УДК 004.89

Халилов Дурбек Аминович к.ф-м.н., профессор кафедры информационных технологий, Ферганский филиал, Ташкенский университет информационных технологий им. Мухаммада аль-Хорезми,
Аттокуров Урмат Тологонович, к.т.н., профессор кафедры Информатика,
Ошский технологический университет им. М.М. Адышева

МЕТОДЫ РАЗРАБОТКИ ПРОГРАММНЫХ СРЕДСТВ РАСПОЗНАВАНИЯ ИЗОБРАЖЕНИЯ И ТЕКСТА

Настоящая статья посвящена проблемам разработки методов реализации и составления программных средств для обнаружения и распознавания изображений и текстов. Предложены алгоритмы и программы на языке Python с использованием приложения Open CV.

Ключевые слова: искусственный интеллект, нейронные сети, распознавание изображения и текста.

Халилов Дурбек Аминович, ф-м.и.к.,
Фергана филиалынын маалыматтык технологиялар кафедрасынын профессору, Мухаммад аль-Хорезми атын. Ташкент маалымат технологиялар университети,
Аттокуров Урмат Төлөгөнович, т.и.к., информатика кафедрасынын профессору,
М.М. Адышев атындагы Ош технологиялык университети,
E-mail:durbekamintatuff@umail.uz, urmat_at@mail.ru

СҮРӨТТӨРДҮ ЖАНА ТЕКСТИ ТААНУУ ҮЧҮН ПРОГРАММАЛЫК КАРАЖАТТАРДЫ ИШТЕП ЧЫГУУНУН УСУЛДАРЫ

Бул макала сүрөттөрдү жана тексттерди аныктоо жана таануу программалык куралдарын ишке ашыруу жана түзүү ыкмаларын иштеп чыгуу көйгөйлөрүнө арналган. Open CV тиркемесин колдонуу менен Python тилинде алгоритмдер жана программалар сунушталды.

Khalilov Durbek Aminovich, candidate of physical and mathematical sciences, professor of the Department of Information Technologies, Fergana branch of the Tashkent University of Information Technologies named after Muhammad al-Khorezmi
Attokurov Urmat Tologonovich, candidate of technical sciences, professor of the Department of Computer Science, Osh Technological University named after M.M. Adysheva

METHODS FOR DEVELOPING SOFTWARE TOOLS FOR IMAGE AND TEXT RECOGNITION

This article is devoted to the problems of developing methods for implementing and compiling software tools for detecting and recognizing images and texts. Algorithms and programs in Python using the Open CV application are proposed.

Ачкыч сөздөр: жасалма интеллект, нейрон тармактары, сүрөт жана текстти тааныу.

Распознавание образов, в том числе и символов, является на сегодняшний день одной из актуальных задач, возникающих в различных областях человеческой деятельности. Например, оно может использоваться в промышленности для автоматического распознавания деталей, в образовании для автоматизированной проверки бланков тестирования и в других областях.

Современные информационные технологии позволяют нам существенно упростить доступ к информационным ресурсам, накопленным человечеством, при условии, что они будут переведены в электронный вид. Наиболее простым и быстрым является сканирование документов с помощью сканеров. Результат работы является цифровое изображение документа – графический файл.

Следовательно, создание теории распознавания текста и его определения является основной задачей стоящее перед как государственных, в негосударственных, но и программистами создающие алгоритмы определения и распознавания тестов и живой человеческой речи.

Целью настоящей работы является сравнение и про анализированные возможности систем распознавания символов и текстов в целом.

Постановка задачи является рассмотрение основных понятий и определений теории распознавания текста, технологию оптического распознавания символов**.** Сравнить возможности программ распознавания символов.

Методы распознавания и этапы обработки изображения.

Системы распознавания реализуются как классификаторы, использующие различные методы:

- шаблонные (растровые);
- признаковые;
- структурные.

В классификаторе шаблонного типа с помощью критерия сравнения определяется, какой из шаблонов выбрать из базы. Самый простой критерий – минимум точек, отличающих шаблон от исследуемого изображения. К достоинствам шаблонного классификатора относятся хорошее распознавание дефектных символов («разорванных» или «склеенных»), простота и высокая скорость распознавания.

Недостатком является необходимость настройки системы на типы и размеры шрифтов. В признаковых классификаторах анализ проводится только по набору чисел или признаков, вычисляемых по изображению. Этот метод позволяет распознавать различные начертания символов, т.е. различные подчерки шрифты и т.д.

Этот метод неизбежно вызывает некоторую потерю информации, так как используется топологическое представление, отражающее информацию о взаимном расположении структурных элементов символа. Эти данные могут быть представлены в грифовой форме. При этом данный метод обеспечивает инвариантность относительно типов и размеров шрифтов. Недостатками являются трудность распознавания дефектных символов и медленная работа.

Основой структурно-пятнённого метода является структурно-пятненный эталон. Он имеет вид набора пятен с попарными отношениями между ними. Данное представление нечувствительно к различным начертаниям и дефектам символов.

Алгоритм основан на сочетании шаблонного и структурного методов распознавания образов.

При анализе образца выделяются ключевые точки объекта – так называемые «пятна». В качестве пятен, например, могут выступать:

- концы линий;
- узлы, где сходятся несколько линий;
- места изломов линий;
- места пересечения линий; крайние точки.

Основными этапами являются:

1. Предобработка. На этом этапе выполняются следующие задачи: повышение качества изображения за счет фильтрации, шумоподавления и других, имеющих своей целью повысить качество изображения. На этом этапе происходит очистка изображения от дефектов сканирования. В частности, в самом начале работы к изображению в целях шумоподавления часто применяется фильтр Гаусса. Важную роль играет пороговая бинаризация, то есть перевод изображения в чёрно-белый формат из цветного или оттенков серого. Это позволяет резко разделить текст и фон, упрощает в дальнейшем применение многих алгоритмов, а также избавляет от некоторых шумов на изображении. При этом используется гистограмма яркости изображения текста, на котором наблюдается два пика: высокий пик, соответствующий белому фону, то есть цвету бумаги, и пик в области тёмных пикселей, соответствующих яркости символов текста.

2. Выделение региона интереса. На этом этапе бинаризации изображении выделяется непосредственно область, на которой находится распознаваемый текст, и отбрасываются элементы, текстом не являющиеся. К ним относятся такие объекты, как кляксы, пятна на бумаге, не удалённые в процессе бинаризации, картинки и др. Для их удаления можно, например, выделять компоненты связности на изображении, вычислять геометрические признаки и на их основе классифицировать компоненту связности как часть текста или дефект, используя методы машинного обучения или эвристики.

3. Сегментация и нормализация текста. На этом этапе текст разделяется, или сегментируется, на удобные для анализа составные части. Наиболее естественными действиями на данном этапе является разделение текста на отдельные строки (сегментация строк) и разделение строк на слова (сегментация слов), а также, теоретически, разделение слов на элементарные составные части. Кроме того, на данном этапе проводится нормализация текста приведение выделенных составных частей к некоторому стандартному виду для снижения вариативности и упрощения распознавания.

4. Сегментация строк. Задача сегментации (разделения) строк в машина печатных документах на сегодняшний день считается полностью решённой. Но в задачах при разделении строк в общем случае возникают сложности, не позволяющие напрямую применять алгоритмы, пригодные для машина печатных текстов:

– строки не только могут не являться параллельными, но и могут изгибаться;

– различные строки могут быть слишком близки, а элементы текста, принадлежащего различным строкам, могут налагаться друг на друга.

Пересечение элементов различных строк представляет собой проблему не только сегментации строк, но и распознавания текста, так как отнесение элемента к неправильной строке очевидно ухудшает его распознаваемость. Пересекающиеся компоненты являются проблемой для методов горизонтальной проекции (так как они увеличивают значение профиля проекции в тех местах, где должен быть его минимум) группированных методов (так как они используют связные компоненты пикселей текста для построения строк), но слабо влияют на некоторые методы выделения

базовых линий. Для поиска пересекающихся элементов из различных строк можно использовать такие признаки, как размер компонент связности текста, факт отнесения одной компоненты к нескольким строкам или, напротив, не относящимся ни к какой строке. После нахождения таких сомнительных компонент нужно определить, относятся ли они к какой-то строке или же их нужно декомпозировать на элементы, относящиеся к разным строкам. Такая вертикальная декомпозиция компонент сложная задача. Простое решение заключается в разрезании компоненты на части горизонтальными линиями, но можно применить и более тонкие подходы, например, выделение отдельных штрихов.

5. Сегментация слов. На этом этапе работы системы распознавания выделенные строки текста разделяются на отдельные слова. В отличие от машинописного текста, в котором расстояние между словами более-менее постоянно, а интервалы между символами внутри слова гораздо меньше, чем интервалы между словами, в рукописном тексте размер интервалов между словами может варьироваться в очень широких пределах. Компоненты связности текста, отнесённые к одной строке на предыдущем этапе работы системы распознавания, объединяются в слова на этом этапе.

Варианты использования – схема

- Сканирование печатных документов в версии, которые можно редактировать с помощью обычных редакторов текста.
- Индексирование печатного материала для поисковых систем.
- Автоматизированная обработка и ввод данных.
- Расшифровка документов в текст, который может быть прочитан вслух для пользователей с нарушениями зрения.
- Архивирование исторической информации (газет, журналов), а также поиск по ним.
- Извлечение данных и передача в бухгалтерские программы (квитанции, счета).
- Размещение важных подписанных юридических документов в электронной базе данных.
- Распознавание номерных знаков с помощью камеры контроля скорости и программного обеспечения камеры с подсветкой.
- Сортировка писем для доставки почты.
- Перевод слов в изображении на заданный язык.
- Обеспечение поиска отсканированных книг.

*Задачи распознавания текста***.** Несмотря на то, что в настоящее время большинство документов составляется на компьютерах, задача создания полностью электронного документооборота ещё далека до полной реализации. Как правило, существующие системы охватывают деятельность отдельных организаций, а обмен данными между организациями осуществляется с помощью традиционных бумажных документов.

Задача перевода информации с бумажных на электронные носители актуальна не только в рамках потребностей, возникающих в системах документооборота. Современные информационные технологии позволяют нам существенно упростить доступ к информационным ресурсам, накопленным человечеством, при условии, что они будут переведены в электронный вид.

Наиболее простым и быстрым является сканирование документов с помощью сканеров. Результат работы является цифровое изображение документа - графический файл. Более предпочтительным, по сравнению с графическим, является текстовое представление информации. Этот вариант позволяет существенно сократить затраты на хранение и передачу информации, а также позволяет реализовать все возможные сценарии использования и анализа электронных документов. Поэтому наибольший интерес с практической точки зрения представляет именно перевод бумажных носителей в текстовый электронный документ. На вход системы распознавания

поступает растровое изображение страницы документа. Для работы алгоритмов распознавания желательно, чтобы поступающее на вход изображение было как можно более высокого качества. Если изображение зашумлено, нерезкой, имеет низкую контрастность, то это усложнит задачу алгоритмов распознавания.

В итоге проведенных исследований и разработок были получены следующие основные научные результаты.

1. Разработан метод сегментации объектов изображений, позволяющий выявить структуру сложных изображений за счет использования подхода «сверху вниз» и комбинации операций фильтрации и заливки, позволяющий правильно сегментировать строки в случаях слипания и перекрытия по вертикали символов соседних строк, а также при появлении на факсимильном документе тонкой вертикальной полосы. При работе с изображениями текстов, изменяя параметры фильтрации, можно сегментировать текстовые блоки на странице, строки внутри текстового блока или слова в строке. При этом данный метод сегментации лишен недостатка коммерческой программы Fine Reader, обнаруживающей две текстовых строки на изображении одной.
2. Реализован метод сегментации строк на изображении документа, использующий подход «снизу-вверх» и формирующий строки из набора сегментированных символов текстового блока, упорядоченных по возрастанию их координаты x, менее трудоемкий, и в то же время, позволяющий правильно находить строки в условиях перекоса страницы.
3. Разработан метод идентификации типов бланков факсимильных сообщений по характерному графическому фрагменту (логотипу), позволяющий производить быструю сортировку документов в системах обработки потоков факсимильных сообщений.
4. Разработан комбинированный нейросетевой метод распознавания, включающий предварительную классификацию символов по высоте и положению в строке и окончательное распознавание одиночных символов и слипшихся пар производящееся различными нейросетями, что обеспечивает распознавание строчных и прописных букв сходного начертания, позволяет выбирать оптимальный набор различаемых пар соединенных символов без переобучения соответствующей сети одиночных символов упрощает структуру сетей и обеспечивает качество распознавания до 99,4%.
5. Предложен способ извлечения полной информации, содержащейся в выходном векторе персептрона за счет использования не только максимального его элемента, но и близких к нему по значению, в качестве набора классификационных решений с разной степенью достоверности, что повышает качество последующего контекстного распознавания символов.
6. Разработан метод орфографической коррекции результатов нейросетевого распознавания символов. В отличие от метода на основе алгоритма Витерби, в данном методе при переборе вариантов распознавания текущего символа ищется N лучших цепочек символов среди всех, которые могут быть порождены этими вариантами, что повышает качество контекстного распознавания символов.
7. Произведены оценки трудоемкости основных этапов обработки факсимильных сообщений, определены пути распараллеливания этих этапов в части декомпозиции соответствующих алгоритмов и установления взаимосвязей элементов, соответствующих данных. Это является определяющей стадией разработки параллельных алгоритмов распознавания текста и позволяет на этой основе строить модели параллельных вычислений для постановки задач по обработке факсимильных сообщений на многопроцессорных системах, в том числе специализированных.

Литература:

1. Богданов В., Ахметов К. Системы распознавания текстов в офисе. // Компьютер-пресс -- 1999 №3, с.40-42.

2. Павлидис Т. Алгоритмы машинной графики и обработки изображений. М.: Радио и связь, 1986
3. Халилов Д.А. Конспект лекций. ФФ ТУИТ. 2020.