МИНОБРНАУКИ РОССИИ
Федеральное государственное бюджетное образовательное учреждение
высшего образования
«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н.Г.ЧЕРНЫШЕВСКОГО»

Кафедра информатики и программирования

Создание системы распознавания рукописного текста с последующим исправлением ошибок

АВТОРЕФЕРАТ МАГИСТЕРСКОЙ РАБОТЫ

студентки 2 курса 273 группы
направления 02.04.03 Математическое обеспечение и администрирование информационных систем
факультета компьютерных наук и информационных технологий
Елисеевой Елизаветы Дмитриевны

Научный руководитель
к. ф.-м. н. , доцент _____ М.В. Огнева
(подпись, дата)

Заведующий кафедрой
к. ф.-м. н., доцент _____ М. В. Огнева
(подпись, дата)

Саратов 2021

ВВЕДЕНИЕ

С развитием технологий, увеличением объема потребляемой информации и ускорением её обмена, текстовая информация на электронных носителях становится всё более востребованной. Сканирование и последующее сохранение в памяти компьютера текста является давно решенной задачей, однако полученный текст хранится в виде изображения, что усложняет, а зачастую и вовсе исключает, возможность работы с текстом: исправление, добавление, поиск и т.п.

Для решения вышеперечисленных задач существуют OCR-системы: они производят анализ и распознавание текста на изображении. Распознавание текста – это процесс перевода изображений рукописного, машинописного или печатного текста в текстовые данные. В настоящее время существуют высокоточные системы для распознавания машинописных и рукопечатных текстов, например, ABBYY FineReader, Tesseract, GOCR. Гораздо более сложной и нерешенной задачей является распознавание рукописных текстов, HWR или HTR. Существуют два класса задач HWR:

- онлайн-распознавание – распознавание текста ведётся параллельно с вводом текста;
- оффлайн-распознавание – распознавание текста ведётся на уже синтезированном изображении.

Онлайн-распознавание часто используется на планшетных ПК и позволяет в режиме реального времени отслеживать различные параметры, необходимые для более точного распознавания, в том числе сам процесс начертания отдельного символа, что позволяет легче сегментировать рукописный текст. Некоторые системы предоставляют обработку и исправление орфографических ошибок «на лету».

Оффлайн-распознавание является более сложной задачей с весомым списком проблем, нерешенных до сих пор:

- орфографические ошибки в тексте;

- затруднение процесса сегментирования в связи с пересечением элементов текста, наложением частей текста друг на друга, например, близко расположенные строки;
- помарки, кляксы, исправления и артефакты, возникающие при сканировании;
- высокая вариативность начертания символов – размер, наклон, набор составных частей и связи между ними.

Для оффлайн распознавания рукописного текста существуют такие системы как ABBYY FineReader, он хорошо распознает рукописный текст, но является платным продуктом, точность распознавания неизвестна, Tesseract имеет хорошую точность распознавания английского текста, но для русского языка точность составляет лишь 22%, GOCR имеет точность 13%, кроме того его производительночть очень низкая, Tenserflow обучен на данный момент только распознавания цифр.

Целью работы является разработка приложения, производящего обработку изображения сосканированного рукописного текста.

Задачи работы:

- изучение методов предобработки изображений;
- изучение методов сегментирования текста и распознавания символов;
- анализ текущих инструментов исправления ошибок в тексте;
- создание программного продукта, осуществляющего предобработку изображения сосканированного рукописного текста: избавление от артефактов, возникающих при сканировании; поворот строк текста до горизонтального положения; сегментирование, а также распознавание символов текста и последующее исправление ошибок.

Методологические основы. Теоретические основы распознавания рукописного текста представлены в работах: Шапиро, Л., Стокман [1], Дж.,

Plamondon, R´ejean, Srihari, Sargur N [2], Pl¨otz, Thomas and Fink, Gernot A. Markov [3], Li, Yi and Zheng, Yefeng and Doermann, David and Jaeger, Stefan [4], сегментирвоание текста представлено в работах: Bar-Yosef, Itay and Hagbi, Nate and Kedem, Klara and Dinstein, Itshak [5], Likforman-Sulem, Laurence and Hanimyan, Anahid and Faure, Claudie [6], Fortune S[7], Wang Z[8].

Практическая значимость магистерской работы. В ходе выполнения практической части магистерской работы были реализованы алгоритмы предварительной обработки изображений, нормализации ркописного текста на изображении, сегментирвоания теста на строки, слова и буквы, а также обучение инстумента Tesseract OCR после чего значительно улучшилось качество распознавания рукописного текста. Так как при распознавании букв могут быть ошибки, к результирующему тексту был применен алгоритм исправления ошибок в тексте.

Для текстов средней сложности независимое применение данных библиотек Tesseract OCR и SymSpell не приводит к эффективному распознаванию рукописных текстов на русском языке (не более 40%). Вместе с тем, совместное применение предложенных в работе методов сегментации на основе частотного анализа и диаграмм Вороного и стандартных средств библиотек Tesseract OCR и SymSpell после «тонкой настройки» и обучения на относительно небольшом множестве рукописных текстов позволяет в ряде случаев провести распознавание рукописных текстов средней сложности на русском языке с эффективностью 100%. Результаты могут быть использованы IT-компаниями при разработке программных модулей для распознавания рукописных текстов на русском языке, что определяет практическую значимость работы.

Структура и объём работы. Магистерская работа состоит из введения, 2 разделов, заключения, списка использованных источников и 1 приложения. Общий объем работы – 92 страницы, из них 66 страниц – основное содержание, включая 44 рисунка и 3 таблиц, приложение, список использованных источников информации – 29 наименований.

КРАТКОЕ СОДЕРЖАНИЕ РАБОТЫ

Первый раздел «Распознавание рукописного текста» посвящен обзору существующих методов предобработки изображений, сегментации

рукописного текста на строки, слова и символы, а также распознавания рукописного текста и исправления ошибок в тексте.

Основные понятия первого раздела:

OCR – Optical Character Recognition – оптическое распознавание символов, механический или электронный перевод изображений рукописного, машинописного или печатного текста в текстовые данные, использующиеся для представления символов в компьютере.

HWR (HTR) – Handwriting Recongnition (Handwritten Text Recognition) – оптическое распознавание символов, механический или электронный перевод изображений рукописного текста в текстовые данные, использующиеся для представления символов в компьютере.

Диаграмма Вороного конечного множества точек S – разбиение плоскости, при котором каждая область разбиения образует множество точек, более близких к одному из элементов множества S, чем к любому другому элементу множества.

Распознавание рукописного текста можно разделить на этапы. Первым этапом является **предобработка изображения** (preprocessing). На данном шаге происходит первоначальное преобразование изображения для повышения его качества и удобства проведения последующих этапов. Производятся изменение палитры изображения, выделение заднего фона и переднего слоя (текста, картинок и т.п.), приведение строк, слов и букв рукописного текста к близкому горизонтальному расположению. Для предобработи изображений были описаны методы выделения цветовой палитры изображения, определения цвета фона, изолирования переднего плана от фона, выделения основных цветов метом *k*-средних и нормализации изображений.

Следующий этап, **сегментация** (segmentation), выделяет текст на изображении и производит его разделение на составные части: строки, слова, буквы. Есть два подхода выделения текста на изображении: сверху-вниз, когда изначально выделяются абзацы, далее строки, после чего слова и,

наконец, отдельные буквы; и снизу-вверх, когда изначально выделяют буквы, после собирая их в слова, которые, в свою очередь, собираются в строки или предложения, из которых составляют абзацы. В данной работе описаны ращличные способы сегментации текста: методом частотного анализа, методом дигаграм Вороного, и интегральное представление изображения.

Далее, на следующем шаге, производится **извлечение признаков** (feature extraction). Происходит поиск и анализ особенностей сегментов, выделенных на предыдущем этапе.

На четвертом этапе, **классификации** (classification), по найденным и проанализированным признакам принимается рашание о том, к какому известному классу отнести рассматриваемый сегмент. Зачастую на данном шаге используются нейронные сети. Для третьего и четвертог этапов был описан инструмент Tesseract OC.

Пятый, и последний, этап, **постобработка** (postprocessing), производит построение итогового текста по результатам классификации сегментов. На этом этапе в работе был описан алгоритм SymSpell для исправления ошибок в распознанных словах.

После изучения теоретических материалов, были сделаны выводы, что предобработка изображения – важный этап распознавания текста, от котрого зависит дальнейшая успешное сегментирование и распознавание рукописного текста. Для сегментации строк и слов был выбран метод частотного анализа, для сегментации слов была выбрана диаграмма Вороного. Для распоназавания текста был выбран инструмент Tesseract OCR. Для исправления ошибок был выбран алгоритм SymSpell.

Второй раздел «Программная реализация» посвящен реализации предобработки изображений, сегментации текста методом частотного анлиза на предложения и слова и методом диаграмм Вороного на буквы. На следующем этапе работе была обучена нейросеть в инструменте Tesseract OC. А также после распознавания слов, был применен алгоритм SymSpell для исправления ошибок в итоговом тексте. Были приведены подробные

примеры тестирования работы алгоритма, а также отдельные примеры тестирования предобработки и распознавания рукописного текста. В результате чего были сделаны выводы, представленные ниже:

Результаты работы продукта показали гибкость программы: возможность достижения максимально эффективной и точной обработки изображения с помощью изменения опций. Однако результаты с фотографией текста, снятой со включенной вспышкой, оставляют желать лучшего: здесь необходим хитрый алгоритм, отделяющий фон и передний план не по цветовым характеристикам. Алгоритм сегментации, созданный в данной работе, лучше своих аналогов, особенно при использовании полученных сегментов для последующего распознавания символов.

Точность распознавания рукописных символов на несегментированных данных у Tesseract OCR составляет 22%. После обучения программа была протестирована на 100, обучающих текстах и 100 случайных текстах. Точность распознавания – отношение правильно распознанных символов к общему количеству символов в текстах. Для изображений отдельных символов, сегментированных одним из способов, описанных в данной работе, точность равна 41%. При проверке на обучающих данных точность достигла отметки 64%. Также были протестированы тексты разного уровня сложности, в некоторых случаях не удалось распознать ничего, такие тексты тяжело прочитать даже человеку, в сложных текстах точность распозавания – 30%, в текстах средней сложности алгоритм показал себя хорошо, точность распознавания 73%, в легких – 87%. С прменением алгоритма SymSpell для текстов средней и легкой сложности итоговый текст мог достигнуть 100% точности. После чего было сделано предположение, что успешное распознавание сильно зависит от почерка. Поэтому эффективность созданного в данной работе алгоритма сегментации выше представленного в инструменте Tesseract.

Следует заметить, что разнообразие почерков остается главной проблемой распознавания рукописных символов. Какая бы огромная

обучающая выборка ни была дана на вход искусственной нейронной сети, она будет недостаточна для точного определения символов специфического почерка. На данный момент даже самые продвинутые искусственные нейронные сети не справляются с данной задачей.

ЗАКЛЮЧЕНИЕ

В ходе работы были проанализированы особенности рукописного текста, выделены проблемы, возникающие по ряду причин на каждом этапе создания и обработки изображения. При сканировании зачастую появляются артефакты: пятна, засвеченности при неплотно закрытой крышке сканера, просвечивающаяся задняя сторона листа. При фотографировании – чаще всего засвеченность фрагментов изображений из-за применения вспышки, неравномерного падения света на лист в момент съема, приводящие зачастую к тому, что в разных фрагментах фон и передний план (текст) становятся неразличимы. Также встречаются такие проблемы, как отклонение строк текста от горизонтального положения, нахлест символов друг на друга, затрудняющие процесс сегментации. Кроме того, люди имеют разный почерк, в итоге существуют сотни вариантов написания одного символа, это существенно осложняет задачу распознавания рукописного текста, на данный момент не существует универсального решения этой задачи. Исправление ошибок после распознавния текста неройнной сетью также является нетривиальной задачей, так как для существенного улучшения итого результата, необходимо иметь большой словарь с разными формами слов, а также анализировать частотность употребления слов, семантические и грамматические особенности языка.

Была создана программа, справляющаяся с большинством вышеуказанных проблем, избавляя исходное изображение от большинства артефактов, проводящая нормализацию и сегментацию текста, а также распознавание и исправление ошибок в нем.

Предобработка изображений включает в себя последовательно выполняемые алгоритмы выделения цветовой палитры, уменьшения размерность цветового канала, приближение цветов методом *k*-средних, усечение цветовой палитры и повышения яркости. Дополнительной опцией здесь является преобразование найденного фона к белому цвету.

Нормализация текста была произведена с помощью алгоритма достижения максимально возможного количества наиболее белых горизонтальных линий на изображении. Используемый в работе метод успешно приводит строки текста в горизонтальное положение, предлагается его дальнейшее использование при нормализации отдельных элементов текста (отдельных строк и слов).

В качестве одного из алгоритмов сегментации текста был выбран метод диаграмм Вороного. Программа строит диаграмму алгоритмом Форчуна, однако на вход ей требуются центры масс символов текста. Пример работы сегментации при заданных центрах масс был приведен на рисунке 18.

Вторым методом сегментации текста стал алгоритм троекратно повторяющегося частотного анализа, описанного в данной работе. При каждом его применении последовательно выделяются строки, слова и символы.

Алгоритмом распознавания символов выбран Tesseract-OCR. Его собственный метод сегментации оказался на порядок менее точным, чем представленные в данной работе. Однако даже с использованием лучших алгоритмов сегментаций точность распознавания – 41%.

Сочетание инструмента SymSpell, методов N-грамм и сглаживания позволило с высокой точностью находить и исправлять ошибки в тексте. При этом распознанный алгоритмом Tesseract текст содержит большое количество ошибок и не может быть полностью исправлен инструментом SymSpell.

Результаты работы продукта показали гибкость программы: возможность достижения максимально эффективной и точной обработки

изображения с помощью изменения опций. Однако результаты с фотографией текста, снятой со включенной вспышкой, оставляют желать лучшего: здесь необходим хитрый алгоритм, отделяющий фон и передний план не по цветовым характеристикам. Алгоритм сегментации, созданный в данной работе, лучше своих аналогов, особенно при использовании полученных сегментов для последующего распознавания символов. Следует заметить, что разнообразие почерков остается главной проблемой распознавания рукописных символов. Какая бы огромная обучающая выборка ни была дана на вход искусственной нейронной сети, она будет недостаточна для точного определения символов специфического почерка. На данный момент даже самые продвинутые искусственные нейронные сети не справляются с данной задачей.

Отдельный части магистерской работы были представлены на конференции: «Сегментация текста методом диаграмм Вороного», студенческая научная конференция факультета КниИТ, 23 апреля 2021г.

Основные источники информации:

1. Kim, Gyeonghwan and Govindaraju, Venu and Srihari, Sargur N. An architecture for handwritten text recognition systems. International Journal on Document Analysis and Recognition, 2, 1, (37–44), 1999.
2. Oztop, E and M¨ulayim, Adem Yasar and Atalay, Volkan and Yarman-Vural, Fatos. Repulsive attractive network for baseline extraction on document images. Signal Processing, 75, 1, (1–10), 1999.
3. Plamondon, R´ejean and Srihari, Sargur N. Online and off-line handwriting recognition: a comprehensive survey. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 22, 1, (63–84), 2000.
4. Vinciarelli, Alessandro and Luettin, Juergen. A new normalization technique for cursive handwritten words. Pattern Recognition Letters, 22, 9, (1043–1050), 2001.
5. Bar-Yosef, Itay and Hagbi, Nate and Kedem, Klara and Dinstein, Itshak. Line segmentation for degraded handwritten historical documents.

Document Analysis and Recognition, 2009. ICDAR’09. 10th International Conference on, (1161–1165), 2009.

6. Likforman-Sulem, Laurence and Hanimyan, Anahid and Faure, Claudie. A Hough based algorithm for extracting text lines in handwritten documents. Document Analysis and Recognition, 1995., Proceedings of the Third International Conference on, 2, (774–777), 1995
7. Fortune S. A sweepline algorithm for Voronoi diagrams / S. Fortune // Proceedings of the second annual symposium on Computational geometry. – 1986. – P. 313 – 322.
8. Wang Z. Word Extraction Using Area Voronoi Diagram. / Z. Wang, Y. Lu, C. Lim // CVPRW ‘03. – 2003. – P. 31 – 36.