Computational Linguistics and Intellectual Technologies:
Proceedings of the International Conference "Dialogue 2023"

June 14–16, 2023

HWR200: New open access dataset of handwritten texts images in Russian

Ivan Potyashin
Antiplagiat
potyashin@ap-team.ru

Mariam Kaprielova
Antiplagiat, FRC CSC RAS
kaprielova@ap-team.ru

Yury Chekhovich
Antiplagiat, FRC CSC RAS
chehovich@ap-team.ru

Alexandr Kildyakov
Antiplagiat
kildyakov@ap-team.ru

Temirlan Seil
Antiplagiat
seilov@ap-team.ru

Evgeny Finogeev
Antiplagiat
finogeev@ap-team.ru

Andrey Grabovoy
Antiplagiat, FRC CSC RAS
grabovoy@ap-team.ru

Abstract

Handwritten text image datasets are highly useful for solving many problems using machine learning. Such problems include recognition of handwritten characters and handwriting, visual question answering, near-duplicate detection, search for text reuse in handwriting and many auxiliary tasks: highlighting lines, words, other objects in the text. The paper presents new dataset of handwritten texts images in Russian created by 200 writers with different handwriting and photographed in different environment[1]. We described the procedure for creating this dataset and the requirements that were set for the texts and photos. The experiments with the baseline solution on fraud search and text reuse search problems showed results of results of 60% and 83% recall respectively and 5% and 2% false positive rate respectively on the dataset.

Keywords: OCR; handwriting; text reuse detection; computer vision; handwritten text recognition; HTR
DOI: 10.28995/2075-7182-2023-22-452-458

1 Introduction

There are many different tasks that require working with images of handwritten texts. For example, visual question answering (Mathew et al., 2020). Optical character recognition (OCR) for handwritten texts is an important (Nurseitov et al., 2021) and challenging problem(Yousef and Bishop, 2020; Coquenet et al., 2023), especially in languages where there is a lack of labelled data. An example of such a case can be any Cyrillic language. The solution to this problem can be applied in many areas: healthcare (Fogel et al., 2020), education (Yanikoglu, 2017; Bakhteev et al., 2021), digitization of historical documents (Wigington et al., 2018).

Standard datasets for the OCR problem are IAM (Marti, 2002) and Bentham (Gatos et al., 2014). The IAM-database is based on the Lancaster-Oslo/Bergen corpus. It consists of 13353 images with handwritten lines of text written in special forms by 657 people. The database is labeled at the sentence, line, and word levels. The total number of words in the collection is 115320. Bentham dataset contains over 6000 documents and over 25000 pages of text written by a philosopher Jeremy Bentham. An analogue of (Gatos et al., 2014) in Russian is Digital Petr (Potanin et al., 2021). It, like Bentham, consists of scanned historical documents written by a single person with line-level text segmentation. It contains about, 10000 image-text pairs corresponding to lines in historical documents.

The datasets (Potanin et al., 2021; Gatos et al., 2014) have a feature that all the texts are written by one person. Such data can be useful for solving problems of recognition of texts written in the same handwriting. Also, it can be helpful for researchers to compare different handwriting recognition models. But if one needs to develop a model that deals with different handwriting these datasets may not be the best choice.

Datasets of another type, which contain different handwritings are IDP-forms (idp,), HKR (Nurseitov et al., 2021), school_notebooks (sch,). When compiling the sber-idp-forms dataset, the assessors were asked to manually write the given words or phrases on special forms. In total, there are 5203 images of rectangles with written text and their annotations in this dataset. The collection may be used for text segmentation, handwriting recognition, writer identification and writer verification tasks. When creating

[1] Our dataset is available at: https://huggingface.co/datasets/AntiplagiatCompany/HWR200

the HKR collection, the same idea with form was used. The dataset includes 63000 phrases written by 200 people. The text is written in Russian and Kazakh where about 95% is presented in Russian. The datasets (idp, ; Nurseitov et al., 2021) consist of handwritten texts digitized in the same environment, while in some cases it is natural to work with handwritten texts converted to an image in different ways.

Another significant problem is reuse in handwritten texts such as essays in schools and universities (Wrigley, 2019). We separate this problem into two categories. The first one is text reuse and the second one is submitting the same writing photographed in a different environment which will be further referred to as *fraud*.

The collection (sch,) contains 1857 images of school notebooks with word-level polygonal markup. This dataset is quite small, and as in the previous datasets, the environment in which the photos are taken is the same.

Cyrillic languages are not so actively studied in the problem of handwritten text recognition (HTR). There is a small amount of marked up data in these languages. Thus for the Russian language there are no approaches with sufficient quality of handwriting recognition.

For a text reuse search task, even a small number of typos in the text leads to a degradation in the search quality. Thus it is important to have either high-quality OCR or a new approach that takes into account the imperfection of the OCR model.

To create new models, it is essential to have a highly diverse dataset of handwritten texts. We contribute to solving handwritten text recognition problem by introducing the HWR200 dataset: a collection of handwritten texts in Russian for HTR and the search for reuse in handwritten texts. This collection provides texts created by 200 different writers, photographed under different conditions. The peculiarity of this dataset is that the same texts were written by more than one assessors, and this information can be used when training a robust OCR model. Texts are written by 200 assessors and photographed in three different ways. In total there are 30030 images with handwritten texts in our dataset.

2 Description of the dataset

2.1 Text generation algorithm and markup structure

The basis of the dataset is 35 different unique texts further referred to as *originals*. They are used to generate most of the dataset: texts further referred to as *reuses*. The *reuses* consist of two types of sentences: sentences that appeared in *original* texts and unique sentences. The text generation algorithm is as follows:

1. We generate a number from 28 to 32 for the amount of sentences to be reused;
2. We randomly select one or two *originals* to be used;
3. We generate how many sentences will be taken from the first *original* text and how many from the second;
4. Unique sentences are added to the beginning, to the end, or both to the beginning and to the end.

In total, 2650 *reuses* are generated. In addition, there are 35 more unique texts further referred to as *fprs*. An example of json with metadata:

```
// for original texts:
{
    sentences: [{id: <id>, text: <sentence>}, ...],
    words_count: <word count>,
    full_text: <full text>
}

// for reuse texts:
{
    reuse_0: {
        sentences: [{id: <id>, text: <sentence>}, ...],
        id: <original text file name>
        intersection_score: <intersection_score>
```

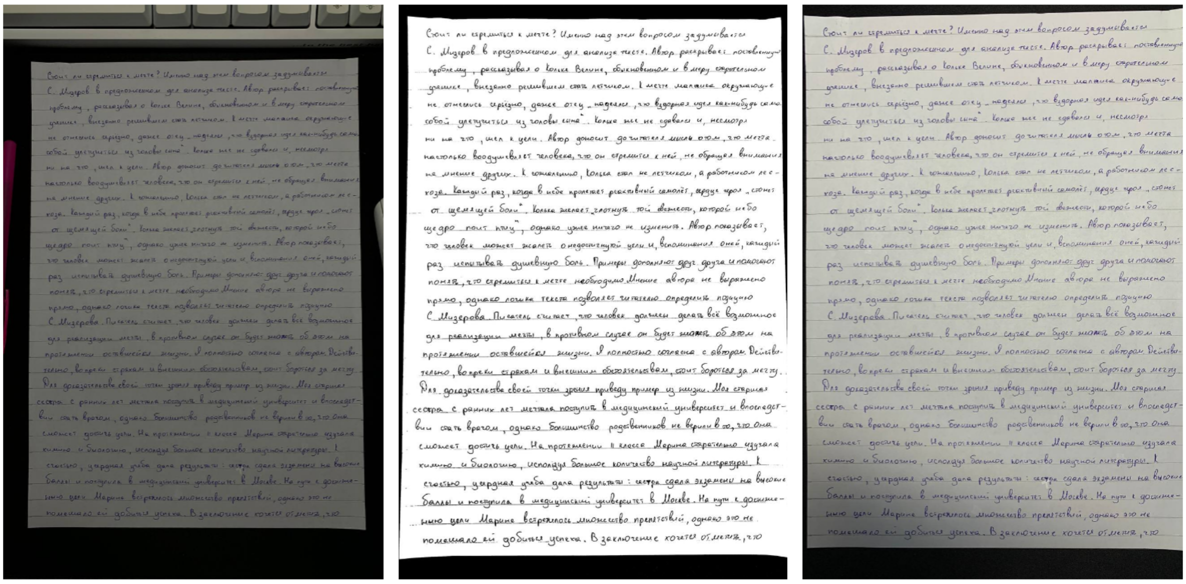

Figure 1: Three types of images: (left) photographed in poor light and with other objects, (center) scanned, (right) photographed in good light and without other objects.

```
    }
    reuse_1: {  // if exists
        sentences: [{id: <id>, text: <sentence>}, ...],
        id: <original text file name>
        intersection_score: <intersection_score>
    }
    start clear sentences: [<sentence>, <sentence>, ...]  // if exists
    end clear sentences: [<sentence>, <sentence>, ...]  // if exists
    words_count: <word count>
    full_text: <full text>
}

// for fpr texts:
{
    sentences: [{id: <id>, text: <sentence>}, ...],
    words_count: <word count>,
    full_text: <full text>
}
```

2.2 Image types

Each page of handwritten text had to be converted into an image in three different ways. First, it had to be scanned. Second, the assessor had to take a photo in good light. There should have been no glare, the page should not be cut off, extra objects should not fall into the frame. Third, the text had to be photographed in poor light. In this case, it was desirable that objects on the table fall into the frame, but the main part of the frame should have been occupied by the page. It was important that each page fits completely into the frame. See examples in Figure 1.

	Ours	**Bentham, Digital Petr**	**IAM**	**School notebooks**	**Sber-idp-forms, HKR**
Texts or phrases	texts	texts	phrases	texts	phrases
Word / line level markup	-	+	+	+	+
Different handwriting	+	-	+	+	+
Different environment	+	-	-	-	-

Table 1: Characteristics of datasets

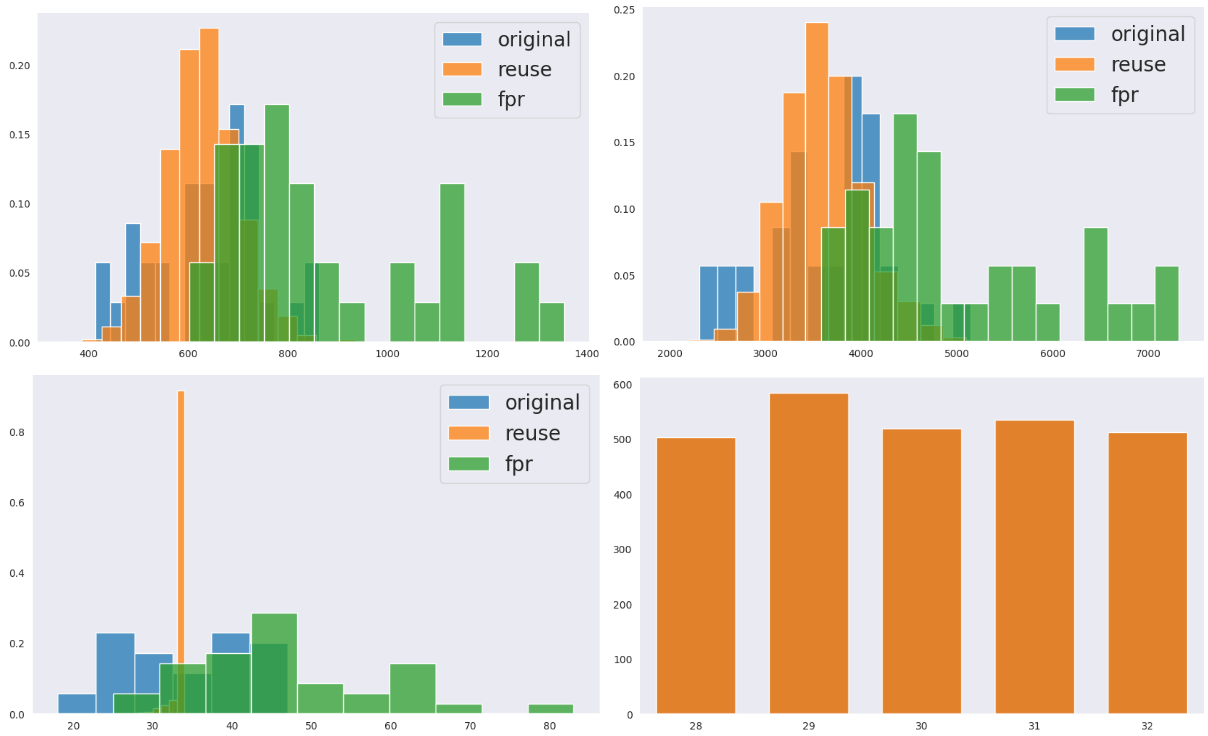

Figure 2: (top left) distribution of the number of words in texts. (top right) distribution of the lengths of texts. (bottom left) distribution of the number of sentences in texts. (bottom right) distribution of the number of duplicate sentences in *reuses*.

2.3 Distribution of texts by assessors

Each of 200 assessors wrote 15 texts. The first 175 assessors wrote one *original*, one *fpr* and 13 *reuses*, at that each *original* and each *fpr* are handwritten by 5 assessors. The rest 25 wrote 15 *reuses*. Thus 2650 *reuses* are handwritten once and 35 *originals* and 35 *reuses* are handwritten five times.

2.4 Characteristics of the dataset

The dataset contains 2720 handwritten texts with an average word count of 631, an average text length of 3617, and an average sentence count of 34. In addition, 47% of texts with duplicates have two *original* essays, the rest have only one. The distribution is show in Figure 2. The total number of images with text is 30030, so on average each text takes up 3.3 pages. A comparative table with the characteristics of various datasets can be seen in Table 1.

Dataset	Task	Recall@1	FPR
Ours	fraud detection	60%	5%
Bentham+IAM	fraud detection	80%	5%
Ours	reuse detection	83%	3%

Table 2: Results of our baseline solutions.

3 Experiments

We trialled the HWR200 dataset in two tasks: fraud search and text reuse detection. To evaluate our solutions we used recall@1 and fpr metrics. These tasks are actually binary classification tasks: is a given image a reuse or a fraud or not. Formally, metrics are defined as follows:

$$recall@1 = \frac{TP}{TP + FN}, \tag{1}$$

$$fpr = \frac{FP}{FP + TN}, \tag{2}$$

where TP is the number of true positive predictions, FN is the number of false negative predictions, FP is the number of false positive predictions, and TN is the number of true negative predictions. Results can bee seen in Table 2.

3.1 Fraud detection

As described above, every handwritten page is photographed three times in three different ways. We considered one of them as an original, and the other two as fraud. So, the task for each fraud page is to find the original page.

Our baseline solution for this task consists of three stages: embedding generation, candidate search and similarity estimation between query and candidates to find the closest one. We use a neural network to transform a handwritten document into embedding. For candidates search we use Faiss framework (Johnson et al., 2017) (the faiss index is filled with original photos) and similarity estimation is performed using deep learning approach inspired by (Sun et al., 2021).

This approach showed 60% recall@1 and 5% fpr. Similar approach on IAM and Bentham dataset showed 80% recall and 5% fpr. It should be taken into account that in that experiment, fraud images were generated from images in the dataset, whereas in our experiment, fraud images are part of the dataset.

3.2 Reuse detection

Every *reuse* contains some sentences from one or two *original* texts. The task for every *reuse* text is to find at least one original text.

The solution for this task also consists of three stages. First, the algorithm tries to recognize handwritten text. The input page is divided into lines, and text is extracted from each line using a deep learning OCR model optimized in supervised learning mode inspired by (Coquenet et al., 2020). Second, we split the text into bigrams and search for candidates based on them using a shingle index based on (Broder et al., 1997; Broder, 1997). Last, we compare the candidates with the input text and find the text with the highest reuse rate.

This approach showed 83% recall and 2% false positive rate.

4 Conclusion

We have introduced the dataset of handwritten texts in Russian. This collection contains texts written in different handwriting and photographed under various conditions. One of the key features of this collection is that one text can be written by several assessors, which may be very useful for tasks where models have to be robust. Besides, the dataset can also be helpful for solving more specific tasks such as text reuse search or near-duplicate detection.

Acknowledgements

This research was supported by FASIE (Project No. 79068, application No. 208298).

References

Oleg Bakhteev, Dorodnicyn CC Antiplagiat, Rita Kuznetsova, Andrey Khazov, Aleksandr Ogaltsov, Kamil Safin, Tatyana Gorlenko, Marina Suvorova, Andrey Ivahnenko, Pavel Botov, et al. 2021. Near-duplicate handwritten document detection without text recognition.

Andrei Z. Broder, Steven C. Glassman, Mark S. Manasse, and Geoffrey Zweig. 1997. Syntactic clustering of the web, Sep.

A.Z. Broder. 1997. On the resemblance and containment of documents.

Denis Coquenet, Clément Chatelain, and Thierry Paquet. 2020. End-to-end handwritten paragraph text recognition using a vertical attention network. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45:508–524.

Denis Coquenet, Clément Chatelain, and Thierry Paquet. 2023. Dan: a segmentation-free document attention network for handwritten document recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Sharon Fogel, Hadar Averbuch-Elor, Sarel Cohen, Shai Mazor, and Roee Litman. 2020. Scrabblegan: Semi-supervised varying length handwritten text generation. // *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June.

Basilis Gatos, Georgios Louloudis, Tim Causer, Kris Grint, Verónica Romero, Joan Andreu Sánchez, Alejandro H. Toselli, and Enrique Vidal. 2014. Ground-truth production in the transcriptorium project. // *2014 11th IAPR International Workshop on Document Analysis Systems*, P 237–241.

Idp-forms (2021). Available at: `https://github.com/ai-forever/htr_datasets/tree/main/IDP-forms`.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2017. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, 7:535–547.

U.-V. Marti. 2002. The iam-database: an english sentence database for offline handwriting recognition. *International Journal on Document Analysis and Recognition*, 28(1):114–133.

Minesh Mathew, Dimosthenis Karatzas, R. Manmatha, and C. Jawahar. 2020. Docvqa: A dataset for vqa on document images. 07.

Daniyar Nurseitov, Kairat Bostanbekov, Daniyar Kurmankhojayev, Anel Alimova, Abdelrahman Abdallah, and Rassul Tolegenov. 2021. Handwritten kazakh and russian (hkr) database for text recognition. *Multimedia Tools and Applications*, P 1–23.

M. B. Potanin, Denis Dimitrov, A. Shonenkov, Vladimir Bataev, Denis Karachev, and Maxim Novopoltsev. 2021. Digital peter: New dataset, competition and handwriting recognition methods. *The 6th International Workshop on Historical Document Imaging and Processing*.

School_notebooks_ru (2021). Available at: `https://github.com/ai-forever/htr_datasets/tree/main/school_notebooks`.

Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. 2021. LoFTR: Detector-free local feature matching with transformers. *CVPR*.

Curtis Wigington, Chris Tensmeyer, Brian Davis, William Barrett, Brian Price, and Scott Cohen. 2018. Start, follow, read: End-to-end full-page handwriting recognition. // *Computer Vision – ECCV 2018"*, P 372–388, Cham. Springer International Publishing.

Stuart Wrigley. 2019. Avoiding ‘de-plagiarism’: Exploring the affordances of handwriting in the essay-writing process. *Active Learning in Higher Education*, 20(2):167–179.

Berrin Yanikoglu. 2017. Use of handwriting recognition technologies in tablet-based learning modules for first grade education. *Educational Technology Research and Development*.

Mohamed Yousef and Tom E. Bishop. 2020. Origaminet: Weakly-supervised, segmentation-free, one-step, full page textrecognition by learning to unfold. // *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June.