On recognition of Cyrillic Text

Kostiantyn Liepieshov
The Machine Learning Lab, Ukrainian Catholic University
Lviv, Ukraine
liepieshov@ucu.edu.ua

Oles Dobosevych
The Machine Learning Lab, Ukrainian Catholic University
Lviv, Ukraine
dobosevych@ucu.edu.ua

Abstract

We introduce the largest (among publicly available) dataset for Cyrillic Handwritten Text Recognition and the first dataset for Cyrillic Text in the Wild Recognition, as well as suggest a method for recognizing Cyrillic Handwritten Text and Text in the Wild. Based on this approach, we develop a system that can reduce the document processing time for one of the largest mathematical competitions in Ukraine by 12 days and the amount of used paper by 0.5 ton.

1 Introduction

Text is one of the main ways to transfer the information and it can take many forms. It can be handwritten or printed, in the form of business documents, notes, bills, historical documents, advertisements, logos etc. Therefore, the method for its recognition should be flexible enough to work with different text styles and under the different external conditions. Although for the English language the task of text recognition is well studied [1], [2], for Cyrillic languages such studies are almost missing, the main reason being the lack of extensive publicly available datasets. To the best of our knowledge, the only public Cyrillic dataset consists only of individual letters [3], while others [1], [4], [5], are unavailable.

In our research, we will focus on developing a single model for Handwritten Text Recognition and Text Recognition in the Wild, as the extreme case of printed text.

2 Datasets

Currently the amount of data provided by the existing Cyrillic datasets for text recognition is insufficient for training deep networks. That is why we have created a synthetic dataset and annotated datasets for Handwritten Text and Text in the Wild recognition.

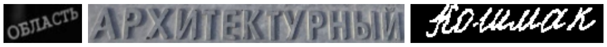

Figure 1: Samples of datasets: Synthetic Cyrillic text, Cyrillic Text in the Wild, Handwritten Cyrillic text (from left to right)

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

2.1 Synthetic Cyrillic text

We approach the problem of generating Synthetic Cyrillic text recognition dataset by adapting the idea proposed by Jaderberg et al. [6], but in our Cyrillic setup. For the text sampling stage, a word is randomly sampled from the UberText Corpus with the requirement to include only Cyrillic letters. For the font rendering, a font is randomly selected from the subset of fonts from UKR-Fonts, Google Fonts, Font Squirrel including all glyphs needed for generation of the sampled word. In total, the dataset consists of 881309 samples of 180994 different words (Figure 1).

2.2 Cyrillic Text in the Wild

As mentioned above, at the moment there are no existing datasets for Text in the Wild recognition of Cyrillic words. We addressed this problem by collecting the dataset by annotating photos from the Internet depicting different cities all around Ukraine. The dataset consist of 505 samples in different orientations from 151 different photos and in total has 454 different words (Figure 1). The extended version of this dataset consists not only of rectangles of words, but also their locations in the images (although this information is not used in the research).

2.3 Handwritten Cyrillic text

The handwritten dataset was collected by processing data from one of the largest mathematical competitions in Ukraine. The dataset is extracted from the forms that were filled by children aged 7 to 18 from different parts of Ukraine and contains mainly their surnames. It consists of 82061 samples and 37007 words (Figure 1) and is divided into 3 parts (train 60%, validate 20% and test 20%), with the same distribution of each word in each part of the dataset.

3 Proposed method and experiments

We developed a single model that was pretrained on Synthetic Cyrillic text and then finetuned on a mixture of Synthetic Cyrillic text and Handwritten Cyrillic text. The architecture consists of two Conv blocks with CRelu and Relu followed by six Conv blocks with Instance Normalisation and Leaky Relu. The obtained word error rate and character error rate are reported in Table 1.

Table 1: Model results

	Word Error Rate	Char Error Rate
Cyrillic Text in the Wild	59.8	27.6
Handwritten Cyrillic text (test)	21.1	4.8

4 Industry Application

The method was used for the development of the new system for processing responses of participants of one of the largest math competitions in Ukraine (International Mathematics Contest "Kangaroo"). The new system allows to remove part of the form that was used by participants for manual annotation of their surnames and to reduce the size of the form from B5 to A5. This competition is organized twice a year and attracts around 500000 participants each time. The new system will allow to reduce the processing (scanning) time by 12 days, and the amount of used paper by 0.5 tons on each round.

5 Conclusion

We collected unique datasets that can advance Cyrillic text OCR methods in different forms, developed the method that achieves the quality comparable to English text on Handwritten Text Recognition [7], and use this method to solve one industry problem.

Acknowledgements: The authors were supported by ELEKS Ltd. and Ukrainian Catholic University.

References

[1] Victor Carbune et al. Fast multi-language lstm-based online handwriting recognition. 2019.

[2] Bušta et al. E2e-mlt - an unconstrained end-to-end method for multi-language scene text. 2019.

[3] Grégory Vial. Comnist: a crowd-sourced version of nist dataset for cyrillic and latin alphabets, 2017.

[4] Elmira Mustakimova. Offline recognition of russian handwriting, 2016.

[5] Daniel et al. Multi-language online handwriting recognition. 2016.

[6] Jaderberg et al. Synthetic data and artificial neural networks for natural scene text recognition. 2014.

[7] Ingle et al. A scalable handwritten text recognition system. 2019.