УДК 004.932.75'1

ADVANCING HANDWRITTEN PARAGRAPH RECOGNITION: A SPATIAL APPROACH WITH THE RUSSIAN NOTEBOOKS DATASET

Mohammed.S (ITMO university)
Supervisor –Associate Professor Ph. D, Teslya N.N
(SPC RAS)

Introduction. Handwritten paragraph recognition plays a crucial role in improving accuracy and usability in various applications within handwritten document analysis. However, recognizing paragraphs in handwritten documents poses challenges due to layout variations. Accurate paragraph segmentation relies on spatial information, encompassing the relationships between text elements. While recent studies in handwritten Russian recognition have mainly focused on character and line levels, this research pioneers paragraph-level recognition. The Vertical Attention Network (VAN) [1], utilizing a hybrid attention method, is employed for this purpose. The study introduces the first Russian dataset at the paragraph level, comprising approximately 6293 paragraph images with PAGE XML-encoded ground truth, it is prepared from the Russian school notebooks' dataset [2] at word level, which contains approximately 1857 images detailed in JSON format. The VAN model is fine-tuned for comprehensive paragraph recognition, and its performance is compared with alternative non-layout-aware approaches. This work represents a significant advancement in layout-aware recognition for handwritten Russian documents, addressing an underexplored area in the field.

Main part. Due to the absence of available datasets for recognizing Russian handwriting at the paragraph level, and with existing datasets primarily focused on word-level, such as the Russian school notebooks' dataset, or line-level, as seen in the 'Digital Peter' dataset [3], we prepare the First Russian handwritten dataset at the paragraph level.

The dataset aimed to facilitate paragraph-level recognition for Russian handwritten documents. Leveraging the Russian school notebooks' dataset as the source, it contained 1557 training, 150 validation, and 150 testing images from various Russian students, primarily double-page, with COCO format annotations. The ground truth, provided in JSON format, delineated categories, image details, and annotations for words in text lines. Preparing this dataset entailed a Python script to extract word-level ground truth using baseline points and word polygons, merging polygons to construct line coordinates. Next, it segmented lines onto respective pages based on baseline point inclusion within page polygons, handling potential overlaps between page margins. Finally, it separated pages into paragraphs by considering shared horizontal coordinates and minimal vertical distances between lines, employing convex hull polygons to enclose each paragraph's coordinates. Finally, the prepared dataset contains 5375 training, 450 validation, 468 testing paragraph images.

For dataset evaluation, we opted for the VAN model as it is the first end-to-end encoder-decoder segmentation-free architecture using hybrid attention. The model is designed for joint recognition of text and layout in handwritten documents, with a focus on capturing the spatial relationships between text lines within a paragraph.

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. It is the sum of levenshtein distance (d_{lev}) among the ground truth y^{text} and the predictions $\hat{y}^{text}$ normalized by the total length of the ground truth $y^{text}_{len_i}$. Word Error Rate (WER) is also used to evaluate text recognition and it is computed in the same way but at word level where we also consider punctuation characters as words. Since all the proposed approaches to recognize Russian handwritten at line level have been conducted on Digital Peter dataset only. Thus, there are no results reported in the literature on the any Russian datasets at paragraph level and comparisons with approaches under similar conditions can't be conducted as this work is the first attempt to recognize Russian handwritten dataset at paragraph level.

The VAN model achieves 20.87% of CER and 38.52 % of WER on the test set compared to 17.69% of CER and 32.61% of WER on the valid set. The values of CER, WER are slightly worse for the line level dataset, and this can be explained by the different complex layouts and irregularity such as multi-column text lines. The challenges in this domain (Paragraph recognition) are substantial, stemming from the variations in layout, line spacing, and textual arrangements within handwritten documents.

Conclusion. Key contributions of this work include the preparation of the first Russian dataset at the paragraph level, The meticulous annotation in PAGEXML format servs as a benchmark for training and evaluating recognition models.

Also, we fine-tune the VAN model for whole paragraph recognition. Comprehensive experiments have been conducted at line level and paragraph level. We couldn't reach state-of-the-art recognition results at line level, but the results still indicate promising performance in paragraph recognition, the complexity of handwritten remains a challenge for future research.

References:

1. D. Coquenet, C. Chatelain, and T. Paquet, "End-to-end handwritten paragraph text recognition using a vertical attention network," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 45, no. 1,2023, pp. 508–524.
2. Ai-Forever, "Ai-forever/school_notebooks_ru · datasets at hugging face," ai-forever/school_notebooks_RU · Datasets at Hugging Face, https://huggingface.co/datasets/ai-forever/school_notebooks_RU (accessed Sep. 10, 2023).
3. M. Potanin et al., "Digital peter: New dataset, competition and handwriting recognition methods," The 6th International Workshop on Historical Document Imaging and Processing, 2021, pp. 43–48.