УДК 004.932.75

СОЗДАНИЕ ОТКРЫТОГО НАБОРА ДАННЫХ ДЛЯ РАСПОЗНАВАНИЯ КИРИЛЛИЧЕСКИХ ИСТОРИЧЕСКИХ РУКОПИСЕЙ

Никишин А. П.[1]
Научный руководитель – канд. физико-математических наук, доцент Графеева Н. Г.[1]
[1]Университет ИТМО
apnikishin@itmo.ru

Работа выполнена в рамках темы НИОКТР №425041 «Разработка приложения по оцифровке и атрибуции кириллических рукописных текстов с применением методов компьютерного зрения и мультимодальных моделей.»

Введение

Автоматический анализ рукописных исторических документов является одной из ключевых задач сохранения культурного наследия и развития цифровых гуманитарных исследований. Тем не менее, отсутствие надежных ресурсов для анализа структуры документов и сегментации строк текста ограничивает разработку эффективных систем распознавания рукописного текста.

В рамках данной работы был создан и опубликован открытый набор данных CyrillicHist, предназначенный для анализа структуры документов в исторических кириллических рукописях XV–XVI вв. Текущая работа посвящена обобщению результатов CyrillicHist, а также описанию направлений развития второй версии корпуса, в которой уточняется разметка и добавляется текстовая аннотация на уровне строк.

Основная часть

Первая версия корпуса CyrillicHist включает 1000 полноразмерных изображений страниц из трёх палеографических традиций XV–XVI веков. Разметка выполнена в иерархической двухуровневой структуре: на уровне текстовых блоков и на уровне отдельных строк текста. Полигональная форма представления строк позволяет корректно учитывать кривизну строк, нестандартную геометрию, плотное письмо и маргиналии, характерные для исторических кириллических источников. Корпус был создан с использованием полуавтоматического пайплайна аннотирования. На первом этапе обучалась модель детекции текстовых блоков, на втором — модель сегментации строк внутри блоков. В качестве архитектурной основы использовались подходы семейства YOLO [1], предварительно обученные на открытых датасетах, таких как DIVA-HisDB [2] и HOME-Alcar [3]. Итоговый корпус опубликован в репозитории Zenodo [4] и предназначен для задач анализа макета документа и подготовки данных для HTR-систем.

В настоящее время ведётся работа над второй версией открытого набора данных. Основные направления развития включают: корректировку и унификацию существующей разметки с учётом выявленных систематических погрешностей и повышение согласованности полигональных масок. Также вторая версия предполагает связку строк с их текстовой транскрипцией, что позволит использовать корпус для дообучения моделей распознавания исторической кириллицы.

Выводы

Создание CyrillicHist v1.0 позволило восполнить дефицит открытых размеченных данных для исторических кириллических рукописей и заложило основу для стандартизированной оценки методов сегментации и анализа структуры страницы. Разрабатываемая версия направлена расширение прикладной ценности корпуса, что в перспективе создаёт основу для построения полноценных систем распознавания исторической кириллицы и дальнейшего развития цифровых исследований рукописного наследия.

Литература

1. J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “You Only Look Once: Unified, Real-Time Object Detection,” arXiv:1506.02640 (2016).
2. Simistira F. et al. DIVA-HisDB: A Precisely Annotated Large Dataset of Challenging Medieval Manuscripts. ICFHR, 2016.
3. Stutzmann D., Torres Aguilar S., Chaffenet P. HOME-Alcar: Aligned and Annotated Cartularies. Zenodo, 2021.
4. DOI: 10.5281/zenodo.18066471