойдут
i то въда труднѣе iх iз так дале
каго краяпаки оных посылат
   </body>
 </section>
</page>
```

Fig. 3. Example of the input and prediction of the DAN model

Table 5. Ablation study of the DAN model on the Digital Peter dataset, %

Case number	DAN model	CER	WER	LOER	mAPCER
1	Base	84.34	90.73	39.24	3.95
2	Without synthetic data	86.37	93.82	50.32	0.61
3	Without pre-training stage	88.96	89.43	45.57	2.49

Then, we provide an ablation study as detailed in Table 5. All the experiments are carried out for 2 days. In case 2, the model is trained only on the available real documents without generating synthetic documents. In case 3, we train the DAN model on Digital Peter dataset from scratch, without transfer learning from a prior pre-training step.

As one can notice, results are dramatically worse, which highlights that the obtained results are very dependent on the synthetic data and its quality, i.e., they must be close to the original dataset, notably in terms of layout.

Conclusion

In this paper, we evaluated the DAN model on the Digital Peter dataset. DAN is the first end-to-end free segmentation model to tackle Handwritten Document Recognition (HDR) which corresponds to the joint recognition of text and layout. HDR is a new step toward the holistic understanding of whole handwritten documents; meanwhile the recognition at line-level and even paragraph-level can't preserve the global coherence of the content which is only feasible with an understanding of the layout. The obtained results for text recognition are comparable to those obtained at line level, this can be explained by the higher necessity to understand the layout. Also, this increases the speed of deciphering historical documents. For example, it took a team of 10–15 historians about 3 months to decipher 662 full digital copies of Peter the Great's manuscripts but when working on the same dataset on a single GPU RTX 3090, the average decryption speed was 13 pages/min, which is encouraging by historical scientists. In future works, we will work on improving the result.

Also, it would be interesting to go a step further, by recognizing handwritten documents with heterogeneous sizes and layouts.

References

1. Sánchez J., Romero V., Toselli A.H., Vidal E. ICFHR2016 competition on handwritten text recognition on the READ dataset. *Proc. of the 15th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, 2016, pp. 630–635. https://doi.org/10.1109/icfhr.2016.0120
2. Coquenet D., Chatelain C., Paquet T. DAN: a segmentation-free document attention network for handwritten document recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023, in press. https://doi.org/10.1109/tpami.2023.3235826
3. Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A.N., Kaiser Ł., Polosukhin I. Attention is all you need. *Advances in Neural Information Processing Systems 30 (NIPS 2017)*, 2017, pp. 5998–6008.
4. Pletschacher S., Antonacopoulos A. The PAGE (Page Analysis and Ground-truth Elements) format framework. *Proc. of the 20th International Conference on Pattern Recognition*, 2010, pp. 257–260. https://doi.org/10.1109/icpr.2010.72
5. Clausner C., Pletschacher S., Antonacopoulos A. Aletheia - An advanced document layout and text ground-truthing system for production environments. *Proc. of the International Conference on Document Analysis and Recognition*, 2011, pp. 48–52. https://doi.org/10.1109/ICDAR.2011.19
6. Potanin M., Dimitrov D., Shonenkov A., Bataev V., Karachev D., Novopoltsev M., Chertok A. Digital Peter: New dataset, competition and handwriting recognition methods. *Proc. of the HIP'21: The 6th International Workshop on Historical Document Imaging and Processing*, 2021, pp. 43–48. https://doi.org/10.1145/3476887.3476892
7. Shonenkov A., Karachev D., Novopoltsev M., Potanin M., Dimitrov D. StackMix and blot augmentations for handwritten text recognition. *arXiv*, 2021, arXiv:2108.11667. https://doi.org/10.48550/arXiv.2108.11667
8. Teslya N., Mohammed S. Deep learning for handwriting text recognition: Existing approaches and challenges. *Proc. of the 31st Conference of Open Innovations Association (FRUCT)*, 2022, pp. 339–346. https://doi.org/10.23919/FRUCT54823.2022.9770912
9. Bluche T., Messina R. Gated convolutional recurrent neural networks for multilingual handwriting recognition. *Proc. of the 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*. V. 1, 2017, pp. 646–651. https://doi.org/10.1109/ICDAR.2017.111
10. De Sousa Neto A.F., Bezerra B.L.D., Toselli A.H., Lima E.B. HTR-Flor: A deep learning system for offline handwritten text recognition. *Proc. of the 33rd SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, 2020, pp. 54–61. https://doi.org/10.1109/SIBGRAPI51738.2020.00016
11. Shi B., Bai X., Yao C. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017, vol. 39, no. 11, pp. 2298–2304. https://doi.org/10.1109/TPAMI.2016.2646371
12. Bluche T., Louradour J., Messina R. Scan, attend and read: End-to-end handwritten paragraph recognition with MDLSTM attention. *Proc. of the 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*. V. 1, 2017, pp. 1050–1055. https://doi.org/10.1109/ICDAR.2017.174
13. Puigcerver J. Are multidimensional recurrent layers really necessary for handwritten text recognition? *Proc. of the 14th IAPR International Conference on Document Analysis and Recognition (ICDAR). V. 1*, 2017, pp. 67–72. https://doi.org/10.1109/icdar.2017.20
14. Graves A., Fernández S., Gomez F., Schmidhuber J. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. *ICML '06: Proc. of the 23rd International Conference on Machine Learning*, 2006, pp. 369–376. https://doi.org/10.1145/1143844.1143891
15. Li M., Lv T., Chen J., Cui L., Lu Y., Florencio D., Zhang C., Li Z., Wei F. TrOCR: Transformer-based optical character recognition with pre-trained models. *arXiv*, 2021, arXiv:2109.10282. https://doi.org/10.48550/arXiv.2109.10282
16. Dosovitskiy A., Beyer L., Kolesnikov A., Weissenborn D., Zhai X., Unterthiner T., Dehghani M., Minderer M., Heigold G., Gelly S., Uszkoreit J., Houlsby N. An image is worth than 16X16 words: transformers for image recognition. *ICLR 2021*. Available at: https://openreview.net/pdf?id=YicbFdNTTy (accessed: 23.12.2022).

Литература

1. Sánchez J., Romero V., Toselli A.H., Vidal E. ICFHR2016 competition on handwritten text recognition on the READ dataset // Proc. of the 15th International Conference on Frontiers in Handwriting Recognition (ICFHR). 2016. P. 630–635. https://doi.org/10.1109/icfhr.2016.0120
2. Coquenet D., Chatelain C., Paquet T. DAN: a segmentation-free document attention network for handwritten document recognition // IEEE Transactions on Pattern Analysis and Machine Intelligence. 2023. in press. https://doi.org/10.1109/tpami.2023.3235826
3. Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A.N., Kaiser Ł., Polosukhin I. Attention is all you need // Advances in Neural Information Processing Systems 30 (NIPS 2017). 2017. P. 5998–6008.
4. Pletschacher S., Antonacopoulos A. The PAGE (Page Analysis and Ground-truth Elements) format framework // Proc. of the 20th International Conference on Pattern Recognition. 2010. P. 257–260. https://doi.org/10.1109/icpr.2010.72
5. Clausner C., Pletschacher S., Antonacopoulos A. Aletheia - An advanced document layout and text ground-truthing system for production environments // Proc. of the International Conference on Document Analysis and Recognition. 2011. P. 48–52. https://doi.org/10.1109/ICDAR.2011.19
6. Potanin M., Dimitrov D., Shonenkov A., Bataev V., Karachev D., Novopoltsev M., Chertok A. Digital Peter: New dataset, competition and handwriting recognition methods // Proc. of the HIP'21: The 6th International Workshop on Historical Document Imaging and Processing. 2021. P. 43–48. https://doi.org/10.1145/3476887.3476892
7. Shonenkov A., Karachev D., Novopoltsev M., Potanin M., Dimitrov D. StackMix and blot augmentations for handwritten text recognition // arXiv. 2021. arXiv:2108.11667. https://doi.org/10.48550/arXiv.2108.11667
8. Teslya N., Mohammed S. Deep learning for handwriting text recognition: Existing approaches and challenges // Proc. of the 31st Conference of Open Innovations Association (FRUCT). 2022. P. 339–346. https://doi.org/10.23919/FRUCT54823.2022.9770912
9. Bluche T., Messina R. Gated convolutional recurrent neural networks for multilingual handwriting recognition // Proc. of the 14th IAPR International Conference on Document Analysis and Recognition (ICDAR). V. 1. 2017. P. 646–651. https://doi.org/10.1109/ICDAR.2017.111
10. De Sousa Neto A.F., Bezerra B.L.D., Toselli A.H., Lima E.B. HTR-Flor: A deep learning system for offline handwritten text recognition // Proc. of the 33rd SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI). 2020. P. 54–61. https://doi.org/10.1109/SIBGRAPI51738.2020.00016
11. Shi B., Bai X., Yao C. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition // IEEE Transactions on Pattern Analysis and Machine Intelligence. 2017. V. 39. N 11. P. 2298–2304. https://doi.org/10.1109/TPAMI.2016.2646371
12. Bluche T., Louradour J., Messina R. Scan, attend and read: End-to-end handwritten paragraph recognition with MDLSTM attention // Proc. of the 14th IAPR International Conference on Document Analysis and Recognition (ICDAR). V. 1. 2017. P. 1050–1055. https://doi.org/10.1109/ICDAR.2017.174
13. Puigcerver J. Are multidimensional recurrent layers really necessary for handwritten text recognition? // Proc. of the 14th IAPR International Conference on Document Analysis and Recognition (ICDAR). V. 1. 2017. P. 67–72. https://doi.org/10.1109/icdar.2017.20
14. Graves A., Fernández S., Gomez F., Schmidhuber J. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks // ICML '06: Proc. of the 23rd International Conference on Machine Learning. 2006. P. 369–376. https://doi.org/10.1145/1143844.1143891
15. Li M., Lv T., Chen J., Cui L., Lu Y., Florencio D., Zhang C., Li Z., Wei F. TrOCR: Transformer-based optical character recognition with pre-trained models // arXiv. 2021. arXiv:2109.10282. https://doi.org/10.48550/arXiv.2109.10282
16. Dosovitskiy A., Beyer L., Kolesnikov A., Weissenborn D., Zhai X., Unterthiner T., Dehghani M., Minderer M., Heigold G., Gelly S., Uszkoreit J., Houlsby N. An image is worth than 16X16 words: transformers for image recognition // ICLR 2021 [Электронный ресурс]. URL: https://openreview.net/pdf?id=YicbFdNTTy (дата обращения: 23.12.2022).

17. Touvron H., Cord M., Douze M., Massa F., Sablayrolles A., Jégou H. Training data-efficient image transformers & distillation through attention. *arXiv*, 2020, arXiv:2012.12877. https://doi.org/10.48550/arXiv.2012.12877
18. Bao H., Dong L., Wei F. BEiT: BERT Pre-training of image transformers. *arXiv*, 2021, arXiv:2106.08254. https://doi.org/10.48550/arXiv.2106.08254
19. Devlin J., Chang M.W., Lee K., Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding. *Proc. of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL). V. 1*, 2019, pp. 4171–4186. https://doi.org/https://aclanthology.org/N19-1423
20. Liu Y., Ott M., Goyal N., Du J., Joshi M., Chen D., Levy O., Lewis M., Zettlemoyer L., Stoyanov V. RoBERTa: A robustly optimized bert pretraining approach. *arXiv*, 2019, arXiv:1907.11692. https://doi.org/10.48550/arXiv.1907.11692
21. Dong L., Yang N., Wang W., Wei F., Liu X., Wang Y., Gao J., Zhou M., Hon H.-W. Unified language model pre-training for natural language understanding and generation. *Advances in Neural Information Processing Systems 32 (NeurIPS)*, 2019.
22. Singh S.S., Karayev S. Full page handwriting recognition via image to sequence extraction. *Lecture Notes in Computer Science*, 2021, vol. 12823, pp. 55–69. https://doi.org/10.1007/978-3-030-86334-0_4
23. Rouhou A.C., Dhiaf M., Kessentini Y., Ben Salem S. Transformer-based approach for joint handwriting and named entity recognition in historical document. *Pattern Recognition Letters*, 2022, vol. 155, pp. 128–134. https://doi.org/10.1016/j.patrec.2021.11.010
24. Schreiber S., Agne S., Wolf I., Dengel A., Ahmed S. DeepDeSRT: Deep learning for detection and structure recognition of tables in document images. *Proc. of the 14th IAPR International Conference on Document Analysis and Recognition (ICDAR). V. 1*, 2017, pp. 1162–1167. https://doi.org/10.1109/icdar.2017.192
25. Ares Oliveira S., Seguin B., Kaplan F. DhSegment: A generic deep-learning approach for document segmentation. *Proc. of the 16th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, pp. 7–12. https://doi.org/10.1109/icfhr-2018.2018.00011
26. Yang X., Yumer E., Asente P., Kraley M., Kifer D., Giles C.L. Learning to extract semantic structure from documents using multimodal fully convolutional neural networks. *Proc. of the 30th IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 4342–4351. https://doi.org/10.1109/cvpr.2017.462
27. Xu Y., Li M., Cui L., Huang S., Wei F., Zhou M. LayoutLM: Pre-training of text and layout for document image understanding. *Proc. of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2020, pp. 1192–1200. https://doi.org/10.1145/3394486.3403172
28. Xu Y., Xu Y., Lv T., Cui L., Wei F., Wang G., Lu Y., Florencio D., Zhang C., Che W., Zhang M., Zhou L. LayoutLMv2: Multi-modal pre-training for visually-rich document understanding. *Proc. of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing. V. 1*, 2021, pp. 2579–2591. https://doi.org/10.18653/v1/2021.acl-long.201
29. Biswas S., Banerjee A., Lladós J., Pal U. DocSegTr: An Instance-level end-to-end document image segmentation transformer. *arXiv*, 2022, arXiv:2201.11438. https://doi.org/10.48550/arXiv.2201.11438
30. Li J., Xu Y., Lv T., Cui L., Zhang C., Wei F. DiT: Self-supervised pre-training for document image transformer. *MM '22: Proc. of the 30th ACM International Conference on Multimedia*, 2022, pp. 3530–3539. https://doi.org/10.1145/3503161.3547911
31. Coquenet D., Chatelain C., Paquet T. End-to-end handwritten paragraph text recognition using a vertical attention network. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023, vol. 45, no. 1, pp. 508–524. https://doi.org/10.1109/tpami.2022.3144899
32. Grosicki E., Abed H.E. ICDAR 2011 - French handwriting recognition competition. *Proc. of the International Conference on Document Analysis and Recognition (ICDAR)*, 2011, pp. 1459–1463. https://doi.org/10.1109/icdar.2011.290
33. Everingham M., Gool Van L., Williams C.K.I., Winn J. The PASCAL Visual Object Classes (VOC) Challenge. *International Journal of Computer Vision*, 2010, vol. 8, no. 2, pp. 303–338. https://doi.org/10.1007/s11263-009-0275-4

17. Touvron H., Cord M., Douze M., Massa F., Sablayrolles A., Jégou H. Training data-efficient image transformers & distillation through attention // arXiv. 2020. arXiv:2012.12877. https://doi.org/10.48550/arXiv.2012.12877
18. Bao H., Dong L., Wei F. BEiT: BERT Pre-training of image transformers // arXiv. 2021. arXiv:2106.08254. https://doi.org/10.48550/arXiv.2106.08254
19. Devlin J., Chang M.W., Lee K., Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding // Proc. of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL). V. 1. 2019. P. 4171–4186. https://doi.org/https://aclanthology.org/N19-1423
20. Liu Y., Ott M., Goyal N., Du J., Joshi M., Chen D., Levy O., Lewis M., Zettlemoyer L., Stoyanov V. RoBERTa: A robustly optimized bert pretraining approach // arXiv. 2019. arXiv:1907.11692. https://doi.org/10.48550/arXiv.1907.11692
21. Dong L., Yang N., Wang W., Wei F., Liu X., Wang Y., Gao J., Zhou M., Hon H.-W. Unified language model pre-training for natural language understanding and generation // Advances in Neural Information Processing Systems 32 (NeurIPS). 2019.
22. Singh S.S., Karayev S. Full page handwriting recognition via image to sequence extraction // Lecture Notes in Computer Science. 2021. V. 12823. P. 55–69. https://doi.org/10.1007/978-3-030-86334-0_4
23. Rouhou A.C., Dhiaf M., Kessentini Y., Ben Salem S. Transformer-based approach for joint handwriting and named entity recognition in historical document // Pattern Recognition Letters. 2022. V. 155. P. 128–134. https://doi.org/10.1016/j.patrec.2021.11.010
24. Schreiber S., Agne S., Wolf I., Dengel A., Ahmed S. DeepDeSRT: Deep learning for detection and structure recognition of tables in document images // Proc. of the 14th IAPR International Conference on Document Analysis and Recognition (ICDAR). V. 1. 2017. P. 1162–1167. https://doi.org/10.1109/icdar.2017.192
25. Ares Oliveira S., Seguin B., Kaplan F. DhSegment: A generic deep-learning approach for document segmentation // Proc. of the 16th International Conference on Frontiers in Handwriting Recognition (ICFHR). P. 7–12. https://doi.org/10.1109/icfhr-2018.2018.00011
26. Yang X., Yumer E., Asente P., Kraley M., Kifer D., Giles C.L. Learning to extract semantic structure from documents using multimodal fully convolutional neural networks // Proc. of the 30th IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017. P. 4342–4351. https://doi.org/10.1109/cvpr.2017.462
27. Xu Y., Li M., Cui L., Huang S., Wei F., Zhou M. LayoutLM: Pre-training of text and layout for document image understanding // Proc. of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2020. P. 1192–1200. https://doi.org/10.1145/3394486.3403172
28. Xu Y., Xu Y., Lv T., Cui L., Wei F., Wang G., Lu Y., Florencio D., Zhang C., Che W., Zhang M., Zhou L. LayoutLMv2: Multi-modal pre-training for visually-rich document understanding // Proc. of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing. V. 1. 2021. P. 2579–2591. https://doi.org/10.18653/v1/2021.acl-long.201
29. Biswas S., Banerjee A., Lladós J., Pal U. DocSegTr: An Instance-level end-to-end document image segmentation transformer // arXiv. 2022. arXiv:2201.11438. https://doi.org/10.48550/arXiv.2201.11438
30. Li J., Xu Y., Lv T., Cui L., Zhang C., Wei F. DiT: Self-supervised pre-training for document image transformer // MM '22: Proc. of the 30th ACM International Conference on Multimedia. 2022. P. 3530–3539. https://doi.org/10.1145/3503161.3547911
31. Coquenet D., Chatelain C., Paquet T. End-to-end handwritten paragraph text recognition using a vertical attention network // IEEE Transactions on Pattern Analysis and Machine Intelligence. 2023. V. 45. N 1. P. 508–524. https://doi.org/10.1109/tpami.2022.3144899
32. Grosicki E., Abed H.E. ICDAR 2011 - French handwriting recognition competition // Proc. of the International Conference on Document Analysis and Recognition (ICDAR). 2011. P. 1459–1463. https://doi.org/10.1109/icdar.2011.290
33. Everingham M., Gool Van L., Williams C.K.I., Winn J. The PASCAL Visual Object Classes (VOC) Challenge // International Journal of Computer Vision. 2010. V. 8. N 2. P. 303–338. https://doi.org/10.1007/s11263-009-0275-4

34. Sánchez J.A., Romero V., Toselli A.H., Vidal E. ICFHR2016 competition on handwritten text recognition on the READ dataset. *Proc. of the 15th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, 2016, pp. 630–635. https://doi.org/10.1109/icfhr.2016.0120

34. Sánchez J.A., Romero V., Toselli A.H., Vidal E. ICFHR2016 competition on handwritten text recognition on the READ dataset // Proc. of the 15th International Conference on Frontiers in Handwriting Recognition (ICFHR). 2016. P. 630–635. https://doi.org/10.1109/icfhr.2016.0120

Authors

Samah Mohammed — PhD Student, ITMO University, Saint Petersburg, 197101, Russian Federation, SC 57700909900, https://orcid.org/0000-0002-8009-7222, Samahslimanmhmd93@gmail.com
Nikolay Teslya — PhD, Senior Researcher, St. Petersburg Federal Research Center of the Russian Academy of Sciences, Saint Petersburg, 199178, Russian Federation, SC 56946917500, https://orcid.org/0000-0003-0619-8620, teslya@iias.spb.su

Авторы

Мохаммед Самах — аспирант, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, SC 57700909900, https://orcid.org/0000-0002-8009-7222, Samahslimanmhmd93@gmail.com
Тесля Николай — кандидат технических наук, старший научный сотрудник, Санкт-Петербургский Федеральный исследовательский центр Российской академии наук, Санкт-Петербург, 199178, Российская Федерация, SC 56946917500, https://orcid.org/0000-0003-0619-8620, teslya@iias.spb.su

Received 09.11.2022
Approved after reviewing 01.03.2023
Accepted 16.05.2023

Статья поступила в редакцию 09.11.2022
Одобрена после рецензирования 01.03.2023
Принята к печати 16.05.2023