Историческая информатика
Правильная ссылка на статью:
Пригодич Н.Д., Коробко С.С. — Применение программных методов для автоматизированной обработки источников личного происхождения // Историческая информатика. – 2023. – № 1. DOI: 10.7256/2585-7797.2023.1.40376 EDN: OJJZUU URL: https://nbpublish.com/library_read_article.php?id=40376

Применение программных методов для автоматизированной обработки источников личного происхождения

Пригодич Никита Дмитриевич

кандидат исторических наук

старший преподаватель, Национальный исследовательский университет ИТМО; старший научный сотрудник, Санкт-Петербургский государственный университет

197101, Россия, Санкт-Петербург, г. Санкт-Петербург, пр. Кронверкский, 49, литер А

ndprigodich@gmail.com

Коробко Семен Сергеевич

бакалавр, кафедра "Информатика и программирование", Национальный исследовательский университет ИТМО

197101, Россия, Санкт-Петербург, г. Санкт-Петербург, пр. Кронверкский, 49, литер А

semenkorobko2@gmail.com

Статья из рубрики "Новые методы и технологии обработки исторических источников"

DOI:

10.7256/2585-7797.2023.1.40376

EDN:

OJJZUU

Дата направления статьи в редакцию:

01-04-2023

Дата публикации:

08-04-2023

Аннотация: Предметом настоящего исследования являются программные методы автоматизированной предобработки исторических источников и разработка эффективного решения задач при работе с источниками личного происхождения. В рамках статьи проанализировано актуальное положение в области использования современных программных методов. Авторы демонстрируют основной круг аргументов, по которым такие исторические источники с технической точки зрения необходимо рассматривать отдельно. Проведен методологический разбор особенностей применения

оптического распознания символов на основе предобработанных данных. Особое внимание уделено преимуществам и ключевым параметрам эффективности конечного результата работы при использовании автоматизированной преодобработки текстов, в том числе при дальнейшем использовании OCR-методов. Научная новизна исследования заключается в предложении и подробном описании программного решения сложившейся проблемы на основе методов машинного обучения. Разработанная программа имеет три фазы работы с цифровыми копиями источников личного происхождения. В ее основе заложены использование библиотеки OpenCV и решения ряда задач с помощью преобразования Хафа. Опираясь на общий анализ исследования мы можем выделить основные преимущества автоматизированной предобработки сканированных документов: сокращение времени, повышение точности, борьба с искажениями и оптимизация процесса. Представленные результаты успешной апробации разработанного решения позволяют судить о возможных сферах ее эффективного применения.

Ключевые слова:

Источники личного происхождения, машинное обучение, искусственный интеллект, библиотека OpenCV, преобразование Хафа, предобработка, метод OCR, распознание, архивация, оцифровка

Введение

В современных реалиях значительно увеличились объемы сканируемых исторических источников в государственных, региональных и частных архивах, музеях, библиотеках и многих других организациях. Эта планомерная работа ценна как для исследователей, так и для самих учреждений. В то же время, постепенно, осуществляется переход к следующей стадии, технически более сложной, по механическому или автоматическому распознаванию текста для получения возможности новых форм анализа. В первую очередь это относится к описям архивных фондов, наиболее ценным источникам. Исследователи процесса оцифровки, приема и хранения электронных документов в отечественных архивах подробно описали историю и современные особенности данного процесса [1-4].

Примером сплошной обработки и переведения в электронный формат рукописных и машинописных текстов является работа центра изучения эго-документов «Прожито». Сотрудники центра и волонтеры занимаются пополнением базы данных источников личного происхождения, преимущественно на русском языке. Данная деятельность требует проведения ряда мероприятий по предобработке сканированного документа. Автоматизация процесса подготовительной обработки (кадрирование, разделение на страницы, очистка визуально шума, поворот) в значительной степени повышает скорость и эффективность дальнейшей работы по распознаванию текста как программными методами, так и вручную [5].

Определенный акцент, поставленный на автоматизированной обработке источников личного происхождения, может быть объяснен несколькими особенностями работы с такими документами. Во-первых, формат таких записей зачастую представляет собой тетрадные страницы, листы блокнотов, личных дневников различного размера, не подлежащего прямой унификации по сравнению с делопроизводственной документацией и архивными источниками. Во-вторых, большинство рассматриваемых текстов являются

рукописными, что накладывает дополнительные вводные при постановке задач при автоматизации. Наконец, источники личного происхождения часто обладают наложениями визуального шума, отделение которого требуют дополнительных временных затрат при обработке вручную.

Рис. 1. Пример источника личного происхождения.

Настоящее исследование посвящено анализу современного состояния положения подготовки сканированных копий исторических источников к процессу распознавания, а также поиску программного решения для автоматизации ряда задач и приведения их к единому формату методами искусственного интеллекта и машинного обучения. В этой связи следует обозначить наиболее значимые задачи для повышения эффективности при разработке программы: определить документ на сканированной копии, провести классификацию в зависимости от положения документа, автоматизировать поворот текста, разделить по страницам и кадрировать.

Методология и особенности использования

Автоматизированная обработка сканированных изображений перед распознанием текста – это необходимый этап в процессе преобразования бумажных документов в цифровой формат. Этот процесс, включающий в себя оптическое распознавание символов (OCR), позволяет преобразовать бумажные документы в текстовый формат для дальнейшей работы с ними в электронном виде [6]. Безусловно, потребность в качественном решении такой задачи существует не только в исторической науке. Например, банки, медицинские учреждения, правительственные учреждения и другие организации нуждаются в преобразовании бумажных документов в электронный формат для улучшения процессов документооборота и повышения эффективности работы в целом [7]. Однако при работе с историческими источниками есть ряд характерных особенностей, которые требуют адаптированного подхода.

Одним из главных преимуществ автоматизированной обработки сканированных изображений является уменьшение количества ошибок при распознавании символов. Этот процесс позволяет убрать лишние очертания и фон, повышая качество изображения и снижая вероятность ошибок в OCR, которые могут возникнуть при распознавании нечетких символов [6, с. 47]. Автоматизированная обработка сканированных изображений

также повышает скорость конвертации бумажных документов в электронный формат. Это возможно благодаря использованию автоматического распознавания области текста и удалению шума по всему изображению, что позволяет сократить время на ручную обработку документов. Кроме того, автоматизированная обработка сканированных изображений позволяет обрабатывать большие объемы документов. Это особенно актуально для работы с большими массивами исторических источников.

Рассматривая данную специфику следует выделить ряд характерных особенностей, которым, в итоге, содействует решение поставленных задач. Так, автоматическое распознавание текста на основе предобработанных образцов помогает исследователям более точно анализировать тексты, быстро получать доступ к широкому диапазону документов на основе конкретных слов и фраз [8, с. 214]. Более того, автоматическое распознавание текста обеспечивает уточненную интерпретацию смысла фраз и предложений в исторических документах. Оно способствует и сохранению культурного наследия, так как позволяет хранить электронную версию исторического документа, которая может быть отложена в базе данных или архиве, что гарантирует охрану документов для будущих поколений. Наконец, автоматическое распознавание текста повышает эффективность исторических исследований. В первую очередь за счет упрощения сбора и анализа материалов.

Программное решение

Перед началом разработки программного решения обозначенной задачи следует провести двухмерную типологизацию. Будем считать, что есть лишь два вида фотографий исторических документов, в зависимости от ориентации: документы альбомной ориентации (развороты книг, тетрадей, газет, журналов, горизонтально расположенные листы) и документы книжной ориентации (вертикально расположенные листы, блокноты, отдельные страницы тетрадей и др.). Такое разделение необходимо с точки зрения определения деления на отдельные страницы или его отсутствия. На основе установленных задач следует определить, что разрабатываемая программа должна состоять из трёх фаз: распознавание самого документа на сканированной копии, классификация документа в зависимости от ориентации и последующая обработка, включающая в себя поворот, кадрирование и разрезание на страницы.

Распознавание объектов на фотографии, вне контекста сканированных источников личного происхождения, – это хорошо изученная задача в сфере Machine Learning [9]. На сегодняшний день существует множество различных библиотек в области Computer Vision: fastai, PyTorchCV, OpenCV и некоторые другие. В данном случае, наиболее эффективном решению задачи отвечала библиотека OpenCV. Среди ее очевидных преимуществ необходимо обозначить высокую производительность, поддержку целого ряда языков программирования (в том числе Python, Java и C++), что будет полезно при включении данной программы в нагруженные системы, а также большой массив уже реализованных алгоритмов [10-11].

В ходе получения результатов первой фазы работы программы должен получиться набор координат минимального по площади прямоугольника, который покрывал бы все точки документа на фотографии. В этом контексте может возникнуть проблема, что получившийся прямоугольник будет повернут не так, как требует того пользователь, будь то 90 или 180 градусов. Такую проблему призвана решить вторая фаза, в рамках которой вычисление нужного поворота возможно при вводном знании ориентации документа.

Первая фаза может быть реализована двумя путями:

1) Разработка и обучение модели, которая бы распознавала документ на изображении.

2) Использование метода преобразования Хафа, задействованного внутри библиотеки OpenCV. Этот метод предназначен для идентификации геометрических объектов на растровых изображениях [12]. В нашем случае таким объектом является прямоугольник.

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

В ходе проведения предварительной апробации выяснилось, что вторая фаза программного решения также может быть реализована за счет метода преобразования Хафа [12, с. 829]. Необходимо более подробно остановиться 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 число наклонов полученных прямых, то можно вычислить искомую ориентацию.

При этом отдельно следует указать, что представленный способ реализации второй фазы имеет существенный недостаток: невозможно отличить документ от такого же, но перевернутого на 180 градусов. Однако при ближайшем рассмотрении становится понятно, что такая проблема актуальна и для человека, который не может понять правильно ли ориентирована рукопись, при условии, что он не обладает исходными данными о языке, на котором написан текст. В результате 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.

Наконец завершающая, третья фаза программного решения, может быть реализована при помощи вспомогательных инструментов библиотеки OpenCV по работе с изображениями [11, с. 18]. Для этого необходимо «вырезать» прямоугольник, полученный на первом этапе, повернуть его и «разрезать» на страницы. Последнее действие проводится в зависимости от той ориентации, которая была получена на второй фазе.

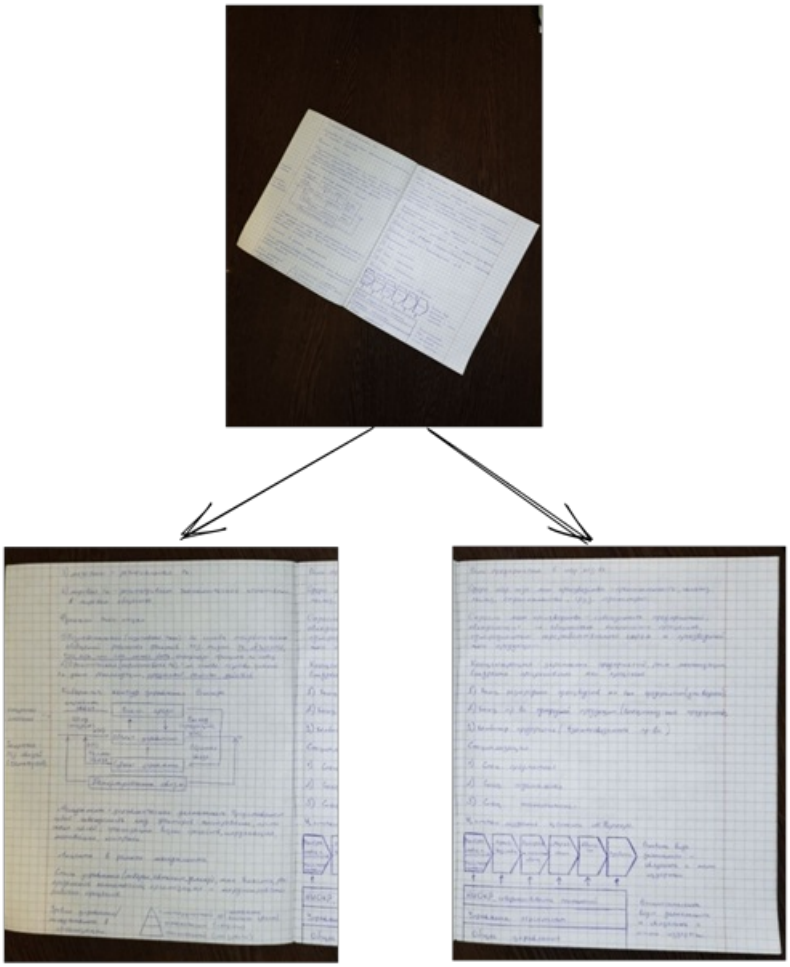

Рис. 2. Пример работы программы.

Выводы

В результате решения практической задачи было разработано программное решение, позволяющее упростить и ускорить процесс оцифровки сканированных источников личного происхождения. Опираясь на общий анализ исследования мы можем выделить основные преимущества автоматизированной предобработки сканированных документов:

1. Сокращение времени распознавания. Автоматическая предобработка изображения с текстом позволяет уменьшить количество ошибок при распознавании символов и значительно сократить время, необходимое для этого процесса.

2. Повышение точности распознавания. Автоматическая предобработка изображения помогает избавиться от шума и других артефактов, что улучшает качество распознавания символов.

3. Борьба с искажениями. Технология автоматической предобработки изображения позволяет бороться с различными искажениями, такими как искажения перспективы, искажения из-за сканирования, искажения в связи с плохим качеством печати.

4. Оптимизация процесса распознавания. Предварительная обработка изображения помогает оптимизировать процесс распознавания, сокращая затраты на вычислительные ресурсы и повышая скорость распознавания.

5. Повышение эффективности работы системы. Совокупность приведенных факторов приводит к общему росту эффективности результатов.

Анализ современного состояния автоматической предобработки цифровых копий источников личного происхождения показывает, что ни одно из существующих разработанных решений не позволяет в полной мере ответить на те запросы, которые стоят перед данной специфической областью. Определенные особенности, существующие в работе такого рода, вынуждают производить эти действия вручную, что приводит к замедлению общего процесса. Разработка конкретного решения на основе методов машинного обучения и искусственного интеллекта, при адресном использовании открытых библиотек, способствует достижению более качественного результата. Данное положение прошло этап успешной апробации на массивном объеме сканированных копий из фондов проекта оцифровки эго-документов «Прожито» при Европейском университете в Санкт-Петербурге и получило позитивные отклики специалистов.

В итоге, разработанное программное решение по автоматизированной обработке сканированных изображений перед распознанием текста – это необходимый этап для преобразования бумажных документов в электронный формат. Этот процесс не только повышает качество изображения, снижая вероятность ошибок OCR и улучшая скорость конвертации документов, но также упрощает работу организаций и позволяет работать с большими объемами бумажных документов. В то же время у разработанного решения есть некоторые ограничения, которые могут быть исправлены при дальнейшей оптимизации при использовании более совершенных методов обучаемого оптического распознавания символов.

Библиография

1. Мирошниченко М. А., Шевченко Ю. В., Охрименко Р. С. Сохранение исторического наследия государственных архивов путем оцифровки архивных документов // Вестник Академии знаний. 2020. № 37(2). С. 188-194. DOI 10.24411/2304-6139-2020-10163.
2. Куткин А. В., Назаров А. Н. Оцифровка документов в архивах Российской Федерации: анализ применяемого оборудования и программного обеспечения // Вестник ВНИИДАД. 2022. № 6. С. 41-52. DOI 10.55970/26191601_2022_6_41.
3. Решетько К. М., Халамей К. Н. Применение искусственного интелекта в банковском секторе // Потенциал российской экономики и инновационные пути его реализации: материалы всероссийской научно-практической конференции. 2021. Т. 2. С. 87-89.
4. Чурсина А. А. Российская практика цифровой обработки исторических источников: направления и результаты // Цифровое измерение новой социальной реальности: сборник научных студенческих статей. М.: Финансовый университет при Правительстве Российской Федерации, 2022. С. 167-176.
5. 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.
6. Ваксина И. Р., Канев А. И., Латыпова К. Н. Оптическое распознавание символов рукописных текстов и табличных данных // Тенденции развития науки и образования. 2022. № 86-1. С. 45-49. DOI 10.18411/trnio-06-2022-15.
7. Нестеров А. С. Анализ рынка современных информационных систем оптического распознавания символов (OCR) // Студенческий вестник. 2020. № 25-3(123). С. 82-85.
8. Шабанов А. В. Обработка изображений при создании цифровых копий рукописей с

угасающим текстом // Труды ГПНТБ СО РАН. 2013. № 5. С. 213-218.
9. Максимов В. Ю., Клышинский Э. С., Антонов Н. В. Проблема понимания в системах искусственного интеллекта // Новые информационные технологии в автоматизированных системах. 2016. № 19. С. 43-60.
10. Gevorkyan M. N., Demidova A. V., Demidova T. S., Sobolev A. A. Review and comparative analysis of machine learning libraries for machine learning // Discrete and Continuous Models and Applied Computational Science. 2019. Vol. 27, No. 4. P. 305-315. – DOI 10.22363/2658-4670-2019-27-4-305-315.
11. Бурмистров А. В., Ильичев В. Ю. Распознавание объектов на изображениях с использованием базовых средств языка Python и библиотеки opencv // Научное обозрение. Технические науки. 2021. № 5. С. 15-19.
12. Фаворская М. Н. Преобразование Хафа для задач распознавания // DSPA: Вопросы применения цифровой обработки сигналов. 2016. Т. 6, № 4. С. 826-830.

Результаты процедуры рецензирования статьи

В связи с политикой двойного слепого рецензирования личность рецензента не раскрывается.
Со списком рецензентов издательства можно ознакомиться здесь.

Рецензия на статью
Применение программных методов для автоматизированной обработки источников личного происхождения

Журнал: Историческая информатика

Предметом исследования в представленной статье является применение программных методов для автоматизированной обработки источников, хранящихся в различных фондах.

Статья посвящена проблеме обработки и переведения в электронный формат рукописных и машинописных текстов личного происхождения. Автор описывает особенности работы с подобными документами, основываясь на конкретных примерах. По сути, исследование посвящено анализу современного состояния положения подготовки сканированных копий исторических источников к процессу распознавания, что является серьезной проблемой, так как именно источники личного происхождения отличаются значительным количеством визуальных «шумов», затрудняющих процесс машинной обработки. Не менее важной задачей автор считает поиск программного решения для автоматизации процесса сканирования и обработки, а также выделяет основные категории задач, необходимых для решения данной проблемы.

Методология исследования является обоснованной и оптимальной для анализа проблемы. Использован комплекс методов, как общенаучных, так и узкопрофессиональных, позволивших предложить программное решение для упрощения процесса оцифровки документов личного происхождения.

Актуальность данного исследования обусловлена тем, что в современных реалиях значительно увеличились объемы сканируемых исторических источников в государственных, региональных и частных архивах, музеях, библиотеках и многих

других организациях. Эта планомерная работа ценна как для исследователей, так и для самих учреждений. Появление новых, более совершенных программных средств, требует постоянного изменения методик работы, в связи с чем представленная статья имеет также важное практическое значение.

Научная новизна работы безусловна. Автор разрабатывает и представляет результаты апробации нового программного решения, которое позволяет упростить и ускорить процесс оцифровки сканированных источников личного происхождения. Автор убедительно демонстрирует, что разработка конкретного решения на основе методов машинного обучения и искусственного интеллекта, при адресном использовании открытых библиотек, способствует достижению более качественного результата. Это подтверждается результатом обработки массива эго-документов «Прожито» при Европейском университете в Санкт-Петербурге. Автоматизированная предобработка документов, проведенная предложенным автором способом, способствует улучшению качества созданных цифровых копий.

Стиль работы соответствует высоким требованиям научного подхода к изложению результатов исследования. Его характеризуют логичность, строгая последовательность изложения, смысловая точность, информативная насыщенность, объективность. Структура изложения не вызывает нареканий и характеризуется взаимосвязанностью частей, логичностью переходов от одного раздела к другому. Предложенные иллюстрации информативны, важны для понимания выводов статьи, демонстрируют последовательность предлагаемых действий.

По содержанию данная статья является логически завершенным исследованием актуальной проблемы, осуществленным посредством применения комплекса научных методов. В статье содержится отсылка к предшествующим научным работам по данной теме, дается квалифицированная оценка полученных ранее результатов – как положительных, так и отрицательных. Автор корректно выявляет дефициты предшествующих исследований и предлагает свою методику решения проблемы автоматизированной обработки источников личного происхождения.
Предложенное автором программное решение обосновано и подтверждено практикой. Автор убедительно демонстрирует тот факт, что разработка конкретного решения на основе методов машинного обучения и искусственного интеллекта, при адресном использовании открытых библиотек, способствует достижению более качественного результата, подчеркивая в то же время, что у предложенного решения есть ограничения, которые могут быть ликвидированы при дальнейшей работе с использованием методов обучаемого оптического распознавания символов.
Библиография включает 12 источников, посвященных конкретной проблеме. Источники цитируются в тексте статьи.

Статья может быть полезна специалистам, так как выводы имеют несомненное практическое значение. Также статья, несомненно, вызовет интерес у широкого круга читателей.