DOI 10.37882/2223-2966.2023.09.10

АЛГОРИТМ РАСПОЗНАВАНИЯ И КЛАСТЕРИЗАЦИЯ РУКОПИСНЫХ ТЕКСТОВ

ALGORITHM FOR RECOGNIZING AND CLUSTERING HANDWRITTEN TEXTS

A. Kasymov
Yu. Maximov

Summary. The paper presents recognition capabilities and clustering of handwritten texts with help modern technology. Separately, the paper pays attention to the comparison of handwritten text recognition from computer.

Keywords: handwritten texts, recognition, clustering, classification, data entry.

Касымов Алексей Алексеевич
Аспирант, Воронежский государственный технический университет
kasimlele@live.ru
Максимов Юрий Максимович
Аспирант, Воронежский государственный технический университет
yuramaximo@mail.ru

Аннотация. В работе представлены возможности распознавания и кластеризация рукописных текстов посредством современных технологий. Отдельно в работе уделено внимание сравнению распознавания рукописных тестов от компьютерных.

Ключевые слова: рукописные тексты, распознавание, кластеризация, классификация, ввод данных.

Введение

Необходимо подчеркнуть, что адекватное усвоение грамматических структур дозволяет оценить временные и видовые аспекты лексических единиц. В определенных случаях, морфологическая структура единичных рукописных символов может обеспечить ограниченное количество информации для безошибочной идентификации примерно 98 % общего объема рукописных текстов. Механизм автоматизированного распознавания образцов и его реализация в контексте оптической системы распознавания символов представляют собой важнейшие компоненты эффективных парадигм искусственного интеллекта.

В рамках данной интерпретации понимается под распознаванием текста автоматический процесс выявления изображений письменных символов, будь то печатные или рукописные, (примером может служить ввод данных сканером в компьютер) с применением специализированного программного обеспечения, и их последующее преобразование в форматы, пригодные для обработки текстовыми редакторами.

Термин «OCR» может быть расшифрован как «оптическое распознавание символов» [1]. В данном контексте, термин OCR обозначает процесс оптического выявления и аппаратное оборудование, предназначенное для автоматизированного считывания текстовых данных (см. рисунок 1). В настоящее время, в индустриальной

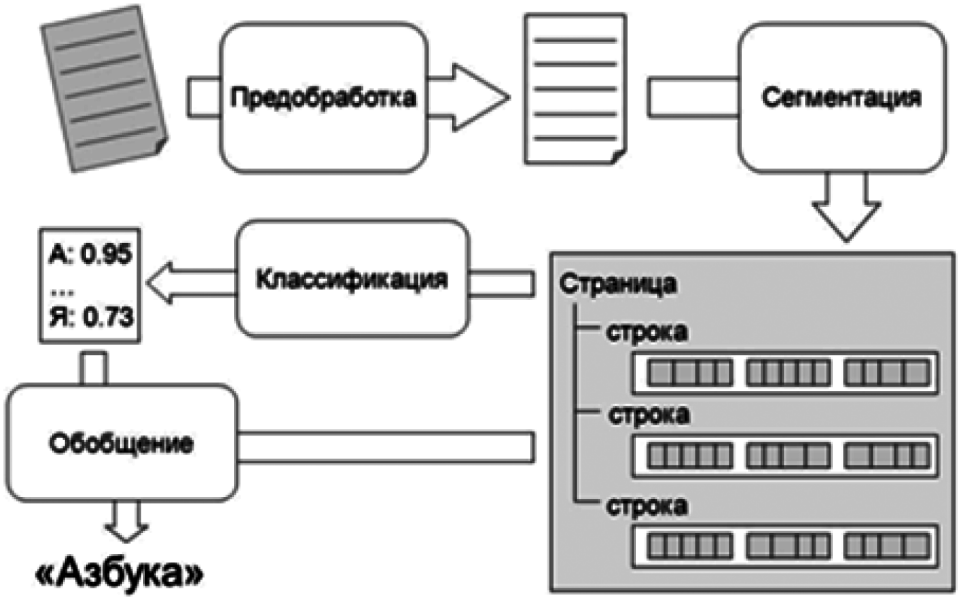

Рис. 1. Структура системы распознавания текста

среде подобное оборудование способно обрабатывать до 100 000 документов в сутки. Под индустриальным применением следует понимать последовательное внедрение в систему дополнительных документов. Этот процесс обусловлен потребностями разнообразных операций обработки (например, обработка переписных листов, налоговых деклараций и т.д.).

Одним из экземпляров подобных систем представляется когнитивный подход, разработанный с целью управления масштабированным вводом документов определенной структуры (подобных, например, налоговым декларациям, формам бухгалтерского учета, документам о платежах и др.).

Система анализа текста на основе комплекса программных компонентов предоставляет возможность параллельного ввода пакетов документов с соответствующими стандартизированными формами. Возможность взаимодействия между модулями, реализованными в локальной сетевой среде, способствует интеграции данных этапов в единую обработку, обеспечивая производительность, способную обработать более 10 тысяч страниц в течение суток [2].

Актуальность

В современном цифровом мире важность автоматизированного распознавания и обработки рукописных текстов приобретает особое значение. Данная статья рассматривает актуальные проблемы, связанные с расшифровкой разнообразных форм рукописного контента, начиная от отдельных символов и до целых текстов.

Статья глубоко исследует процесс распознавания и кластеризации рукописных текстов, разъясняя взаимодействие методов распознавания на основе образов и анализа структуры. В период, когда искусственный интеллект и технологии машинного обучения продолжают развиваться, алгоритмы, представленные в статье, имеют значительное значение для повышения точности и эффективности систем распознавания рукописного текста.

Через изучение сложностей контекстного распознавания, структурного анализа и кластеризации образов статья вносит вклад в развитие интеллектуальных алгоритмов, способных распознавать разнообразные формы рукописных вводов, независимо от изменений шрифтов, размеров или форматов. Более того, исследование методов устранения шума и стратегий коррекции ошибок подчёркивает практическую применимость статьи в реальных условиях.

Релевантность статьи распространяется на такие области, как цифровая документализация, поиск информации и обработка языка, где точное и эффективное распознавание рукописных текстов является ключевым фактором. В условиях перехода отраслей к цифровой трансформации алгоритмы и методологии, обсуждаемые в статье, предоставляют ценные научные предпосылки для улучшения качества систем автоматизированного распознавания, тем самым оптимизируя процессы управления информацией.

Метод. Процесс внесения документов в стандартизированную форму включает две стадии: начальную и основную. В начальной фазе происходит формирование шаблона для целевого документа. Указанный шаблон определяет свойства документа, включая состав данных, его структуру, размер страницы, а также местоположение и размер соответствующих полей, тип данных, формат представления, доступный спектр значений и иные детали. Средство Cognitive Form Designer обеспечивает средства для создания и модификации подобных шаблонов. Данный процесс разбивается на шесть последовательных этапов.

Первый этап охватывает процесс цифровой трансформации печатных документов через сканирование и их последующий перевод в электронный формат. Комплексные модули Cognitive Form Designer, включая модуль пакетного сканирования и модуль автоматического сканирования каждой страницы, руководствуют данным процессом.

Второй этап охватывает классификацию и отбор. Подразумевается, что документ может содержать множество страниц, связанных с различными стандартизированными шаблонами. На данном этапе изображения страниц группируются в коллекции, соответствующие конкретным документам. Проблема автоматически решается через модуль обработки когнитивных форм, который включает следующие шаги:

- предварительная обработка графического представления и выделение ключевых графических элементов (границы полей, текстовые строки и пр.).
- выбор наиболее подходящего стандартизированного документа.
- выявление и идентификация данных, относящихся к проверке целостности файла.
- проверка целостности основывается на соответствии последовательности типов страниц заранее установленной структуре, описанной в стандартизированном шаблоне.

Третий этап охватывает пересмотр результатов классификации. Этот этап выполняется оператором, и на него передаются недостаточно полные документы. Приоритетным является выявление и устранение причин возникших проблем [3].

Четвертый этап заключается в идентификации основной информации. Данный процесс осуществляется через модуль обработки когнитивных форм. Графическое представление страницы и уникальное значение элемента данных записываются в системную базу данных. Для улучшения точности распознавания производится логическая проверка и анализ результатов.

Пятый этап включает проверку результатов идентификации. Документы, содержащие нераспознанные или однозначно распознанные элементы данных, направляются оператору. Для проверки и коррекции результатов применяется модуль Cognitive Form Editor.

Шестой этап заключается в передаче утвержденных документов для дальнейшей обработки во внешних приложениях.

Посредством представленных выше этапов обосновывается вывод о наличии недостатков на каждой из стадий в рамках системных решений. К примеру, инструмент FineReader продемонстрировал выдающиеся результаты при распознавании рукописного текста, включающего индивидуально выделенные символы, но выявил значительное количество ошибок в случае текста слиянием написанных слов.

Также когнитивный механизм распознавания текста обладает собственными недостатками. Первостепенно, этот подход ограничен в своей применимости к неструктурированному тексту, так как в первую очередь рассчитан на обработку текстов, оформленных в форме стандартизированных документов, специализированных форм и подобных артефактов.

С учетом вышеуказанных характеристик предметной области, следующие аспекты предстают ключевыми в рамках рассмотрения системы распознавания текста:

— диспаратные стили и размеры символов;
— искажения в символьных изображениях, представляющие собой различные артефакты (например, скачки масштаба символа при изменении размера, объединение смежных символов и так далее);
— искажения, возникающие в процессе сканирования;
— вариативные категории символов, требующие распознавания при наличии дополнительной контекстуальной информации.

Процедура автоматизированного извлечения информации из печатных и рукописных текстов представляет собой частный случай автоматизированного восприятия текста и графических изображений. Обширный корпус исследований свидетельствует о том, что наличие интеллектуального распознавания, то есть понимания и смысла, оказывается неотъемлемым условием для выполнения данной задачи [4].

Тем не менее, на настоящем этапе развития технических систем для распознавания текста проблематика упрощается и сводится к задаче классификации на основе характеристик базовых элементов. Подобное распознавание представляет собой процесс выбора пороговой границы, опираясь на стройный математический аппарат — разделительную гиперплоскость [2].

Наиболее перспективное решение в области распознавания текста включает в себя методику человеческого распознавания. Человеческое восприятие образов представляет собой процесс многостадийного анализа.

Практически все системы распознавания текста базируются на трех важных принципах.

Принцип интегральности изображения: всегда существует взаимосвязь между ключевыми компонентами изучаемых объектов. Локальные действия, проводимые над отдельными частями изображения, находят объяснение лишь в контексте интерпретации полного фрагмента и всего изображения в целом.

Принцип цели: процесс идентификации представляет собой направленную последовательность проверок гипотез (формирование и проверка ожиданий по отношению к объектам).

Принцип адаптации: система распознавания должна обладать способностью к саморегулированию и обучению [3].

Графическое представление символа после сканирования принимает форму точечной матрицы, доступной для последующего пошагового редактирования. Иллюстрация 2 демонстрирует типичный образец буквы «л» или «п».

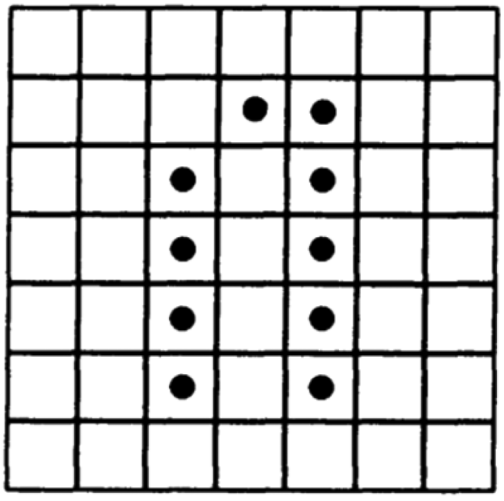

Рис. 2. Пример шейпа

В настоящем контексте, используется информация, извлеченная из результата распознавания смежных текстовых компонентов, для целей выявления аномальных файлов в формате шейп-файла. В наиболее элементарном варианте, контекстом служит само лексическое единство.

Тем не менее, предоставленной информации о конкретном слове нередко недостаточно для осуществления принятия взвешенного решения. В этом случае область анализа контекста расширяется до включения предложения или даже нескольких предложений (частей текста). Внедрение соответствующих механизмов подразумевает включение в рассмотрение трудности интерпретации естественного языка.

Присутствует три видов классификаторов:
— шаблонные.
— признаковые.
— структурные.

В шаблонном классификаторе происходит определение наиболее подходящих шаблонов из имеющейся базы (см. Рисунок 3). Самым элементарным критерием служит минимальное количество различающих точек между узором и текстовой или изображенной информацией.

Рис. 3. Шаблонный классификатор

В процессе анализа осуществляется лишь сопоставление посредством набора числовых характеристик, либо признаков, вычисленных изображением. Таким образом, субъектом распознавания становится не конкретный символ в изоляции, а агрегированный комплекс его свойств, означающих данные, полученные из обследуемого символа. Следовательно, данное установление неизбежно сопровождается утратой некоторой информационной составляющей на более поздних этапах [5].

Классификатор структурного состояния трансформирует шейп-файл символа в его топологическое изображение, содержащее детали об организации структурных компонентов символа. При данном методе обеспечивается инвариантность относительно шрифтового размера и типа. Тем не менее, следует учесть, что данный метод сталкивается с затруднениями в распознавании символов с дефектами, а также может обусловливать замедление процесса обработки.

Современные системы оптического распознавания символов, как правило, воплощают все три видов классификаторов, среди которых особое внимание уделяется применению структурных классификаторов в связи с их актуальной эффективностью. Для обеспечения более высокой скорости работы и улучшения качества распознавания, системы используют комбинированный подход с применением сетей и классификаторов признаков [6].

Автоматические системы включают в свой состав специфическую модель — структурно-пятенный эталон. Данный эталон состоит из набора пятен, снабженных попарными зависимостями, что представлено на Рисунке 4 [7].

Рис. 4. Схема расположения структурных пятен

Принципиальный алгоритм функционирует в основе симбиоза между приемами образов, воплощенными через паттернинг, и строением структурных образцов. В ходе изучения выборочных образов выделяются узловые акценты объекта, характеризуемые как «пятна». Сущностные примеры пятен включают:
— узлы, агрегирующие пересечения нескольких линий;
— точечные дефициты в сегментах линий;
— точки пересечения линий.

В процессе выбора «пятен» проводится детерминированное выявление связей между сегментами и арками. Следовательно, финальным описанием становится графическая схема, выступающая как цель поиска в рамках «Структурного точечного образца» [7]. На этапе поиска осуществляется выявление соответствия между точками выборки и эталоном, с последующим определением необходимой степени деформации маркера, необходимой для приведения исследуемого объекта к эталону, подвергаемому сравнению. Уменьшение требуемой деформации коррелируется с повышением вероятности точной идентификации символа [2].

В целях усовершенствования эффективности процесса распознавания применяются различные методы предварительной обработки изображений с включенным текстом, включая, например, методы шумоподавления. Источники шума визуализированного изображения могут охватывать:

— аналоговый шум:
— пыль;
— царапины;
— цифровой шум:
— матрица теплового шума;
— шум в процессе передачи;
— квантование шума аналого-цифровым преобразователем.

Для подавления пространственного шума наблюдается применение в цифровой обработке изображений. В рамках данного контекста выделяют следующие методы:
— адаптивная фильтрация — применение среднего значения к соседним пикселям;
— фильтрация по окрестностям;
— математическая морфология;
— операция размытия по Гауссу;
— метод главных компонент.

По завершении этапа распознавания возможны дополнительные корректировки с целью повышения точности распознавания спорных символов (подразумевающих наличие нескольких кандидатов с примерно одинаковыми степенями соответствия многорежимному шаблону). Примерами таких корректирующих методов могут быть:
— анализ лингвистических особенностей сочетания символов;
— использование языковых словарей;
— грамматический анализ;
— иные стратегии.

Неустойчивость автоматизированного визуального восприятия на текущий момент еще не преодолена в полной мере сравнительно с человеческим способом восприятия текста. Основной коренной сущностью данной ситуации является выявление трудностей в построении компьютерных моделей предметной области, способных отразить полноту и семантику человеческого понимания [8].

Заключение

Анализируя существующие методы оптического распознавания текста, можно заключить, что метод «точечной матрицы структуры» представляет собой наилучший выбор в силу своей способности комбинировать преимущества различных методик, обеспечивая тем самым достаточную гибкость в рамках процесса распознавания рукописного ввода.

ЛИТЕРАТУРА

1. Шамис А.Л. Принципы интеллектуализации автоматического распознавания / А.Л. Шамис. — К: 2019 — 312 с.
2. Шлезингер М. Десять лекций по статистическому и структурному распознаванию: лекции / М. Шлезингер, В. Главач. — М.: Наука, 2020 — 112 с.
3. Шлезингер М. Структурное распознавание / М. Шлезингер, В. Главач. — Киев: Наука, 2019 — 300 с.
4. Гаврилов Г.П. Логический подход к искусственному интеллекту / Г.П. Гаврилов. — М.: Мир, 2019 — 256 с.
5. Wilkinsonet R.A. The First Census Optical System / R.A. Wilkinsonet. — Gaithersburg: NIST, 2020 — 242 с.
6. Электронный ресурс по искусственному интеллекту. — URL: Research Library.
7. Электронный ресурс по нейронным сетям. — URL: StatSoft
8. Абраменко А. Принципы распознавания / А. Абраменко. — К: Кнорус, 2020 — 123 с.

Журнал «Современная наука: актуальные проблемы теории и практики»