УДК 004.912

ПРЕДОБРАБОТКА ЦИФРОВЫХ КОПИЙ ИСТОРИЧЕСКИХ ИСТОЧНИКОВ

Коробко С.С. (Университет ИТМО)
Научный руководитель - кандидат исторических наук, Пригодич Н.Д. (Университет ИТМО)

Введение. Оцифровка исторических документов, как и публикация изображений документов в вебе, нуждаются в механической предобработке сканированных копий: поворота, кадрирования, разделения на страницы и т.п. Так как документы могут быть объёмными, количество однотипной работы сильно возрастает[1]. Данная работа посвящена автоматизации процесса предобработки цифровых копий исторических источников и приведения их к единому формату методами искусственного интеллекта и машинного обучения. Целью работы является создание автоматизированной системы, приводящей переданные цифровые копии документов к единому формату. В связи с этим следует выделить ключевые задачи: обнаружение документа на цифровой копии, классификация цифровой копии в зависимости от ориентации документа, поворот, кадрирование, разделение на страницы. Полученная в результате исследования программа должна стать удобным и точным вспомогательным инструментом в задачах оцифровки исторических источников[2].

Основная часть. Будем считать, что есть только два вида фотографий исторических документов, в зависимости от ориентации: документы альбомной ориентации (развороты книг, тетрадей и т.п.) и документы книжной ориентации (листы А4, блокноты, отдельные страницы тетрадей). Смысл такого разделения в том, что фотографии первого типа должны быть разделены на страницы; фотографии второго типа — нет.

Таким образом, полученная программа будет состоять из трёх фаз: распознавание самого документа на фотографии, классификация документа в зависимости от ориентации и последующая обработка, включающая в себя поворот, кадрирование и разрезание на страницы.

Распознавание объектов на фотографии — это хорошо изученная задача в сфере Machine Learning. На сегодняшний день существует множество различных библиотек в области Computer Vision: OpenCV, fastai, PyTorchCV и др. С целью эффективного решения задачи была выбрана библиотека OpenCV, несомненными плюсами которой являются большое количество уже реализованных алгоритмов, производительность и поддержка не только языка Python, но других языков, таких как Java и C++, что будет полезно при включении данной программы в нагруженные системы.

Результатом первой фазы должен быть набор координат минимального по площади прямоугольника, покрывающего все точки документа на фотографии. Проблема в том, что получившийся прямоугольник по-прежнему может быть повернут не так, как нужно пользователю (на 90 или 180 градусов). Данную проблему решает вторая фаза: необходимый поворот можно вычислить, зная ориентацию документа.

Реализовать первую фазу можно двумя способами:

1) Разработать и обучить модель, которая бы распознавала документ на изображении.
2) Воспользоваться уже реализованным в библиотеке OpenCV преобразованием Хафа, предназначенным для идентификации геометрических объектов, в нашем случае прямоугольника, на растровых изображениях.

В рамках исследования был выбран второй вариант, из-за относительной простоты и производительности по сравнению с первым вариантом.

Оказалось, что и вторую фазу можно реализовать при помощи преобразования Хафа. Остановимся поподробнее на том, как именно это сделать. Будем считать, что первая фаза уже была выполнена, то есть вторая фаза выполняется в рамках прямоугольника, содержащего

документ. Тогда, если применить преобразование Хафа ещё раз, и взять среднее наклонов полученных прямых, то можно вычислить искомую ориентацию [3].

У данного способа реализации второй фазы есть один недостаток: невозможно отличить документ от его же перевернутого на 180 градусов. Более того, эта проблема существует и для человека – человек не может понять правильно ли ориентирована рукопись, при условии, что он ничего не знает про язык. Получается, что единственный способ устранить этот недостаток сводится к задаче оптического распознавания символов в рукописных текстах, которая на сегодняшний день не имеет достаточно точного решения.

Третья фаза реализуется при помощи вспомогательных инструментов OpenCV по работе с изображениями: “вырезаем” прямоугольник, полученный на первой фазе, поворачиваем и “разрезаем” на страницы, в зависимости от ориентации, полученной на второй фазе [4].

Выводы. В результате данной работы была продумана и разработана система, позволяющая упростить задачу оцифровки исторических источников путем автоматизации процесса приведения цифровых копий к единому человекочитаемому формату. У описанного решения есть ряд незначительных ограничений, которые могут быть сняты при дальнейшем развитии программы. Тем не менее полученная в результате исследования программа получилась в достаточной степени удобной и точной для применения, в качестве вспомогательного инструмента, в задачах оцифровки документов и дальнейшей их публикации в интернете.

Список использованных источников:

1. Муракас Р. Оцифровка исторических материалов исследований социальных наук как источник данных современных исследований // Коммуникация в социально-гуманитарном знании, экономике, образовании: Материалы V Международной научно-практической конференции. – Минск: Белорусский государственный университет, 2021. – С. 107-110.
2. Чурсина А. А. Российская практика цифровой обработки исторических источников: направления и результаты // Цифровое измерение новой социальной реальности: сборник научных студенческих статей. – Москва: Финансовый университет при Правительстве Российской Федерации, 2022. – С. 167-176.
3. Пример работы преобразования Хафа [Электронный ресурс]. – Режим доступа: https://drive.google.com/file/d/1VLuLm1bInSawyktSsDUym_niqGcAslZX/view?usp=sharing (дата обращения: 15.02.2023).
4. Пример работы программы [Электронный ресурс]. – Режим доступа: https://drive.google.com/file/d/1s2GU0HurcLEMqxiZFEiTWjtsGdez6qCX/view?usp=share_link (дата обращения: 15.02.2023).