Recognition of handwritten archives by A. V. Sukhovo-Kobylin

Морозов Иван Дмитриеивч
ВМК МГУ
morozov-ivan-2003@yandex.ru
&
ВМК МГУ
mestlm@mail.ru

2024

Аннотация

This study addresses the problem of recognizing historical handwritten archives. Since there is no universal model that can optimally process such documents due to their unique characteristics, recognition becomes more effective when models are specifically tailored to the handwriting of a particular author. The aim of this work is to develop training methods and models that are particularly effective for the archive of Alexander Vasilyevich Sukhovo-Kobylin's diaries. This archive presents specific challenges: densely packed lines, multilingual text, and missing words or symbols in the annotations. To overcome these difficulties, various approaches are proposed to improve the overall quality of recognition.

1 Introduction

1.1 Text Recognition Approaches

There are two primary approaches to recognizing text on pages:

- **Line Segmentation and Recognition**

  This approach involves using a separate model or method to detect text lines, which are then passed to a recognition model. Training requires labeled images with line annotations and data on line positions within the page. Solutions in this domain often tackle text recognition in historical archives with a focus on line-based processing.

  For instance, in the *Peter the Great Archives* (1), the main contribution was creating a dataset based on historical handwritten texts from Peter the Great's era. The recognition problem was also framed as a competition, where the primary task involved line recognition. The top-performing model, described in the paper, employed a Convolutional Recurrent Neural Network (CRNN) with beam search and an N-gram language model, trained with Connectionist Temporal Classification (CTC) loss.

In another example, *Yandex: Archive Search* prepared historical documents like metric books and census records for genealogical research. Their model segmented lines using a Gaussian heatmap and performed recognition with a convolutional encoder and an RNN decoder with attention, trained using cross-entropy loss. Line blocks were detected using an Instance Segmentation model.

A promising architectural advancement is found in the *TrOCR* paper (3), which introduced a transformer-based encoder-decoder architecture. This model bypasses convolutional layers, instead feeding reduced images divided into patches into the encoder, with the decoder generating byte-pair encoding (BPE) tokens. It is trained with cross-entropy loss, utilizing pretrained Vision Transformer (ViT) and RoBERTa models for initialization.

- **Page-Level Text Recognition**

  A more recent and emerging approach is end-to-end page recognition. This method simplifies the annotation process and enhances model usability, although it typically assumes a single-column text structure.

  One of the simpler architectures is *SPAN* (2), which uses an encoder composed of convolutional blocks with pooling and depthwise convolutional layers with residual connections. The decoder, a single convolutional layer, generates a 2D character probability matrix, which is then converted into a sequence of character probability vectors using horizontal concatenation. The sequence is processed with CTC to produce the final text output.

  Vertical Attention Network (VAN) (5) builds on a convolutional encoder and introduces a vertical attention mechanism, allowing for sequential aggregation of line representations from the page. These representations are transformed into probability matrices by a convolutional and LSTM decoder. VAN achieves state-of-the-art performance on public datasets and serves as the baseline model for this study.

  OrigamiNet (4) presents an alternative fully convolutional architecture. The model comprises a convolutional backbone and the OrigamiNet module. This module receives the image representation as a 3D tensor and applies vertical expansion, horizontal contraction, and convolutions to generate a vertical sequence of character probabilities. The horizontal dimension collapses while the vertical dimension stretches to cover all possible characters. The output is processed with CTC, and the model is trained directly on full pages, requiring no extra annotations. However, OrigamiNet is limited to images of specific sizes, ensuring that the horizontal dimension can be collapsed correctly after convolutions.

Список литературы

[1] Mark Potanin, Denis Dimitrov, Alex Shonenkov, Vladimir Bataev, Denis Karachev, Maxim Novopoltsev – Digital Peter: Dataset, Competition and Handwriting Recognition Methods, arXiv:2103.09354 [cs.CV]

[2] Denis Coquenet, Clement Chatelain, Thierry Paquet – SPAN: a Simple Predict & Align Network for Handwritten Paragraph Recognition, ICDAR 2021

[3] Minghao Li and Tengchao Lv and Jingye Chen and Lei Cui and Yijuan Lu and Dinei Florencio and Cha Zhang and Zhoujun Li and Furu Wei – TrOCR: Transformer-based Optical Character Recognition with Pre-trained Models, CoRR 2022

[4] MohamedYousef, Tom E. Bishop – OrigamiNet: Weakly-Supervised, Segmentation-Free, One-Step, Full Page Text, Recognition by learning to unfold, CVPR 2020

[5] Denis Coquenet, Clement Chatelain, Thierry Paquet – End-to-end Handwritten Paragraph Text Recognition Using a Vertical Attention Network

[6] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah et al.– Language Models are Few-Shot Learners, NeurIPS 2020