

АДАПТИВНЫЕ АЛГОРИТМЫ РАСПОЗНАВАНИЯ ТЕКСТОВ

*Обухов Вадим Анатольевич
Ахунова Таманнохон Зокир кизи
Сиддиков Муродали Йулдошали угли
Ферганский филиал Ташкентского университета
информационных технологий имени Мухаммада Ал-Хоразмий*

Аннотация: Адаптивные алгоритмы распознавания текстов - это область исследования, связанная с разработкой методов и моделей, способных автоматически адаптироваться к разнообразным типам текстовых данных и условиям окружающей среды. Эти алгоритмы являются ключевыми компонентами в сфере обработки естественного языка (Natural Language Processing, NLP) и машинного обучения.

Целью адаптивных алгоритмов распознавания текстов является повышение точности и эффективности распознавания текстовых данных в различных приложениях, таких как автоматическое распознавание рукописного письма, оптическое распознавание символов (OCR), анализ тональности текста, извлечение информации, классификация текстов и многое другое.

Ключевые слова: адаптивные алгоритмы, распознавание текстов, обучение на данных, автоматическая адаптация, многозадачность, учет контекста, NLP, машинное обучение.

Во всех системах электронного документооборота и системах ввода печатных текстов одним из ключевых этапов является распознавание текстовых символов - перевод информации из графической формы - результата сканирования — в текстовую форму. Несмотря на многолетнюю историю развития алгоритмов распознавания и существование большого количества алгоритмов, хорошо распознающих четко напечатанные тексты, задача распознавания в более сложных случаях далека от решения. Возникает задача дальнейшего увеличения точности распознавания документов низкого качества. В частности, существующие алгоритмы обеспечивают относительно невысокую по сравнению с человеком точность распознавания текстов с графических изображений, полученных сканированием с малыми разрешениями. Стоит отметить класс задач, в которых имеющееся графическое изображение невозможно улучшить путем увеличения разрешения сканирования или изменением параметров сканирования.

«Комбинированные алгоритмы в задачах распознавания текстов»

Главной задачей, стоящей перед системами ввода документов, является распознавание образов символов для представления их в текстовом (цифровом) виде в условиях временных ограничений. Критерии качества распознавания различаются в зависимости от назначения системы ввода. Например, персональным программам распознавания текстов с бумажных носителей свойствен естественный и объективный критерий качества, оцениваемый числом исправлений, необходимых для приведения результатов распознавания к приемлемому для последующей обработки виду. Для профессиональных систем массового ввода документов с проверкой результатов операторами не менее важным является критерий надежности распознавания. Объективные критерии качества распознавания документов требуют от алгоритмов распознавания символов возможности оценки их результатов на различных этапах функционирования. В настоящей работе делается попытка осмысления накопленных к настоящему моменту времени алгоритмов распознавания как частей программных продуктов ввода документов, ориентируясь на объективные критерии качества распознавания документов.

«Обработка и распознавание рукописного текста в системах электронного документооборота»

Входной информацией в системах электронного документооборота могут быть не только документы с печатным текстом, но и рукописные документы (документация паспортно-визовой службы, анкетирование, прием заявлений от населения). Также имеется большое количество унаследованных рукописных документов, содержащих важную техническую информацию, которые желательно было бы перевести в электронный вид.

«Методы распознавания текстов»

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.

«Как происходит распознавание текстов»

Сегодня нет необходимости заново набирать имеющийся текст, тратя на это драгоценное время. С этой работой помогают справиться многофункциональные устройства, которые выполняют ее в несколько этапов, освобождая человека от этой нудной процедуры. Во-первых, нужно ввести отсканированный документ в компьютер. Страница в этом случае выглядит как изображение, еще не готовое для дальнейшей работы с ним. Во-вторых, нужно произвести анализ макета, чтобы определить, где на странице находится текст, а где – таблицы и рисунки. Этот процесс выполняется при помощи OCR-приложения, которое позволяет разить текст на небольшие фрагменты, последовательно дробя их на предложения, слова и, наконец, самые мелкие – символы. Таким образом, конечным результатом данного этапа работы будет совокупность отдельных символов, каждый из которых находится в определенном месте страницы.

В соответствии со вторым методом программа сравнивает распознаваемый символ с тем шаблоном, который хранится в базе данных ее памяти. Этот метод называется методом сопоставления, но не очень удобен, поскольку на распознавание текста уходит много времени. Причиной низкой эффективности является также и то, что при использовании этого метода должно быть стопроцентное соответствие между символом и шаблоном, чтобы программа смогла распознать текст.

«Распознавание текста с помощью решений АБВУУ — все гениальное просто для бизнеса»

Программы для распознавания текста знакомы всем, кто в процессе работы сталкивался с необходимостью перевода печатных символов в электронный формат. Современные решения от лидера отрасли АБВУУ давно вышли за рамки массового сегмента: теперь они помогают бизнесу. Разработки в области распознавания текста востребованы в банковском деле, в образовании, энергетике и т. д. В этой статье мы расскажем о том, какие задачи бизнеса позволяют решать технологии АБВУУ.

«Лучшее бесплатное программное обеспечение для распознавания текста для Windows 10»

«Обзор методов распознавания структурированных символов»

Распознавание структурированных (печатных) символов различных изображений обеспечивает решение ряда научных и прикладных задач при идентификации объектов различной природы. Современные методы распознавания символов используются для решения как типовых задач, например распознавание текста, так и специализированных задач,

ориентированных на распознавание символьной информации, нанесенной на поверхность различных объектов. В настоящее время существует достаточно большое количество программ, предназначенных для распознавания текста (например, FineReader, Readiris, ScanSoft OmniPage, CuneiForm и др.). Каждая из этих программ предлагает свою реализацию решения задачи обработки и распознавания изображений. В основном эти программы являются коммерческими, поэтому методы решения задач, заложенные в них, известны только их разработчикам, и практически невозможно определить для каких задач они подходят и какие задачи им не под силу. Кроме того, все эти программы поставляются в виде исполняемых модулей, что делает невозможными анализ работоспособности программ, качества их работы и модификацию используемых ими математических моделей и алгоритмов.

В данной работе рассмотрено наиболее известные и распространенные методы решения задачи распознавания символов. На данный момент выделяют три основных подхода для решения сформулированной задачи: структурный, признаковый и шаблонный. Каждому методу присущи свои достоинства и недостатки.

В другой представлен обзор методов, используемых для решения задачи распознавания по голосу. Уделяется внимание устоявшейся структуре систем распознавания. Также приводятся краткие описания наиболее распространённых методов извлечения признаков (таких как MFCC и LPCC), а также методов классификации (метод векторного квантования, модель гауссовых смесей, метод опорных векторов). Обсуждаются методы оценки систем распознавания и представления результатов таких оценок.

Распознавание дикторов объединяет идентификацию и верификацию дикторов. Идентификация диктора — процесс определения личности по образцу голоса путём сравнения данного образца с шаблонами, сохранёнными в базе. Результатом процесса идентификации является список кандидатов. Реализующая система может выдавать список фиксированного размера либо принимать решение о включении пользователя в список кандидатов на основании заданного порога. Если предусмотрена возможность того, что в процессе идентификации будет участвовать пользователь, не зарегистрированный в системе, то говорят.

«Исследование и разработка системы распознавания текста на изображении»

В работе рассмотрена технология создания системы распознавания символов (с применением сверточных нейронных сетей) на изображении, а именно символов номерного знака автомобиля. На сегодня существует множество подходов к решению этой проблемы, причем большинство из них

неэффективны для изображений, символы которых расположены на сложном фоне и подвержены шумовым, аффинным и проекционным искажениям.

Проведен анализ существующих методов поиска и локализации текста на изображениях. Определены требования к системе распознавания текста на изображении сложных графических сцен, характеризующиеся неоднородностью фона, отсутствием чётких критериев отличия фона от текста, большой вероятностью разнообразных искажений. Определена возможность использования для решения задачи выделения связных областей изображения метода MSER. Для классификации областей изображения на «текстовые» или «не текстовые» предлагается использовать классификаторы AdaBoost, SVM. В качестве самостоятельного признака возможно использование алгоритма SWT. Он может использоваться как при обучении классификаторов, так и при объединении букв-кандидатов в слова или строки текста.

«Методы распознавания текстов»

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.

«Методы и средства обработки и хранения информации»

Результаты исследований в следующих предметных областях: цифровая обработка изображений, распознавание образов, геоинформационные системы, использование информационных технологий, базы данных и знаний, системы искусственного интеллекта, математическое и компьютерное моделирование.

Предназначен для научно-педагогических работников, инженеров, учащихся старших курсов вузов и специалистов в представленных областях знаний. В сборник включены материалы региональной научно-технической конференции молодых ученых и специалистов.

«Исследование методов распознавания текстов»

В работе рассматриваются и анализируются основные алгоритмы офлайн-распознавания рукописных текстов. В качестве наиболее перспективного предложен алгоритм распознавания на основе структурно-пятненого эталонного метода.

В статье «Анализ подходов и методов распознавания текста» авторами проводится аналитический обзор методов и подходов для распознавания графических символов в текстовых документах на примере наиболее популярных используемых систем. Приводятся достоинства и недостатки рассматриваемых методов и подходов. На основе проведенного анализа намечены возможные пути дальнейшего развития систем распознавания текста.

Распознавание - отнесение предъявляемых объектов к определённым классам с помощью применения известных правил классификации. Задача распознавания объектов на изображении сводится к задаче классификации, для которой применяется математический аппарат построения модели обучения по прецедентам. Модель строится автоматически по заранее собранному набору прецедентов - изображений, для которых известно, являются ли они изображениями объекта. Таким образом, процесс делится на два связанных между собой этапа:

- этап преобразования изображения в вектор признаков, в процессе которого определяется наиболее полное и информативное представление изображения в виде числового вектора;
- этап классификации, в процессе которого проверяется гипотеза принадлежности изображения классу изображений объекта на основании наблюдения (вектор признаков).

Основываясь на собранной информации, система распознавания образов должна правильно классифицировать объекты, характеристики которых заранее известны. Правильность распознавания зависит от объема различающей информации, содержащейся в измеряемых характеристиках, и эффективности использования этой информации.

Оптическое распознавание текста является исследуемой проблемой в областях распознавания образов, искусственного интеллекта и компьютерного зрения. Работы в данном направлении ведутся уже на протяжении нескольких десятилетий, и достигнутые результаты впечатляют. В настоящее время разработано достаточно большое число разнообразных методов. В отличие от

ранних версий, существующие OCR подходы не требуют калибровки для работы с конкретным шрифтом и позволяют обеспечить достаточно высокую вероятность достоверного распознавания большинства шрифтов. Некоторые системы оптического распознавания текста способны восстанавливать исходное форматирование текста, включая изображения, колонки и другие нетекстовые компоненты.

Использованная литература

1. ТОЛИБОВЕВ, I., RAYIMJONOVA, O., ISKANDAROV, U., MAKHAMMADJONOV, A., & TOKHIROVA, S. МИРОВАЯ НАУКА. МИРОВАЯ НАУКА Учредители: ООО" Институт управления и социально-экономического развития", (3), 26-29.
2. Tojiboev, I., Rayimjonova, O. S., Iskandarov, U. U., Makhammadjonov, A. G., & Tokhirova, S. G. (2022). ANALYSIS OF THE FLOW OF INFORMATION OF THE PHYSICAL LEVEL OF INTERNET SERVICES IN MULTISERVICE NETWORKS OF TELECOMMUNICATIONS. *Мировая наука*, (3 (60)), 26-29.
3. Muhammadjonov, A., & Toxirova, S. (2023). YARIMO 'TKAZGICHLARNING TURLARI. ICHKI VA TASHQI YARIMO 'TKAZGICHLAR. *Research and implementation*.
4. Isroilovich, H. A., & Abdimahamatovich, H. A. (2023). KIBERJINOYAT JAMIYAT UCHUN YANGI TAHDID SIFATIDA. *World scientific research journal*, 15(1), 249-252.
5. Abdimahamatovich, H. A., & Anatolyevich, O. V. (2022). SANOAT KORXONALARINING RIVOJLANISH TENDENSIYALARI. *Journal of new century innovations*, 11(1), 195-202.
6. Обухов, В. А., & Хакимов, А. А. (2022). ОСНОВЫ ИСПОЛЬЗОВАНИЯ РЕКУРСИВНЫХ ФУНКЦИЙ В СТРУКТУРАХ ДАННЫХ. *Journal of new century innovations*, 11(1), 92-99.
7. Обухов, В. А., & Хакимов, А. А. (2022). МОДЕЛИРОВАНИЕ РЕГИСТРОВ ПРОЦЕССОРА. *Journal of new century innovations*, 11(1), 169-178.
8. Abdullayeva, M., & Hakimov, A. (2023). ZAMONAVIY AXBOROTLASHGAN JAMIYATDA SANOAT KORXONALARIGA AXBOROT TECHNOLOGIYALARINING TADBIQI. *Research and implementation*.
9. Z. Qadamova TATU Farg'ona filiali magistri D.Sotivoldiyev Fiskal instituti dotsenti BIOLOGIK NEYRONLARNING MODELI, SUN'IY NEYRON TARMOQLARINING INSONIYAT HAYOTIDAGI AXAMIYATI// Международная научно-техническая конференция «Практическое применение технических и цифровых технологий и их инновационных решений», Т
10. Набижонов, Р., & Обухов, В. (2023). ДАЛЬНЕЙШИЙ ВКЛАД БЛОКЧЕЙН-СЕТЕЙ В РАЗВИТИЕ ДИСТАНЦИОННОГО ОБРАЗОВАНИЯ. *Research and Implementation*. извлечено от <https://fer-teach.uz/index.php/rai/article/view/772>

11. Nabijonov, R. M. o'g'li, & Mamayeva, O. I. qizi. (2023). TA'LIM SIFATINI OSHIRISHDA ELEKTRON AMALIY DASTURIY PAKETLARNING AHAMIYATI. GOLDEN BRAIN, 1(25), 51–55. Retrieved from <https://researchedu.org/index.php/goldenbrain/article/view/4782>
12. Обухов , В., Хамидов Э., & Набижонов , Р. (2023). ПОЭТАПНОЕ ВНЕДРЕНИЕ БЛОКЧЕЙН ТЕХНОЛОГИЙ В РЕСПУБЛИКЕ УЗБЕКИСТАН. Research and Implementation. извлечено от <https://fer-teach.uz/index.php/rai/article/view/770>
13. Обухов, В. (2023). 5 СПОСОБОВ, КОТОРЫМИ БЛОКЧЕЙН ПОВЛИЯЕТ НА ИНДУСТРИЮ ОБРАЗОВАНИЯ. Engineering problems and innovations.
14. Ходжиматов, Ж. М., Хамидов, Э. Х., & Собиров, М. М. (2022). ОСНОВНЫЕ СОВРЕМЕННЫЕ ЯЗЫКИ ПРОГРАММИРОВАНИЯ. Journal of new century innovations, 11(1), 136-143.